Lecture Notes in Electrical Engineering 260

Yueh-Min Huang
Han-Chieh Chao
Der-Jiunn Deng
James J. (Jong Hyuk) Park *Editors*

Advanced Technologies, Embedded and Multimedia for Human-centric Computing

HumanCom and EMC 2013

Springer

Lecture Notes in Electrical Engineering

Volume 260

For further volumes:
http://www.springer.com/series/7818

Yueh-Min Huang · Han-Chieh Chao
Der-Jiunn Deng · James J. (Jong Hyuk) Park
Editors

Advanced Technologies, Embedded and Multimedia for Human-centric Computing

HumanCom and EMC 2013

Springer

Editors

Yueh-Min Huang
ES Department
National Cheng Kung University
Tainan
Taiwan, R.O.C.

Han-Chieh Chao
Institute of Computer Science and Information Engineering
National Ilan University
Ilan City
Taiwan, R.O.C.

Der-Jiunn Deng
Department of Computer Science and Information Engineering
National Changhua University of Education
Changhua City
Taiwan, R.O.C.

James J. (Jong Hyuk) Park
Department of Computer Science and Engineering
Seoul University of Science and Technology (SeoulTech)
Seoul
Republic of Korea (South Korea)

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-94-007-7261-8 ISBN 978-94-007-7262-5 (eBook)
DOI 10.1007/978-94-007-7262-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013943942

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Part I HumanCom

Part II Special Session

Part III Exploration of Scientific Evidence on Affective Learning

Part IV Multimedia Technology for Education

Part V Modern Learning Technologies and Applications with Smart Mobile Devices

Part VI Embedded Computing

Part VII Multimedia Computing

Part VIII Smart System

Part IX Cross Strait Conference on Information Science and Technology

Part X Advances in Multimedia Algorithms, Architectures, and Applications

Part XI Virtual Reality for Medical Applications

Part XII Recent Advances on Video Analysis and its Applications

Part XIII All-IP Platforms, Services and Internet of Things in Future

Part XIV Networking and Applications

Part I
HumanCom

An Interface for Reducing Errors in Intravenous Drug Administration

Frode Eika Sandnes and Yo-Ping Huang

Abstract Input errors occur with drug infusion pumps when nurses or technicians incorrectly input the prescribed therapy through the control panel. Such number copying tasks are cognitively and visually demanding, errors are easily introduced and the process is often perceived as laborious. Stressful working conditions and poorly designed control panels will further add to the chance of error. An alternative scheme is proposed herein, termed intravenous prescription phrase, based on dictionary coding. Instead of copying number sequences the user copies sequences of familiar words such that the cognitive load on the user is reduced by a factor of five. The strategy is capable of detecting errors and is easy to implement.

Introduction

The operation of medical devices, in particular the input of numeric values has received much attention [1]. Intravenous drug administration involves the input of rate, dose, time and volume of drugs prescribed by doctors and input errors can be lethal. Studies have shown that error rates in intravenous drug administration can be as high as 50–80 % [2, 3]. The technical complexity of menu structures constitute one problem with such medical devices is [4]. Another key issue is the copying of digits such as when placing phone calls. Often this involves a number copying task where the number is read from some source such as a phone directory

F. E. Sandnes (✉)
Institute of Information Technology, Oslo and Akershus University College of Applied Sciences, N-0130 Oslo, Norway
e-mail: eika.sandnes@hioa.no

Y.-P. Huang
Department of Electrical Engineering, National Taipei University of Technology, 10608 Taipei, Taiwan, Republic of China
e-mail: yphuang@ntut.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_1, © Springer Science+Business Media Dordrecht 2014

and input using the numeric keypad [5]. Airlines often use booking references which users input on self-service check-in terminals in the airport to obtain their boarding cards. These higher base numbers, such as base 36, often comprise combinations of letters and digits.

Shannon [6] identified that a decimal digit is approximately equal to 3 1/3 bits. Similarly, base 36 numbers are equal to 5.2 bits. One may argue that the increased information capacity gained through large base-digits does not justify the increased complexity of the copying tasks. One reason is that certain letters and digits look similar such as 0—O, 1—l, 2—Z, 5—S, 6—G and 8—B. Differentiating such symbols is even more difficult when the user has reduced visual acuity and given certain fonts. High base numbers appear as random strings. The lack of internal structure means that users are unable to resolve ambiguous looking characters. This problem is well known in the optical character recognition literature [7].

It has been found that number input errors are caused by either motor slips where the fingers do not perform as expected, recall slips where numbers are remembered incorrectly, or perception slips where the source number is read incorrectly [8]. These errors can result in digit substitutions, insertions or omissions. Most digit and text input studies operates with typical error rates of 5–10 %, which means that up one in every 10 digits are likely to be incorrect. A study of the frequency of digits in drug infusion pumps showed that 0 is the most frequent digit, followed by 1 and 5, while 3, 4, 6, 7 and 8 are comparatively rare [9]. Numbers ranged in 1–5 digits in length with 3 digits being the most common. One class of errors termed "out by 10 errors" is caused by incorrectly entering a decimal point or a zero [10, 11]. Out of 10 errors greatly affect the magnitude of numbers and will have serious consequences.

Several studies have investigated various number input interfaces including touch based [12] numeric keypads, displays with individual incremental up-down buttons, incremental left–right/up-down buttons or 5-button interfaces [13] that are often attached to mobile equipment in ambulances, hospital wards, etc., and results show that the slower incremental up-down interfaces leads to fewer errors than numeric keypads [14]. Another approach is to avoid the input of digits altogether where the prescription is printed on bar-codes and input using bar-code readers [15]. However, bar codes require printing and they are less flexible, as the information cannot be communicated orally. This study focuses on the manual input of prescriptions.

Number Copying Challenges

Number copying tasks are problematic for several reasons. Miller's limit on humans' short term memory of 7 ± 2 pieces of information [16] is often cited and subsequent studies have narrowed this limit down to a maximum of 5–7 pieces [17]. Studies involving memorizing phone numbers have shown that recall performance rapidly degrades beyond 6 digits [5]. A digit copying task therefore has to be split up into several read-input cycles, each cycle requiring alternating the visual attention between the source and the target. This is difficult if the source digits are not grouped such as 16 digit credit cards numbers presented groups of four. Without grouping additional effort is required to locate where to resume when shifting attention between the source and the target. Another issue is a lack of standard interface layout. If there is a mismatch in both information sequence and information location on the printed prescription and the target it will harder for users to pair source and target.

Second, digit copying tasks are error-prone and parity checks are unable to capture all errors.

Third, digit copying is perceived as time-consuming and laborious. Many portable laptop computers do not have numeric keypads and it has been demonstrated that the input of digits on QWERTY keyboards without numeric keypads are significantly slower than using full keyboards with numeric keypads [18]. It is occasionally necessary and practical to communicate number information from one individual to another orally via phone, such as if receiving a prescription from a medical doctor via phone. Each digit will have to be read out one by one and often repeated several times to confirm the correctness of the data. This further reduces efficiency and increases the chances of error and frustration.

Memory Aids

One popular memory aid is to remember number sequences associated with word input on mobile keypad, for example 2,326 m as ADAM. Instead of recalling the individual digits where each digit counts as one piece of information, the word is remembered counting as one piece of information. Unfortunately, not all number sequences have corresponding words as 0 and 1 are not assigned letters.

This paper proposes to use a fixed dictionary for encoding and decoding. Digit sequences are split into groups, where each group is converted to a linguistic word. The user is presented with a word sequence instead of a number sequence and will thus be able to copy more information per copy-input cycle. Group sizes of 5 digits are chosen as this relies on wordlists with 100,000 entries.

For example, 01234 56789 is first split into two groups of 5 digits, namely 01234 and 56789. Each number is then looked up in the wordlist (see Fig. 1). The digits 01234 could correspond to the English word “stir” and 56789 correspond to “galumphs” and the word sequence “stir galumphs” is presented to the user, who has the memory capacity to simultaneously hold both words in working memory while copying the words to the target. Next, the words are identified in the wordlist on the receiving end and the indices of the two words identified.

The wordlist can be organized such that the magnitude of the number correlates with the word length such that smaller numbers with many prefix zeros are

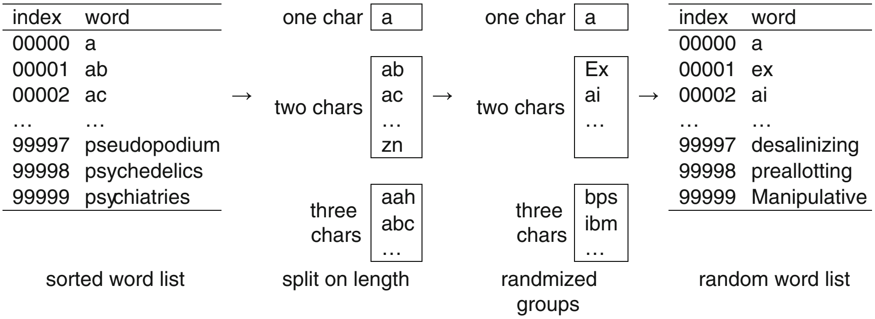


Fig. 1 Composing the error tolerant wordlist

assigned shorter words and larger number are assigned longer words. Another strategy would be to assign the most frequently used numbers shorter words.

A key advantage of linguistic words is that readers read words as whole units, while digit sequences are read individually. The reader recognizes the height signature of lower case letters. This greatly adds to the reading speed. Moreover, users are able to spot simple errors due to the internal word structure. Visual perception errors are reduced since words are read as one unit rather than individual unrelated units. Recall errors are also reduced for the same reason.

Error Detection

The literature on spell checking classifies input errors as deletions, insertions or replacements [19]. This strategy proposes two levels of error detection. The first level of error detection catches misspelled words not found in the wordlist.

Next, the wordlist is organized such that similar words are associated with dissimilar digit sequences. The Levenshtein distance is often used to measure the distance between words [20], but the Hamming distance can be used if the words have equal lengths. The following scheme is employed. The list of 100,000 words is first sorted into groups of equal length and words in each group shuffled into random order. Finally, the groups are recombined into one list organized according to increasing word lengths while exhibiting a random internal structure.

For example, the word "buys" can easily be mistyped as "buts" by substituting the character y with t. The Hamming distance between these words are 1. The corresponding numbers for the two words are 1225 and 3022. These numbers have a Hamming distance of 3 and a numeric distance of 1797. With running numbers the two numbers would be 1352 and 1354, respectively, yielding a Hamming distance of 1 and a numeric distance of 2. Figure 1 illustrates the wordlist construction process.

A second level of error detection is achieved by introducing a parity check [21] by summing each 5-digit chunk multiplied by unique factors and computing the desired modulus 100,000 of the total. By multiplying each number with a factor the parity check will detect if elements are swapped. A modulus 10 scheme misses 10 % of the errors, while modulus 100,000 only misses 0.001 % of the errors.

The linguistic digit representation can also assist detecting errors when comparing numbers. Imagine a customer comparing the dates 23-06-2012 and 23-08-2012, coded as 23062 and 23082, respectively. Each date comprises three information parts, day, month and year. If the user focuses on the day he or she may overlook the difference in month, that is, June versus August since the shape of 6 is similar to 8. This will not happen when comparing their linguistic representations “thrives” and “marxist”.

Information Coding Schemes

This section illustrates how to code values with the proposed scheme. Each unit should be fitted to a chunk to avoid overlap across chunks.

Large numbers with more than 5 digits are split into several chunks. Some values with more than 5-digits can be reduced to 5-digit chunks without information loss. Dates are often described with eight digits, namely day, month and year. One simplification is to use a modulo-10 single digit for the year. This will work as most applications operate within a limited time scope. This date format allows dates to be specified using a single linguistic word which is useful for making comparisons. Comparisons across years are simplified further if the year is dropped altogether.

Numbers comprising 5 digits or less can simply be represented using heading zeros such as small quantities, drug doses, etc. Analysis of commercially available equipment showed that numbers frequently represent volumes in ml or rates in ml/hr with rarely more than three digits and one digit after the decimal point. Such numbers can therefore be represented using the least significant digit of the 5-digit sequence to represent the digit after the decimal point and the four most significant digits to represent the digits on the left of the decimal point. For example 0.1 would be 00001, 10 or 10.0 would be 00010, 450 would be 04500, etc.

However, most drug infusion pumps do not actually have the precision indicated by the operating panels. Some equipment specifies an error in dosage of about 12 % from the set value. It is thus pointless to distinguish between 450.1 and 450.0 ml.

Pairs of related parameters with similar magnitudes can be combined into 5 digit numbers as a type of double floating point numbers. A parameter is represented using its two most significant digits, that is A and B, for example AB0, AB or A.B, and 3 bits are used to indicate the position of the decimal point from 10-1 to 104 represented by the values from 0 to 6, respectively. For example to code the parameter pair 2.0 and 1.5 using this scheme one would get 02015. The first digit 0 indicates the decimal point in the most leftmost position. To code the number pair 45 and 35 the decimal is moved one position to the right and the first digit is therefore 1 yielding the value 14535. Similarly, the pair 450 and 350 can be coded as 24535.

Materials and Methods

An English wordlist published by the SIL International Linguistics Department containing 109,582 entries was used. The entries were sorted according to increasing word length and the 100,000 shortest words were kept. The average word length is 8 and the maximum words length 12 characters. The size of the final wordlist was less than 1 Mb making it applicable to devices with limited memory. A proof of concept coder and decoder were implemented in Microsoft Excel. Excel was chosen in order to demonstrate that it is simple to implement. Several drug infusion pump user manuals were studied to acquire information about common intravenous drug administration practices and prescription parameters, namely Curlin Medical's 4,000 Plus and 4,000 CMS Ambulatory Infusion Systems, Abbott Laboratories' PLUM A+, Eureka's IP and LF infusion pumps and BodyGuard's 323 Multi-Therapy ambulatory infusion pump. Common parameters are listed in Table 1. Of the less obvious parameters titration limit specifies the maximum infusion rate, keep vein open rate specifies the rate of fluid transmitted when the devices is in an open state. This measure is specified as the overall rate is achieved by injecting drugs in smaller doses, and this specifies the rate for each dose.

Table 1 Common drug infusion pump parameters

Parameter	Resolution	Max	Unit	Type	Flag no.
Bag volume	1	9,999	ml	N/A	
Concentration	0.1	999	mg/ml	Required	
Start time	00:01	23:59	Hours: minutes	Optional	1
Titration limit	0.1	400	ml/hour	Optional	2
Amount to infuse	0.1	9,999	ml	Required	
Rate of injection	0.1	400	ml/hr	Required	
Vein open rate	0.1	10	ml/hour	Optional	2
Loading dose	0	50	ml	PCA	3
Bolus dose	0	50	ml	PCA	4
Delta rate	0	50	ml	PCA	5
Delta time	0	60	Minutes	PCA	5
Max bolus dose	0	100	ml	PCA	4
No. bolus dose/hr	0	15	Doses	PCA	4
Min bolus interval	0	60	Minutes	PCA	4
Up-ramp	00:01	99:59	Hours: minutes	Optional	6
Down-ramp	00:01	99:59	Hours: minutes	Optional	6

The following parameters apply to patient controlled analgesia therapies: Loading dose indicates an initial dose of drugs infused at the beginning of a therapy. Bolus dose indicates the amount of drugs in lm. Delta rate and delta time specifies the gradual increase of a dose per time unit. Maximum bolus dose indicates the maximum dose that can be demanded by the patient, minimum bolus interval specifies the minimum time interval between consecutive bolus doses and number of boluses allowed per hour specifies an upper limit to how many bolus doses the patient are allowed per hour. Up-ramp and down-ramp parameters can be used to gradually increase and decrease doses.

This parameter list comprise six optional parameters, therefore a check word can be introduced to indicate both which optional parameters that are included in the message and a parity symbol for error detection purposes. The two most significant digits are used to indicate the parameters included and the three least significant digits constitute a parity symbol.

Results and Discussion

The following example illustrates how the proposed strategy is applied to the following simple fictitious prescription:

Concentration	1.5 mg/ml
Amount to infuse	75.0 ml
Rate of injection	15.0 ml/hr

The corresponding number sequence for this prescription is:
00695 00015 00750 00150
Here 00695 is the control value where the first two heading zeros indicate that no optional parameters are present. The parity value 695 is computed as the modulus 1,000 of the sum of the three parameters computed with weights 3, 2 and 1. The corresponding word prescription phrase sequence is
but md lip pub
Setting three digits may not seem hard, but it would probably require three read-input cycles. However, under stress one may misread one of the digits or swap the numbers that are relatively similar in structure and magnitude.

The corresponding four word phrase “but md lip pub” is easier to remember, and can be input in one read-input cycle. Moreover, any mistakes would be detected, for instance if the word “but” is incorrectly input as “byt” since y is next to u on the QWERTY-keyboard. The letter sequence “byt” would be detected as an invalid word.

Next, imagine the word “bit” input instead of “but” as the character i is next to u on the keyboard. The word bit gives parity 456, which indicates that it is a mistake somewhere since it does not match the parity value 695.

If two of the words are accidentally swapped, for instance md and lip, giving the incorrect phrase "but lip md pub", the parity becomes 430 which does not match 695.

If we include an extra word accidentally, for example md twice, giving "but md md lip pub" the error is detected as there are no flags indicating optional parameters. Also, the parity values do not match. Next, consider the following prescription:

Concentration	1.5 mg/ml
Titration limit	30.0 ml/hr
Amount to infuse	75.0 ml
Rate of injection	20.0 ml/hr
Vein open rate	10 ml/hr
Loading dose	5.0 ml
Bolus dose	1.0 ml
Delta rate	1.0 ml
Delta time	5 min
Max bolus dose	10 ml
No. bolus dose/hr	2
Min bolus interval	30 min
Up-ramp	03:00
Down-ramp	06:00

This prescription involves 14 data items with 27 digits and 5 separators that possibly would have to be input through an intricate procedure involving different menu screen in 14 read-input cycles. This prescription is represented using the following string:
37983 00015 00300 00750 00200 0010 00050 00010 00010 00005 00010 00002 00030 00300 00600
All but the first parameters are included meaning that all bit flags apart from the first bit are set, giving 37. The parity value of the parameters is 983 using weights from 13 to 1 from left to right. The corresponding drug prescription phrase is thus flipper md fen lip lap us od us et us ai on fen doz

The resulting 15 word phrase can be copied in three read-input cycles, as "flipper md fen lip lap", "us od us et us" and "ai on fen doz".
The two examples illustrate a very general and flexible approach. However, some values occur repeatedly such as 10 resulting in monotonous phrases with several similar ("us"). Since most drug infusion pumps do not actually have the precision indicated by the operating panels the final example illustrates the use combined parameter pairs for obtaining shorter intravenous prescription phrases. The following related parameters from the previous example can be paired, titration limit (30 ml/hr) and keep vein open rate (10 ml/hr) coded as 13010, loading dose (5 ml) and bolus dose (1 ml) coded as 05010, delta dose (1 ml/hr) and delta time (5 min) coded as 10105, max bolus dose (10 ml), bolus doses per hour (2) and minimum bolus interval (30 min) coded as 10230, and finally up-ramp (03:00) and down

ramp (6:60) coded as 11836. The resulting string is thus37271 00015 00750 00200 13010 05010 10105 10230 11836
which gives the following intravenous prescription phrase
cappers md lip lap vegans divvy nitty babes milton
This phrase is easily remembered in two steps, namely "cappers md lip lap" and "vegans divvy nitty babes milton". This optimized phrase does not compromise accuracy.

Conclusions

A strategy for reducing the errors in intravenous drug administration is proposed. It comprises a memory aid that simplifies manual number copying tasks. The strategy converts the prescription data comprising digits sequences to sequence of linguistic words. Ordinary drug prescriptions can therefore be memorized for short durations. In addition the strategy allows users to easily verify the correctness of information as it is easier to compare linguistic words than digit sequences.

References

1. Oladimeji P (2012) Towards safer number entry in interactive medical systems. In: Proceedings of the 4th ACM SIGCHI symposium on engineering interactive computing systems (EICS '12), ACM, New York, pp 329–332
2. Barber N, Taxis K (2004) Incidence and severity of intravenous drug errors in a German hospital. Eur J Clin Pharmacol 59:815–817
3. Taxis K, Barber N (2003) Ethnographic study of incidence and severity of intravenous drug errors. BMJ 326:684
4. Nunnally M, Nemeth CP, Brunetti V, Cook RI (2004) Lost in menuspace: user interactions with complex medical devices. IEEE Trans Syst Man Cybern Part A 34:736–742
5. Raanaas RK, Nordby K, Magnussen S (2002) The expanding telephone number part 1: Keying briefly presented multiple-digit numbers. Behav Inf Technol 21:27–38
6. Shannon CE (2001) A mathematical theory of communication. SIGMOBILE Mob Comput Commun Rev 5:3–55
7. Kahan S, Pavlidis T, Baird HS (1987) On the recognition of printed characters of any font and size. IEEE Trans Pattern Anal Mach Intell PAMI-9 2:274–288
8. Wiseman S, Cairns P, Cox A (2011) A taxonomy of number entry error. In: Proceedings of the 25th BCS conference on human-computer interaction (BCS-HCI '11) British Computer Society, UK, pp 187–196
9. Wiseman S (2011) Digit distributions, What digits are really being used in hospitals? In: Proceedings of the fourth York doctoral symposium on computer science, The University of York, pp 61–68
10. Thimbleby H, Cairns P (2010) Reducing number entry errors: solving a widespread, serious problem. J R Soc Interface 7:1429–1439
11. Thyen AB, McAllister RK, Councilman LM (2010) Epidural pump programming error leading to inadvertent 10-fold dosing error during epidural labor analgesia with Ropivacaine. J Patient Saf 6:244–246

12. Isokoski P, Koki M (2002) Comparison of two touchpad-based methods for numeric entry. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI ‘02), ACM, New York, pp 25–32
13. Cauchi A (2012) Differential formal analysis: evaluating safer 5-key number entry user interface designs, EICS’12. In: Proceedings of the 4th ACM SIGCHI symposium on engineering interactive computing systems, ACM, New York, pp 317–320
14. Oladimeji P, Thimbleby H, Cox A (2011) Number entry interfaces and their effects on error detection, LNCS, vol 6949. Springer, Heidelberg, pp 178–185
15. Meyer GE, Brandell R, Smith JE, Milewski FJ, Brucker P, Coniglio M (1991) Use of bar codes in inpatient drug distribution. Am J Health Syst Pharm 48:953–966
16. Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81–97
17. Simon HA (1974) How big is a chunk? Science 183:482–488
18. Sandnes FE (2010) Effects of common keyboard layouts on physical effort: implications for kiosks and Internet banking. In: Proceedings of Unitech 2010, Tapir Academic Publishers, Trondheim, pp 91–100
19. Kukich K (1992) Techniques for automatically correcting words in text. ACM Comput Surv 24:377–437
20. Navarro G (2001) A guided tour to approximate string matching. ACM Comput Surv 33:31–88
21. Wagner NR, Putter PS (1989) Error detection decimal digits. Commun ACM 32:106–110

A Vocabulary Learning Game Using a Serious-Game Approach

Kanako Nakajima and Tatsuo Nakajima

Abstract It is always hard to keep motivated while doing something we must do but we do not want to. However, gamers put so much time into their favorite games, just because its fun. Games have many tricks to keep attracting people, and nowadays these gimmicks are included into education-games. However, not many of the education-games in the markets are fun enough to keep users motivated for playing. In this paper we address this conflict, propose a better education-game created based on a popular smartphone game, and evaluate the improvement of the motivation through playing the game we offer. As a conclusion, we discovered that the examinees are motivated through the experiment using the education-game we created; however, these motivations are passive as they are not actively willing to do, but rather not mind doing it. Supplementations to shift these passive motivations to active motivations are considered in our future work.

Keywords Serious game · Education

Introduction

It is always hard to keep motivated while doing something we must do but we do not want to, such as studying and working. On the other hand, there is something we enjoy doing it although we do not have to: games. Games are loved by many; in the research held in 2009 in Japan, 34 % of all Japanese citizens, and 64 % of Japanese elementally students play games regularly [1], and those gamers are willing to play games for there lives [2]. Good games have power to attract people, thus applying game mechanics to education has already been investigated for

K. Nakajima (✉) · T. Nakajima
Department of Computer Science and Engineering, Waseda University, 3-4-1 Okubo
Shinjuku, Tokyo 169-8555, Japan
e-mail: kanako.n@dcl.cs.waseda.ac.jp

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_2, © Springer Science+Business Media Dordrecht 2014

several times. However, it is doubtful that the education games produced in the market today work well enough. According to *the Annual Video Game Industry Report*, the peak of the education game market was in 2006–2007 in Japan, with the release of *Brain Age: Train Your Brain in Minutes a Day! and English Training: Have Fun Improving Your Skills!. Brain Age: Train Your Brain in Minutes a Day!*, the one sold very well, is full of game mechanics, and even though game was not the main part of *English Training: Have Fun Improving Your Skills!*, it included some of game parts and sold well. Many education games are released after the hit of *Brain Age*, but most of them were not successful. The reason of the failure of the education game industry is based on misusing the concept of "game". According to the *Annual Video Game Industry Report In 2008*, "many of the education game titles in the market today are just reprints of the existing books, thus they can reduce the cost of software development" [1]. These software are sold as they are games, but most of the times there is almost no game mechanics are included. Thus, it could not continue the boom of education games.

In this paper, we produced a English vocabulary learning game, Vocab Draw, by adding some educational elements to the game instead of adding game mechanics to education; with this approach, we believe that the gaming attributes remain while learning materials are provided.

Preliminary Survey

Preliminary Survey has been done for learning examinees' attitude towards games, especially the educational ones. The samples for this survey are 10 males and 7 females, all around 20-year-old.

Eighty eight percent, 15 of 17 samples are active game players. For those game players, such a question is asked: *when they quit playing the game they are playing if it is not fun*. None of the examinees answered they play until the end whether it is good or bad, and 1/3 of them answered they quit playing immediately when they find out the game is boring. 2/3 of them play for a while even tough they cannot find fun in it, waiting for the game to gets better by any change. With this result, it can be said that amusingness is important for games to make players continue playing it.

However, as its discussed in the previous section, recent education games are not attracting people well. 9 of 17 examinees have had experience of playing education games. The education games they played were able to be divided into three groups, brain training, English learning, and other. Taking close look at the brain training and English learning games, the willing to play the game again towards those two types were different; those who played brain training games are willing to play the game again, because it is fun. On the other hand, those who played English learning game are also willing to play the game again, but only to learn english, even though the game was boring. However, it is doubtful that boring game can attract people for long time even if it is efficient for learning.

This assumption can be verified by the popularity of education games today. By the *Annual Video Game Industry Report*, the sales of the education games in 2011 were 300 thousands, 0.56 % of all genres of games, the lowest genre out of 13 others. Also, according to our survey, the popularity of education game was the lowest as well. Even though the players say they will play the education games for their goods, while other interesting games are out there, it is hard for them to pick a boring educational game. And if they cannot continue playing the game, the efficiency of learning would become waste. Thus, we put focus on making the game interesting, rather than improve the efficiency.

Proposal of a Vocabulary Learning Game

Draw Something

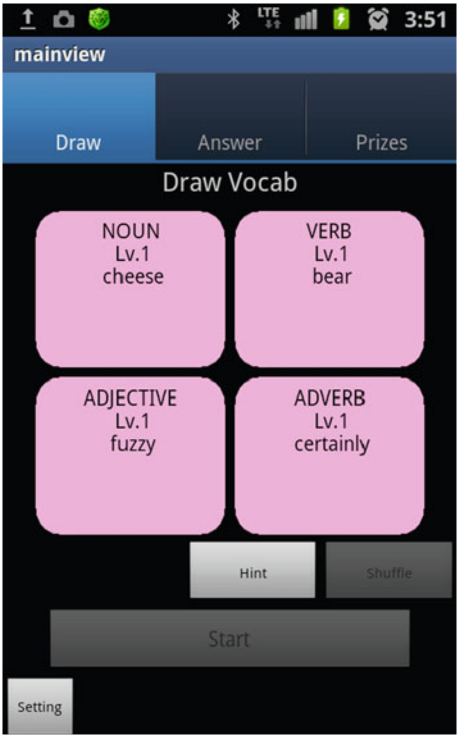
As we developed an interesting English vocabulary learning game, we used an existing popular game as a reference. The game is called *Draw Something*, published by OMGPOP. This game is distributed for iPhone and Android, and 20 million people have played free and non-free version in total on Android. The game is originally provided in English, but since it is for native English speakers there is no learning elements of English.

The basic rule is simple; two players draw and exchange their drawings, and make guess what other players have drawn. First of all, a player chooses an opponent and begins the game. Then a word selection scene is shown, and he chooses a word he might be able to draw. A level is set to each word, and the reward to each level is different. The player must choose an appropriate word from the words he can draw, while considering the rewards he can get. After the player has chosen the word, it is time for him to draw. The drawing is sent to the opponent he chose in the beginning, and then the opponent can guess what he has drawn. If the opponent corrects the answer, then their continuity is counted up as combo and some coins, which depends on the word level, are rewarded to the players. And then the drawing is now sent from the opponent to the player, and they keep exchanging the drawings to add up the combos. This series of action can be done in parallel, with many opponents at a time.

Vocab Draw

The rules we applied to Vocab Draw, our novel English vocabulary learning game, is exactly same as Draw Something. However, we added some learning elements to Draw Something in order to make it as an attractive education game. Also, we

Fig. 1 Draw-part1



added a beneficial change to Draw Something; so players can keep their motivation. In this section, the difference between Draw Something and Vocab Draw is presented.

Interface Change As several changes are made from Draw Something to Vocab Draw, the interface was upgraded. The game part is decided into two main parts, one is Draw-part and another is Answer-Part as shown in Figs. 1 and 2. Each part starts with a word selection part, followed by actual Drawing- and Answering-parts, shown in Figs. 3 and 4.

Improvement from Draw Something At first, we changed a way of drawing exchanging from one-to-one to one-to-many. In Draw Something, a player was be able to send to only one opponent per drawing, and when his opponent answered the game continued. In Vocab Draw this rule is discontinued because with this method the player must wait until the opponent answers, in order to keep playing. This mechanism is shown in the Fig. 5, when exchanging the drawing one by one, if an opponent's answer is slow or given up because he decreased the motivation to play the game, the player cannot get the feedback to his drawing. In such case, it is worried that the player's motivation is also declined. Instead in Vocab Draw, the relationship between players are reconsidered as one-to-many from one-to-one in Draw Something. As shown in Fig. 6, if the drawing is distributed to many rather than one, the chance of getting feedback is increased.

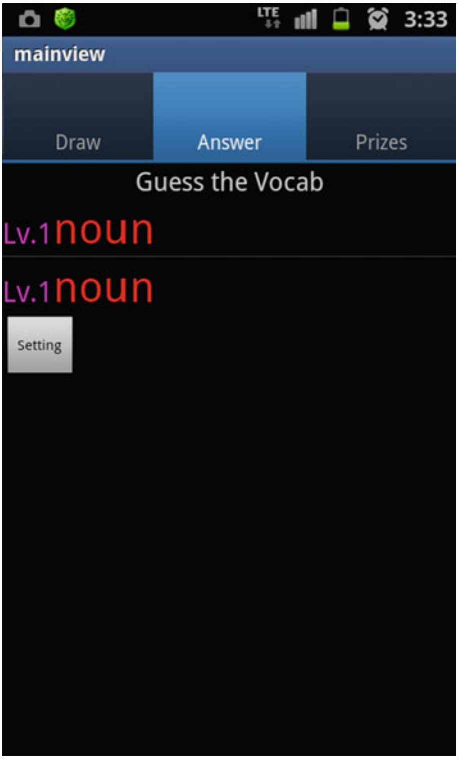


Fig. 2 Answer-part1

Learning Elements In order to include educational functions into Vocab Draw, materials listed below are considered:

1. Player Level
2. Coverage of fundamental part of speech
3. Japanese translation.

Firstly, the concept of level is given to each player. A player sets a level and the system shows the vocabularies according to that level. With this level, a player can learn vocabularies effectively as he tries the problems he can actually understand. In Figs. 1 and 2, the level is set as one.

Secondly, vocabularies are chosen by four fundamental parts of speech: noun, verb, adjective, and adverb. These parts of speech are considered fundamental, based on the English Note, which is published by the Ministry of Education in Japan as the base of English teaching material for Japanese elementary students [3]. In English Note, vocabularies provided for elementary students are nouns, verbs, adjectives, adverbs, and prepositions. Although preposition is as important as other five parts of speech, since it is used to show the relationship between other parts of speech, and we considered that teaching its meaning alone is pointless and confusing for the learners/players; thus, we included the first five parts of speech only.

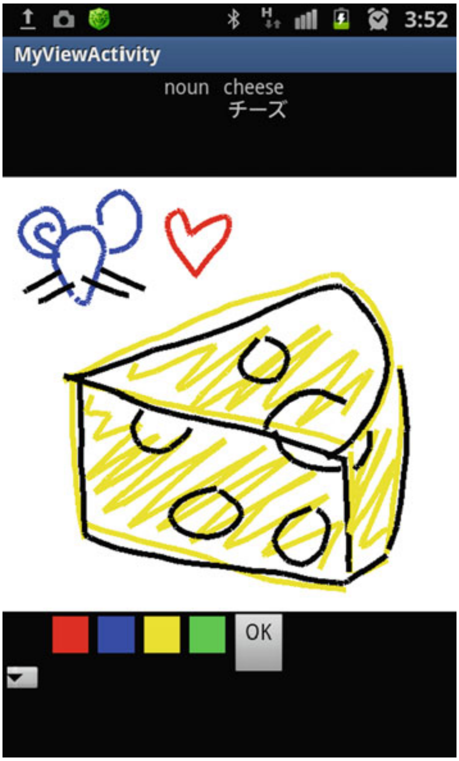


Fig. 3 Draw-part2

Thirdly, Japanese translations are shown to improve the understanding of the vocabularies. These Japanese translations are, not shown in the first time, and they are shown by pressing Hint Button as shown in Fig. 1. If the player cannot draw any of the words provided, they can shuffle the words to get another set; an important thing is that they cannot shuffle the words until they check the Japanese translation, so there would be more chance to see English and Japanese together to learn the vocabularies. Once the player starts to draw with a hint Japanese meaning is provided as shown in Fig. 3, if he/she starts without with a hint, then there would be no translation shown.

Evaluation

After the preliminary survey, examinees were asked to play both Draw-part and Answer-part of Vocab Draw once for each. First of all, 82 % of the examinees responded with positive feedback for the question whether he/she wants to play Vocab Draw again. The specific numbers are as follows: 6 % are strongly willing to play again, 35 % are willing to play again, 41 % can play again, 18 % are not

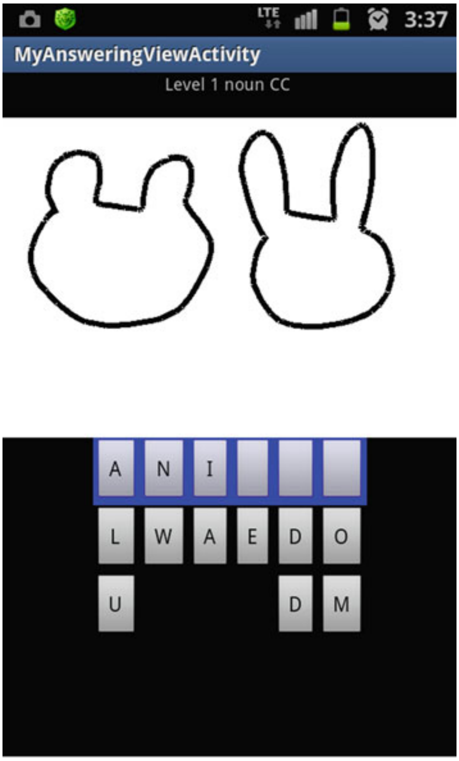


Fig. 4 Answer-part2

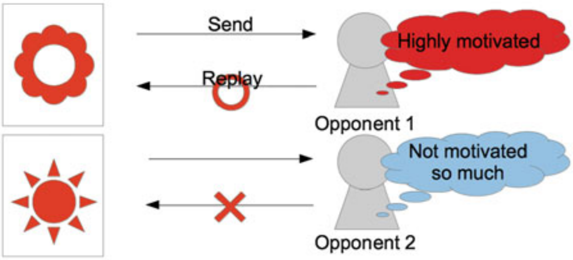
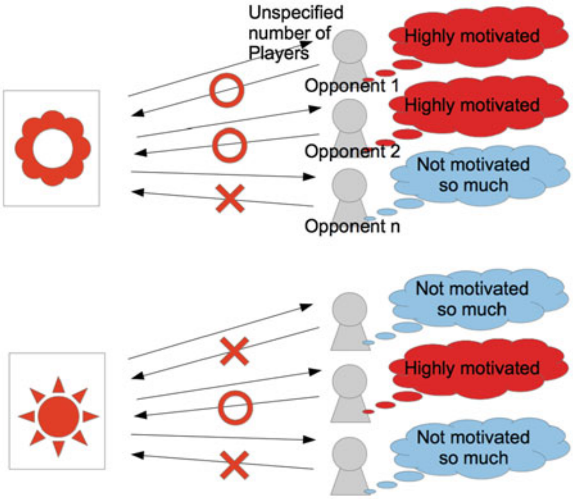


Fig. 5 One-to-one (draw something)

willing to play again. In addition, paying a close attention to the ones who gave negative posture for learning English in the preliminary survey, it is clear that their attitude changed via the usage of Vocab Draw. 72.7 % of those who are not willing to put effort into learning english, 78.6 % of those who hate learning vocabularies, and also 72.7 % of those who are not doing any extra work for learning vocabularies have improved their motivation towards English learning if it is with Vocab Draw. Asking why they want to play it again, most of the answers were because of the amusingness of the game, shown in the Fig. 7. By this result, it is proven that the amusingness would improve the attitude of learners, even if they hate learning

Fig. 6 One-to-many (vocab draw)



it. However, half of the positive feedback are as weak as that they CAN play the Vocab Draw again, and only one player answered that he strongly wants to play it again. From this it can be said that motivations can be improved with the amusingness as much as they do not leave the game, but it is not enough to lead them to strong attitude or willingness to play. Our future direction is to change the examinees' attitude from that they CAN play it again to they WANT to play it again.

We also asked the examinees to feedback freely with pros and cons of the application. For educational elements, some opinions in Table 1 is provided by the examinees. Most of the positive opinions were about the usage of drawing. Four examinees answered that it is easy to see the vocabulary as an image, thus it is helpful to learn it better. As another opinion, one examinee considered that Vocab Draw is more like an usual game rather than a boring education game. This opinion is important because it was our final goal.

On the other hand, some issues became distinct. Our biggest issue is that some of the words are difficult to draw in a picture. Especially the words which are not noun, it is difficult for the players to associate the word with an image. If many

Fig. 7 Features of vocab draw motivate the examinees to play again (multiple answers)

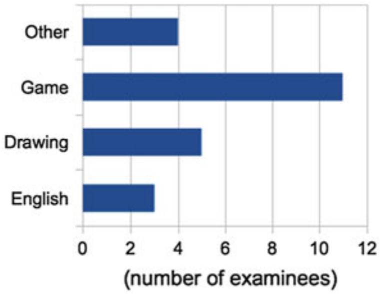


Table 1 Feedback of vocab draw

Number of people	Feedback
Positive feedbacks	
4	Easy to see the vocabulary as an image
3	Helpful to learning vocabularies
1	Does not give the feeling of learning
1	More chance to play with English
1	Might be memorable since playing while guessing
Negative feedbacks	
4	Many words are difficult to draw
1	Words should be used in text, rather than just translated alone
1	Learning efficiency is low, since long time is taken to draw
1	It is more like a testing rather than learning

people cannot draw the word except noun, the number of those words would be reduced in the answering part as well, resulting less opportunity to experience the word. Thus, in our experiment, almost everyone chose a noun to draw. If we keep using these four parts of speech, we must re-consider the easier way to draw, or the sets of the vocabularies.

Conclusion and Future Work

As the conclusion, we have succeeded to encourage the examinees to learn English vocabularies with our novel game, Vocab Draw. Their motivation to play it again was as high as 82 %, however, not all of them were strongly passionate to do so.

In order to change the players attitude from passive to active motivations, introduction of the Bartle's personality classification [4] should be considered. Bartle's personality classification is probably the most famous classification of game player personality, originally applied to MMORPG players. The game players are divided into four personalities of achievers, explorers, socializers, and killers. Each personality has different aim to feel joy in a game, and he plays the game to accomplish his aim. A list of their basic aims are provided in the Table 2. One goal might be a perfect fit for one type, but occasionally it wouldn't be an amenity to another type at all. Great games tend to support all of these goals shown in Table 2, so it could attract any game players. It is proven by many top selling

Table 2 Bartle's personality classification

Player type	Aim to motivate the player
Achiever	Achieving
Explorer	Curiosity
Socializer	Co-operating
Killer	Competition

Table 3 Possible new functions for vocab draw

Player types	Functions to implement	Aiming goals
Achiever	Gallery	Collect all the words
Explorer	Stages	Open up every stages
Socializer	Team play	Co-operate with others
Killer	Ranking	Becoming a top player in the ranking

games that this classification is useful, but it is not proven that it would work to educational games as well yet. Therefore, in our future work, we would like to validate its usefulness in an educational game by adding some new features listed in Table 3.

References

1. The Annual Video Game Industry Report: Media Create (2007–2012)
2. McGonigal J (2011) Reality is broken: why games make us better and how they can change the world. Penguin Books, City of Westminster
3. Kamiya N, Hasegawa N, Hasebe N, Machida N (2011) Part of speech ratio and kinds of verbs in “english note”: scientific approaches to language 1(9):233–258
4. Bartle RA (1996) Hearts, clubs, diamonds, spades: players who suit muds, http://www.mud.co.uk/richard/hcds.htm

An Improved Method for Measurement of Gross National Happiness Using Social Network Services

Dongsheng Wang, Abdelilah Khiati, Jongsoo Sohn, Bok-Gyu Joo and In-Jeong Chung

Abstract Studies on the measurement of happiness have been utilized in a variety of areas; in particular, it has played an important role in the measurement of society stability. As the number of users of Social Network Services (SNSs) increase, efforts are being made to measure human well-being by analyzing user messages in SNSs. Most previous works mainly counted positive and negative words; they did not consider the grammar and emotion. In this paper, we reorganize the mechanism to harness the advantages of (a) Part-Of-Speech (POS) tagging for grammatical analysis, and (b) the SentiWordNet lexicon for the assignment of sentiment scores for emotion degree. We suggest a modified formula for calculating the Gross National Happiness (GNH). To verify the method, we gather a real-world dataset from 405,700 Twitter users, measure the GNH, and compare it with the Gallup well-being release. We demonstrate that the method has more precise computation ability for GNH.

Keywords Social network service (SNS) · Happiness · Well-being · Gross national happiness (GNH)

D. Wang (✉) · A. Khiati · I.-J. Chung
Department of Computer and Information Science, Korea University, Seoul, Korea
e-mail: dswang2011@korea.ac.kr

A. Khiati
e-mail: le_zakkaz@korea.ac.kr

I.-J. Chung
e-mail: chung@korea.ac.kr

J. Sohn
Service Strategy Team, Visual Display, Samsung Electronics, Seoul, Korea
e-mail: jongsoo.sohn@samsung.com

B.-G. Joo
Department of Computer and Information Communications, Hong-Ik University,
Seoul, Korea
e-mail: bkjoo@hongik.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_3, © Springer Science+Business Media Dordrecht 2014

Introduction

Gross National Happiness (GNH) was initially proposed in 1972 by the former King of Bhutan, who wanted to promote the development of people's mental health and happiness according to Bhutan's Buddhist culture [1]. Recently, GNH is utilized in a variety of areas as an evaluative indicator of the measure of happiness in a nation and its regions. Moreover, it plays a predominant role in measuring personal happiness, according to a diversity of studies [2]. Gallup, a representative organization to measure GNH, annually measures happiness for each nation by conducting survey tracking [3].

Evaluating the GNH based on Social Network Services (SNSs) was attempted initially by Kramer [4]. He assumed and proved that the more positive words people used in their updates, the happier they were, and vice versa. Thus, he used LIWC2007 corpus [5] to classify positive and negative words and count them respectively. Additionally, he proposed a formula to normalize the happiness scores, and applied this to his research. Subsequently, various studies were conducted using Kramer's method. However, Kramer's method has two problems.

- Lack of a Part-Of-Speech (POS) analysis: It is hard to distinguish the sentiment polarity. Considering the statements "I like my family" and "She looks like her sister," the term 'like' in the former statement conveys a positive sentiment, whereas its use in the latter one conveys an objective description.
- Lack of sentiment degree analysis: Words that convey emotions may express varying degrees of a particular sentiment. For example, although both 'bad' and 'terrible' are negative words, they convey different degrees of the quality, i.e., 'terrible' conveys a greater degree of negativity than 'bad'.

In order to overcome these barriers, we make use of a POS [6] tagger to grammatically analyze a user's message. This enables us to clarify the meaning of each word and then classify positive and negative words clearly. For example, 'like' is commonly regarded as a positive word; however, in the sentence "she looks like somebody," it is an objective description without any sentimental meaning, which can be identified if the POS of words is provided. Further, POS can be applied to analyze negative phrases such as 'not good' or 'not bad.' Secondly, we employ the *SentiWordNet* [7] lexicon to assign a sentiment scores (a real number from −1 to 1) to each emotion word. These sentiment scores can be used not only in the classification of positive and negative words but also in the utilization of the degree of sentiment, thus overcoming the drawbacks of Kramer's approach. Finally, we suggest the improved formula to measure GNH based on Kramer's method and the sentiment score.

In order to verify the proposed method, we establish Gallup's survey tracking as ground truth, and then compare the method to the results of the *Word Count* method using the same datasets for eight cities in the United States. We show that the suggested method is comparable to that of Gallup and better than the previous *Word Count* methods.

Related Works

The progress of health care in Bhutan was reported in [8], which provided a quantitative meaning to the measurement of GNH for the government. The authors of [9] emphasize the conception GNH and propose a GNH framework to assess the health impact where the health, environmental, and economic impact can be estimated collaboratively.

Survey tracking has been a predominant method to measure happiness and the current survey method employs a self-report methodology [3], which has been found to be reasonably accurate in measuring the well-being of an individual [2]. Gallup [3], has studied human nature and behavior for more than 75 years. He released a well-being index based on a daily tracking survey of 1,000 American adults.

Winton performed a comprehensive investigation from different views of measurement and concluded that no measurement is definitely perfect and the best approach is to use a range of measures [1]. SNS-based measurement is evolving to be an innovative and significant approach for measuring GNH. Some researchers employ the *Word Count* strategy to measure GNH, which extracts words from user's messages in SNSs and then classifies positive and negative words by using the LIWC2007 corpus (as depicted in Fig. 1). For instance, Facebook is used by Kramer to track the well-being of the population in the U.S. that uses Facebook [4]. He assumed that the more positive words people use in their status updates, the happier they are, and vice versa.

For example, the user's message "It's a happy day." gets a positive score of 1/4, and a negative score of 0/4. Not only Facebook, the biggest social media, but also Twitter, the second biggest one, was used to measure happiness. Daniele [10] attempted to measure the Gross Community Happiness from Tweets using a method similar to Kramer's, and validated the feasibility of using Twitter even though Twitter has many of its own unique characteristics. However, the *Word Count* method is more likely to result in inaccurate and incorrect analyses since it is difficult to represent the degree of emotion conveyed using each word and infeasible to analyze the POS of each word.

Various studies were conducted to derive insights using national happiness indices. In [11], an evaluation was made based on 336,802 unique users and 12,781,243 Tweets to explore and visualize topics in Twitter. The drawback of the method is that it analyzes only the frequency of words without considering the POS and the relationship between words.

Fig. 1 *Word count* method

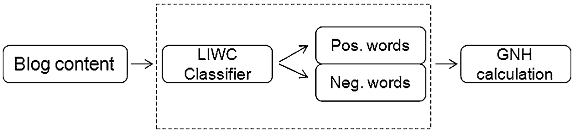
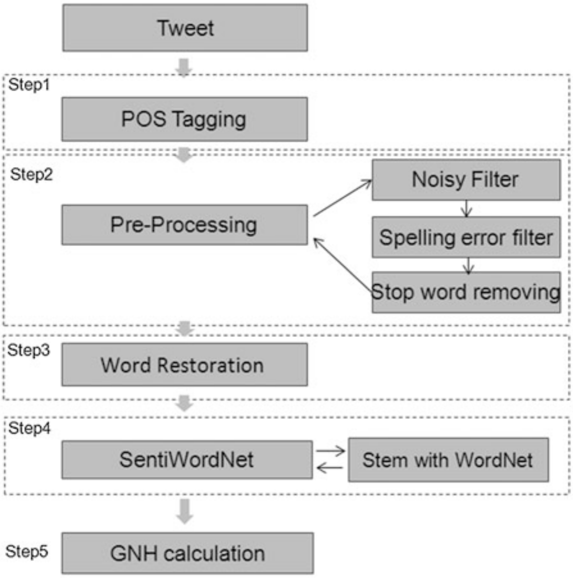


Fig. 2 Overview of the suggested method for GNH measurement



Improved Method for GNH Measurement

In this paper, we proposed a modified version of the GNH measurement method based on the frequency of words in SNS messages. We enhance the measurement method of the GNH by adding POS tagging and utilizing the *SentiWordNet* lexicon. The suggested method evaluates the sentiment scores, as shown in Fig. 2.

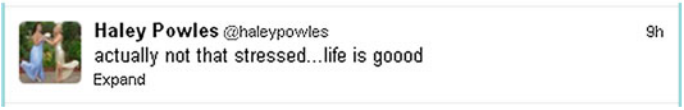
The proposed method consists of five steps: (1) POS tagging, (2) preprocessing, (3) restoration of the original form of each word, (4) assignment of a sentiment score for each word, and (5) computation of GNH based on the suggested formula.

Step 1: POS Tagging. POS tagging divides each sentence into words and restores the original form of each word using *WordNet*. It recognizes the POS, such as noun, adjective, and verb, and tags words respectively. Thus, it reduces the error rate of the classification of positive and negative words and increased the ease in processing synonyms.

For the case depicted in Fig. 3, the result of the tagging is:

$\{(\mathit{actually}\colon R), (\mathit{not}\colon R), (\mathit{that}\colon D), (\mathit{stressed}\colon A), (\ldots\colon,), (\mathit{life}\colon N), (\mathit{is}\colon V), (\mathit{goood}\colon A)\}$

R denotes an adverb; *D*, determiner; *A*, adjective; ‘,’, punctuation; *N*, common noun; and *V*, verb [6].



Haley Powles @haleypowles 9h
actually not that stressed...life is goood
Expand

Fig. 3 Example of a tweet

Step 2: Preprocessing. The preprocessing step encompasses (1) noise filtering, (2) spelling error filtering, and (3) stop word removal. Noise filtering removes symbols such as URLs, @, RTs, emoticons, and punctuations. Words that are often skipped and filtered out by search machines, such as ‘a,’ ‘is,’ and ‘the,’ are removed directly.

For the case depicted in Fig. 3, ‘…’ is filtered according to sub-step (1). ‘that’ and ‘is’ are also filtered during sub-step (3).

Step 3: Word Restoration. This step restores the original, condensed form of each word associated with subjectivity and sentiment, which may have been intentionally lengthened, such as ‘Goooood’ and ‘Coooooool.’

For the case depicted in Fig. 3, ‘goood’ is condensed back to ‘good’ so that it can retrieve a sentiment score from *SentiWordNet*.

Step 4: Sentiment Score Assignment. *SentiWordNet*, a lexical dictionary developed based on *WordNet* [12], describes terms and their semantic relationships. It consists of approximately 117,660 synsets (207,000 word-sense pairs). Each term in its synset is described using a triple of positive, negative, and neutral (objective) sentiment scores. For example, {pos. = 0.35, neg. = 0.65, neutral = 0}, which sums up to 1 (i.e., pos. + neg. + neutral = 1).

WordNet and *SentiWordNet* are applied to represent the sentiment score for each word as a real number between −1 and 1. In addition, the sentiment score of an emotion word that belongs to negative sentence (appears after ‘not,’ ‘never,’ ‘none,’ ‘no,’ ‘hardly,’ ‘seldom,’ etc.) is reversed.

For the case depicted in Fig. 3, after the three subsequent steps are completed, ‘actually,’ ‘not,’ ‘stressed,’ ‘life,’ and ‘good’ remain. The adverb ‘not’ expresses a negation, which reverses the sentiment score of the emotion word after it. The sentiment score of ‘stressed,’ which is −0.34, is reversed as +0.34. The adjective ‘good’ gets a sentiment score of 0.48 and the noun ‘life’ gets a 0.0.

Step 5: Formula Calculation. Formulae (1) and (2) are defined to calculate the sentiment score of a message. p_i is the positive score of the message, and n_i is the negative score of the message. N indicates the number of positive words in formula (1), and the number of negative words in formula (2); $score_w$ is the *SentiWordNet* score of each word.

$$p_i = \sum_{p_w=1}^{N} \frac{score_w}{\text{Number of total words}} \tag{1}$$

$$n_i = \sum_{n_w=1}^{N} \frac{score_w}{\text{Number of total words}} \tag{2}$$

For example, consider the Tweet “the weather is good, but I am sad.” The preprocessing result is {0, 0, 0, 0.478, 0, 0, 0, −0.416} as the scores of ‘good’ and ‘sad’ are 0.478 and −0.416 respectively. By applying formulae (1) and (2), the positive score p_i is 0.478/8, and the negative score n_i is −0.416/8 respectively. For the case depicted in Fig. 3, $p_i = \frac{0.87}{7} = 0.11$, $n_i = \frac{0.0}{7} = 0$.

$$H_i = \frac{p_i - \mu_p}{\sigma_p} - \frac{n_i - \mu_n}{\sigma_n} \tag{3}$$

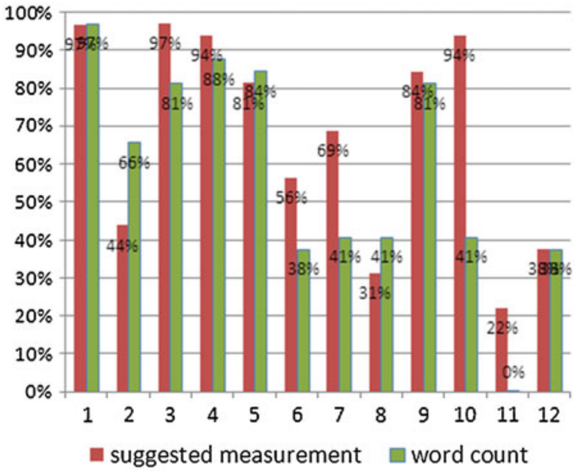
Further, we put this score into the formula (3) [4] in order to normalize the happiness score (in practice, for the control of scale, we aggregate all Tweets of a user daily as ‘one Tweet’). In the formula (3), H_i is the happiness score for user i at a specific time, where p_i and n_i represent the positive and negative sentiment scores; μ_p and μ_n represent the average positive and negative sentiment scores across all users during a specific time; and $\sigma_p(\sigma_n)$ are the corresponding standard deviations. As described above, we compute the happiness score for each user in a particular SNS using the suggested steps and formulae.

Performance Study

For the validation and evaluation of the suggested method, we implemented a system and conducted an experiment. In the experiment, with the aid of Twitter Application Programming Interface (API), we crawled 2,072,329 user profiles across 8 cities in the United States. In order to guarantee that the users are active, we refined the dataset with some strategies, and obtained 405,700 user profiles as the experimental dataset. We collected up to 300 Tweets created by each user in 2011. As a result, we acquired approximately 60 million Tweets from 405,700 users.

We applied the *Word Count* method and the suggested method to the same dataset for the measurement of the GNH results and compared the two results. Given that the Gallup is ground truth, Fig. 4 illustrates the accuracy rates of the *Word Count* method and our suggested measurement with respect to Gallup. The x-coordinate represents the 12 months in 2011, and the y-coordinate is the degree

Fig. 4 Accuracy distribution in each month



of proximity to Gallup. Except for February, May, and August, we find a stronger correlation with Gallup during all other months, and the overall improvement is by approximately 10 %.

Compared to the survey tracking method, SNS-based GNH measurement is able to deal with a huge amount of data while consuming a smaller amount of time. Additionally, the suggested measurement is more accurate and credible than the existing *Word Count* method.

Conclusion

GNH measurement based on data in SNSs has been studied recently. However, previous methods consider only the polarity during classification leading to ambiguous and unreliable results. Accordingly, we consider the emotional degree in the computation of GNH and reorganize the mechanism to be more accurate. First, we employ (a) POS tagging and (b) emotion degree assignment and detection of dual affirmation and negation (score reversing) to obtain a more precise sentiment score for each word. We then validate the proposed method in the experiment by providing a comparison with the *Word Count* method. The experiment illustrated that the suggested method shows better results over existing methods and demonstrated the feasibility of considering the degree of emotion words.

In comparison with the questionnaire-based survey tracking, the suggested SNS-based method for the measurement of GNH was able to reduce the time and cost drastically. Moreover, the method facilitates the establishment of policies and strategies for customer marketing in enterprises and societies based on the computed happiness scores.

In future, based on the experimental results, the proposed method can be applied to analyze the relationship between the happiness score and other mass media such as online news and articles with the aid of semantic web technology.

References

1. Bates W (2009) Gross national happiness. Asian-Pac Econ Lit 23:1–16
2. Diener E, Diener M, Diener C (1995) Factors predicting the subjective well-being of nations. J Pers Soc Psychol 69:851–864
3. Walker SS, Schimmack U (2008) Validity of a happiness implicit association test as a measure of subjective well-being. J Res Pers 42:490–497
4. Kramer ADI (2010) An unobtrusive behavioral model of "gross national happiness." In: 28th international conference on human factors in computing systems, ACM, Atlanta, Georgia, USA, pp 287–290
5. James W, Pennebaker CKC, Ireland M, Gonzales A, Booth RJ (2007) The development and psychometric properties of LIWC2007. LIWC.Net, Austin, TX

6. Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, Heilman M, Yogatama D, Flanigan J, Smith NA (2011) Part-of-speech tagging for Twitter: annotation, features, and experiments. In: 49th annual meeting of the association for computational linguistics: human language technologies: short papers. vol 2. Association for Computational Linguistics, Portland, Oregon, pp 42–47
7. Sebastiani AEAF (2006) SentiWordNet: a publicly available lexical resource for opinion mining. In: Language resources and evaluation (LREC), pp 417–422
8. Tobgay T, Dorji T, Pelzom D, Gibbons RV (2011) Progress and delivery of health care in Bhutan, the land of the thunder dragon and gross National happiness. Trop Med Int Health 16:731–736
9. Pennock M, Ura K (2011) Gross National happiness as a framework for health impact assessment. Environ Impact Asses 31:61–65
10. Quercia D, Ellis J, Capra L, Crowcroft J (2012) Tracking "gross community happiness" from Tweets. In: ACM 2012 conference on computer supported cooperative work. ACM, Seattle, Washington, USA, pp 965–968
11. Brew A, Greene D, Archambault D, Cunningham P (2011) Deriving insights from National happiness indices. In: 2011 IEEE 11th international conference on data mining workshops. IEEE Computer Society, pp 53–60
12. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38:39–41

Advanced Comb Filtering for Robust Speech Recognition

Jeong-Sik Park

Abstract This paper proposes a speech enhancement scheme that leads to significant improvements in recognition performance when used in the Automatic Speech Recognition (ASR) front-end. The proposed approach is based upon adaptive comb filtering. While adaptive comb filtering reduces noise components remarkably, it is rarely effective in reducing non-stationary noises due to its uniformly distributed frequency response. This paper proposes an advanced comb filtering technique that adjusts its spectral magnitude to the original speech, based on the gain modification function, an Minimum Mean Squared Error (MMSE) estimator.

Keywords Advanced comb filtering · Gain modification function · Minimum mean squared error · Robust speech recognition

Introduction

Considerable efforts have been directed toward reducing various kinds of noises including additive and channel noise, with the goal of improving speech quality and intelligibility. Although various algorithms for speech enhancement have been developed over the last several decades, ASR systems in speech-driven interfaces such as mobile devices require further effective reduction of non-stationary noises.

The adaptive comb filtering firstly presented by Frazier enhances noisy speech with the pitch period of a voiced sound [1]. Although adaptive comb filtering has

J.-S. Park (✉)
Department of Intelligent Robot Engineering, Mokwon University, Daejeon, South Korea
e-mail: parkjs@mokwon.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_4, © Springer Science+Business Media Dordrecht 2014

been applied to noise reduction with low complexity and high capacity [2, 3], it is vulnerable to non-stationary noises due to uniformly distributed frequency response of comb-filter as well as the difficulty in estimating the accurate pitch period. This paper introduces some problems of conventional adaptive comb filtering and proposes an improved comb filtering process employing an Minimum Mean Squared Error (MMSE) estimator.

Section Advanced comb filtering explains the principles of adaptive comb filtering and the modified comb filtering. And, sections Estimation of speech absence probability and Improvement of comb-filter by the gain modification function present experimental results and conclusions, respectively.

Advanced Comb Filtering

The comb filtering is based on the observation that wave-forms of voiced speech are periodic, corresponding to the fundamental frequency [1]. As noise components generally exist between the harmonics of speech spectrum, they are reduced by a comb-filter, which passes only the harmonics. The basic operation of a comb-filter can be explained by considering its impulse response as:

$$h(n) = \sum_{k=-L}^{L} \alpha_k \times \delta(n - T) \tag{1}$$

where $\delta(n)$ is an unit impulse function, T corresponds to the pitch period, and the length of the filter is $2L + 1$. α_k is the filter coefficient that is associated with the intelligibility of the filtered speech, satisfying $\sum_{k=-L}^{L} \alpha_k = 1$.

A comb-filter can significantly reduce noise components existing between the harmonics while preserving the speech sounds. Nevertheless, comb filtering occasionally results in poor performance in certain cases. First, the filter may degrade the quality of clean speech, as regularly repeated frequency response can distort the speech signal where the fundamental frequency is changed continuously. Second, severe noise may also cause distortion due to inaccurate pitch estimation.

This study proposes an approach to compensate for the distortions, using the Gain Modification Function (GMF), an MMSE estimator. The GMF denotes the probability of speech-presence in the frequency bin and consists of the Speech Absence Probability (SAP), priori SNR, and posteriori SNR [4]. A frequency bin with low SAP and high SNR represents the speech region rather than the non-speech region, indicating a high value of GMF. The key issue in this work is adjusting the frequency response of the comb-filter to match the local characteristics of the clean speech spectrum on the basis of GMF estimated in each frequency bin.

Estimation of Speech Absence Probability

First, we estimate the SAP given by [5]:

$$q_l(\omega) = \alpha \cdot q_{l-1}(\omega) + (1 - \alpha) \cdot I_l(\omega) \tag{2}$$

which denotes the SAP of l-th frame in the frequency bin ω. $I_l(\omega)$ is a hard-decision parameter that determines whether speech is present in the frequency bin. The decision is made from the posteriori SNR (γ_ω^l) and a threshold (γ_{TH}) as:

$$I_l(\omega) = \begin{cases} 0 \ (\gamma_\omega^l \geq \gamma_{TH}) \\ 1 \ (\gamma_\omega^l < \gamma_{TH}) \end{cases} \tag{3}$$

The frequency bin, where $I_l(\omega)$ equals 0, refers to the speech-present region. In the speech-absent region, $q_l(\omega)$ is close to 1, thus giving high SAP. In [5], the parameters γ_{TH} and α are set as constants (0.8 and 0.95, respectively) based on informal listening tests. As such, the fixed value of γ_{TH} may mislead (3), thus resulting in an incorrect decision of a speech-present/absent region, particularly for heavily damaged speech. We update the parameter continually, considering the posteriori SNR changed per frame and frequency bin, as shown in:

$$\gamma_{TH}^l(\omega) = \sum_{k=l-\beta}^{l-1} \frac{\gamma_\omega^k}{\beta} \tag{4}$$

The average of γ_ω estimated in the previous β frames is expected to advance the hard-decision and thus improve the estimate of SAP applied to GMF. Our experiments determined that the smallest number of β is 5 while preserving the performance. And we confirmed that α doesn't influence the performance a lot and used a fixed value of 0.95.

Improvement of Comb-Filter by the Gain Modification Function

The proposed approach is based on a proposition that the GMF can adjust the frequency response of the comb-filter to the spectrum of the original speech, depending on the speech absence probability for each frequency bin. The adaptive comb filtering modified by the MMSE estimator is summarized as:

$$\widehat{A}_l(\omega) = G_l(\omega) \times A_l(\omega) \tag{5}$$

$A_l(\omega)$ is the spectral energy of the adaptive comb-filter in the frequency bin ω for l-th frame and $G_l(\omega)$ is the gain modification function estimated by [5]:

$$G_l(\omega) = \frac{\Lambda_l(\omega)}{1 + \Lambda_l(\omega)} \approx 1 - q_l(\omega) \tag{6}$$

where $\Lambda_l(\omega)$ is a likelihood ratio calculated from the SAP and $G_l(\omega)$ is close to 1 in the speech present region. According to (5), $G_l(\omega)$ degrades the spectral energy of the comb-filter in the frequency bins corresponding to the speech-absent regions, where $G_l(\omega)$ goes to 0. This modification gives further correct estimate for the frequency response of comb-filter.

It should be noted that the direct use of $G_l(\omega)$ ranging from 0 to 1 may degrade the spectral energy over all of the frequency bins including the speech region. To prevent signal distortion induced by this property, we modify $G_l(\omega)$ based on the SAP estimator as shown in (7).

$$\overset{\frown}{G}_l(\omega) = \{1 - (q_l(\omega) - \varepsilon)\} \times G_l(\omega) \tag{7}$$

where $q_l(\omega)$ gives the GMF a limited range from 0 to $(1 + \varepsilon)$.

By applying $\overset{\frown}{G}_l(\omega)$ instead of $G_l(\omega)$ in (5), the modified GMF degrades the spectral energy of the comb-filter over the region where $q_l(\omega)$ is higher than ε. On the other hand, it retains or emphasizes the energy over the region where $q_l(\omega)$ is equal to or lower than ε. By experiments, we concluded that $\varepsilon = 0.35$ provides the best performance. Based on the modified GMF, the comb-filter is expected to further remove the spectral noise components, eliminating signal distortions.

Experiments and Results

Experimental Environments

We performed the ASR experiments on the Aurora 2 database, following the procedures for training Hidden Markov Model (HMM) under the clean condition and recognizing data by using the HTK software tools [6, 7]. The training and testing set consist of 8,440 and 4,004 digit strings, respectively. In addition, the test data is divided into three sets according to noise types and channel condition. We obtained 12 Mel Frequency Cepstral Coefficients (MFCCs) and a log energy with their first and second derivatives from all training and testing data. Each word was modeled by a simple left-to-right 16-state three-mixture whole word HMM. A three-state six-mixture silence model and a one-state six-mixture pause model were also used.

Recognition Results

We performed two kinds of experiments to compare the performance of our proposed technique to several comb filtering techniques introduced previously.

Table 1 Comparison of WERs (%) for the speech processed by several adaptive comb filtering

	Data set			Avg.	WER reduction
	Set A	Set B	Set C		
Original	44.76	50.35	39.22	44.78	–
BACF	42.14	48.56	37.10	42.60	4.86
FACF	41.42	46.76	37.26	41.85	6.54
MACF	38.82	43.38	35.41	39.21	12.45

Table 2 Comparison of WERs (%) over clean and high SNRs

	SNR(dB)			Avg.	WER reduction
	clean	20 dB	15 dB		
Original	1.68	4.83	17.07	7.86	–
BACF	4.05	6.06	13.94	8.02	−1.99
FACF	3.54	5.98	13.70	7.74	1.53
MACF	2.21	5.25	11.00	6.15	21.71

Table 1 shows results for the first experiment. We performed recognition experiments on four kinds of speech data, one of which is the raw data itself ("original") and the others were processed by the basic adaptive comb-filter ("BACF"), the fully adaptive comb-filter ("FACF") and the modified comb-filter ("MACF"), respectively. FACF was known as the representative method for enhancing the basic comb-filter, by improving the fixed filter coefficients [3]. We report the performance as the Word Error Rate (WER) averaged over Signal-to-Noise Ratios (SNRs) from 0 to 20 dB. In all test sets, the comb-filtered data outperforms the raw speech data due to noise reduction. Compared with the original speech, three kinds of com-filters lead to WER reduction of 4.86, 6.54, and 12.45 %, respectively. MACF shows a significant improvement, 8 and 6.3 % WER reduction, over BACF and FACF, respectively.

Although the above results verify that the proposed approach improves BACF and FACF successfully, more reliable results are given in Table 2, where BACF has even higher WER than Original or MACF over clean and high SNRs (15–20 dB). This poor accuracy results from the distortions of clean speech due to comb filtering. By the way, MACF leads to remarkably 45.4 and 37.6 % WER reduction compared with BACF and FACF, respectively, in clean environment. This result demonstrates that MMSE estimator-based comb filtering guards against signal distortions significantly while previous comb filtering techniques don't prevent from the distortions.

Conclusion

We proposed a speech enhancement scheme that combines MMSE estimator and conventional adaptive comb filtering to protect against signal distortions caused by comb filtering. The proposed method adjusts the spectral magnitude of the comb-filter to the spectrum of the clean speech, based on gain modification function. This approach degrades the spectral energy of comb-filter in the frequency region where no speech components exist while retaining or emphasizing the energy of comb-filter in the speech-present region. We applied our speech enhancement scheme to the ASR front-end based on Aurora 2 database. Recognition results showed that the modified comb filtering yields superior performance compared to no-processing (baseline) and other comb filtering techniques, presenting WER reduction of 12.5 % to baseline. The results also confirmed that our approach notably improves the distortions of clean speech, which is the most serious weakness of comb filtering. Further work will be aimed at verifying our approach with experiments on another data sets and applying to ASR system in speech-driven interfaces.

Acknowledgments This work was supported by the NAP (National Agenda Project) of the Korea Research Council of Fundamental Science and Technology.

References

1. Frazier RH, Samsam S (1976) Enhancement of speech by adaptive filtering. In: Proceedings of IEEE International Conference on ASSP, Philadelphia. PA, pp 251–253
2. Nehorai A, Porat B (1986) Adaptive comb filtering for harmonic signal enhancement. IEEE Trans Acoust Speech Signal Process 34(5):1124–1138
3. Veeneman DE, Mazor B (1989) A fully adaptive comb filter for enhancing block-coded speech. IEEE Trans Acoust Speech Signal Process 37(6):955–957
4. Ephraim Y, Malah D (1985) Speech Enhancement using a minimum mean-square error log spectral amplitude estimator. IEEE Trans Acoust Speech Sig Process 33(2):443–445
5. Malah D, Cox RV, Accardi AJ (1999) Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments. In: Proceedings of ICASSP, pp 201–204
6. Hirsch HG, Pearce D (2000) The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. Proc ICSLP 4:29–32
7. Young S, Evermmana G, Gales M, Hain T et al (2006) The HTK Book for HTK Version 3.4., Cambridge University Engineering Department,Cambridge

Fall Detection by a SVM-Based Cloud System with Motion Sensors

Chien-Hui (Christina) Liao, Kuan-Wei Lee, Ting-Hua Chen, Che-Chen Chang and Charles H.-P. Wen

Abstract Recently, fall detection has become a popular research topic to take care of the increasing aging population. Many previous works used cameras, accelerometers and gyroscopes as sensor devices to collect motion data of human beings and then to distinguish falls from other normal behaviors of human beings. However, these techniques encountered some challenges such as privacy, accuracy, convenience and data-processing time. In this paper, a motion sensor which can compress motion data into skeleton points effectively meanwhile providing privacy and convenience are chosen as the sensor devices for detecting falls. Furthermore, to achieve high accuracy of fall detection, support vector machine (SVM) is employed in the proposed cloud system. Experimental results show that, under the best setting, the accuracy of our fall-detection SVM model can be greater than 99.90 %. In addition, the detection time of falls only takes less than 10^{-3} s. Therefore, the proposed SVM-based cloud system with motion sensors successfully enables fall detection at real time with high accuracy.

Keywords Fall detection · Kinect · Support vector machine (SVM) · Cloud computing

C.-H. (Christina) Liao (✉) · K.-W. Lee · T.-H. Chen · C.-C. Chang · C. H.-P. Wen
Department of Electrical and Computer Engineering, National Chiao Tung University,
1001 University Road, Hsinchu, Taiwan, Republic of China
e-mail: liangel.cm97g@g2.nctu.edu.tw

K.-W. Lee
e-mail: behigheveryday.eed99@nctu.edu.tw

T.-H. Chen
e-mail: tinghua000@gmail.com

C.-C. Chang
e-mail: boy76229@yahoo.com.tw

C. H.-P. Wen
e-mail: opwen@g2.nctu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_5, © Springer Science+Business Media Dordrecht 2014

Introduction

Nowadays, various health-care systems have been rapidly developed to take care of the increasing aging population. Particularly, since falls are dangerous and even fatal to the elder, fall detection [1–4] has became an important topic for elder healthcare. Most previous works used accelerometers, gyroscopes and cameras to distinguish falls from normal behaviors. However, these techniques encountered some challenges from privacy, accuracy, convenience and runtime to its practicality.

Many of previous works [5–9] used tri-axial accelerometers or gyroscopes to detect fall accidents. However, these wearable accelerometers and gyroscopes are inconvenient to users and also hard to accurately differentiate falls from normal behaviors such as sitting down, lying and hunkering. On the other hand, using cameras [10–12] to detect falls can achieve higher accuracy than using accelerometers and gyroscopes. However, cameras will collect massive data, either useful or useless. Thus, heavy image-processing is required, which makes it hard to be applied for many real-time applications.

Recently, cloud applications for human-centered computing (HCC) arise rapidly. Compared to traditional approaches [1–4], a cloud system intends to handle only useful data for efficiency. Figure 1 shows an examples for illustration. However, a cloud system for HCC still comes with three problems: *big data*, *massive communication* and *heavy computation*. To overcome these three problems to enable a real-time system for highly-accurate and highly-private fall detection, this work uses *motion sensors* (i.e. Microsoft Kinects [13]) to collect data from human beings and applies *support vector machine* (SVM) models [14] in a cloud system for fall detection. Motion sensor devices like Microsoft Kinect have many advantages and are outlined as below:

1. An image of a human being can be transformed into skeleton points, providing better privacy. So our fall-detection systems can even be installed in the bathroom where accidents often occur for the elder.
2. Motion data from multiple users can be captured simultaneously.
3. Data size of skeleton points is much smaller than that of the image.

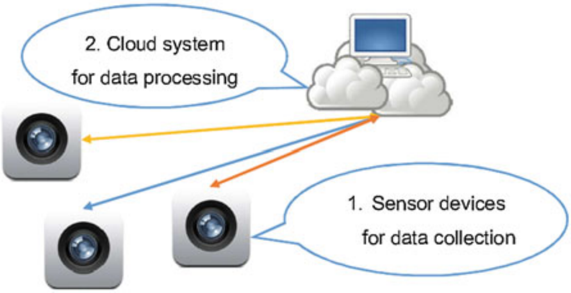


Fig. 1 Cloud systems with sensor devices

Developing a SVM model to determine human behaviors on the cloud system with motion sensors can be widely used at home, hospital, kindergartens, and etc. Our experimental results show that using Kinect as a sensor device, our SVM-based cloud system can reach greater than 99.90 % accuracy under the best setting for fall detection. In addition, using the linear kernel or polynomial kernel for the SVM learning, the response time (for prediction) takes less than 10^{-3} s on average, making such system suitable for real-time applications.

The rest of this paper is organized as follows. Section SVM-based Cloud Systems with Motion Sensors for Fall Detection describes the architecture of our SVM-based cloud system with motion sensors. In this section, Microsoft Kinect is introduced and the proposed SVM model for fall detection is also elaborated. In section Experimental Results, experimental results show the training time and the compression ratio of the proposed SVM model. Furthermore, performance including accuracy and response time (for prediction) for the application of fall detection is also demonstrated. Finally, section Conclusion concludes this paper.

SVM-Based Cloud Systems with Motion Sensors for Fall Detection

Along with the rapidly development of cloud computing and growing popularity of sensor devices, a cloud system with sensor devices have been prevailed in many applications. However, a cloud system combined with sensor devices will encounter three major issues including (1) *big data* (2) *massive communication* and (3) *heavy computation* during applications. To optimize the performance of such system for real-time fall detection, these three challenges need to be taken care of properly.

The idea of data compression is employed to deal with the first two problems (i.e. big data and massive communication). Motion sensors like Microsoft Kinect applies data compression to convert an image to skeleton points and thus are used as the sensor devices in our fall-detection cloud system. Moreover, to alleviate heavy computation, support vector machine (SVM), a machine-learning algorithm to extract only important data points, is also incorporated in this work. Therefore, Fig. 2 shows the overall architecture of our cloud system where motion sensors and the SVM model are the two key components.

Skeleton-Based Motion Sensor

In the proposed cloud system, motion sensors like Microsoft Kinect are responsible for extracting motion data from human beings. Compared to camera sensors which need to process huge amount of pixels, to wait for long uploading time and to perform complex pattern recognition, motion sensors leave only three-dimensional

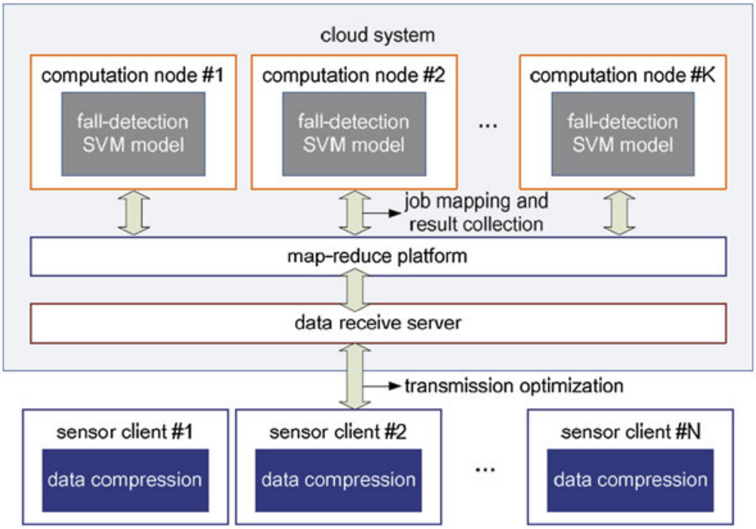


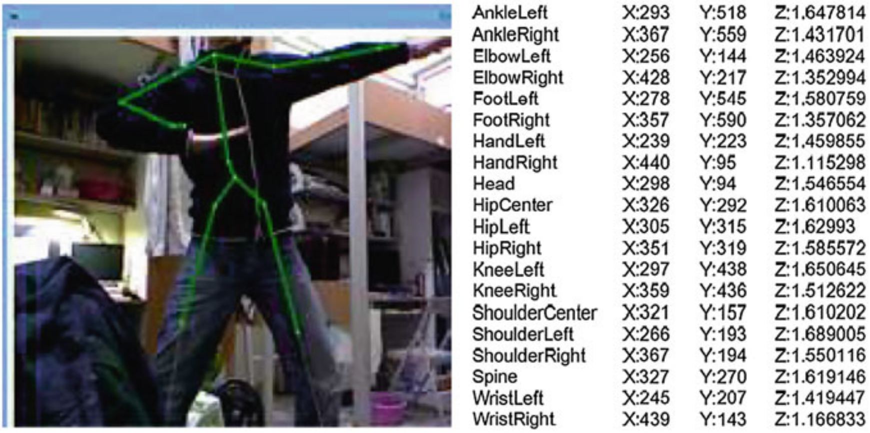
Fig. 2 Architecture of the SVM-based cloud system with motion sensors

Table 1 Comparison between cameras, accelerometers/gyroscopes and motion sensors

Sensor device	Cameras	Accelerometers/gyroscopes	Motion sensors
Data transmission loading	Large	Small	Small
Accuracy (%)	93.3	90.0	97.0
Application flexibility	High	Low	High
Convenience	High	Low	High

skeleton points to represent a human being, and thus compress data, increase data-uploading efficiency and lower computation greatly. Table 1 shows the comparison among cameras [12], accelerometers/gyroscopes [9] and motion sensors from different perspectives. As a result, the motion sensor is the best candidate for the sensor device in our fall-detection cloud systems.

Microsoft Kinect is one new type of motion sensors and can represent the body structure of a human being as twenty three-dimensional skeleton points shown in Fig. 3. Compared to the other types of motion sensors, such as ASUS Xtion [15], Kinect can capture skeleton points more accurately with a wider angle. Furthermore, Kinect supports up to thirty time-frames per second where each time-frame contains twenty points (x, y, z)'s (each point contains three coordinate information, x, y and z). Therefore, Kinect motion sensors can leave only $30 \times 20 \times 3 \times 4 = 7{,}200$ bytes data per second, several-order smaller than the typical image data size.



AnkleLeft	X:293	Y:518	Z:1.647814
AnkleRight	X:367	Y:559	Z:1.431701
ElbowLeft	X:256	Y:144	Z:1.463924
ElbowRight	X:428	Y:217	Z:1.352994
FootLeft	X:278	Y:545	Z:1.580759
FootRight	X:357	Y:590	Z:1.357062
HandLeft	X:239	Y:223	Z:1.459855
HandRight	X:440	Y:95	Z:1.115298
Head	X:298	Y:94	Z:1.546554
HipCenter	X:326	Y:292	Z:1.610063
HipLeft	X:305	Y:315	Z:1.62993
HipRight	X:351	Y:319	Z:1.585572
KneeLeft	X:297	Y:438	Z:1.650645
KneeRight	X:359	Y:436	Z:1.512622
ShoulderCenter	X:321	Y:157	Z:1.610202
ShoulderLeft	X:266	Y:193	Z:1.689005
ShoulderRight	X:367	Y:194	Z:1.550116
Spine	X:327	Y:270	Z:1.619146
WristLeft	X:245	Y:207	Z:1.419447
WristRight	X:439	Y:143	Z:1.166833

Fig. 3 Twenty three-dimensional skeleton points captured by Microsoft Kinect

Support Vector Machine

Support vector machine (SVM) is an advanced algorithm, which is widely used for machine learning problems [14] and has three major characteristics:

- SVM can find a global optimal solution for a convex problem easily.
- SVM can compute high-dimensional data effectively.
- SVM can use only a small subset to derive a decision boundary for a collection of data where each sample in the subset called a support vector.

SVM is often used to derive a function as the decision boundary with minimal errors and a maximal margin to separate data in a multi-dimensional space. An illustration of separating the data set with many possible decision boundaries is

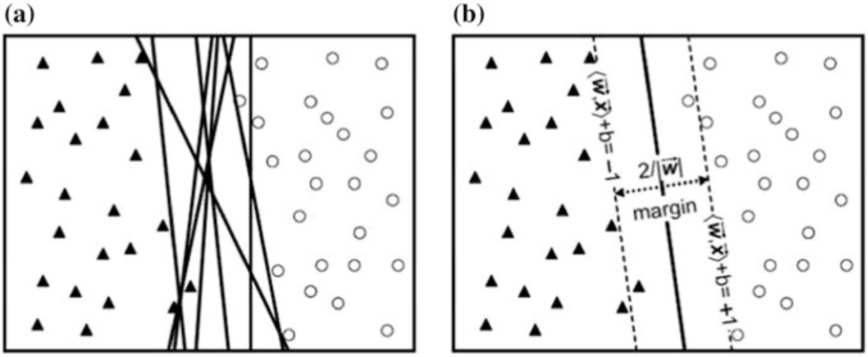


Fig. 4 Linear decision boundaries for a two-class data set [16]

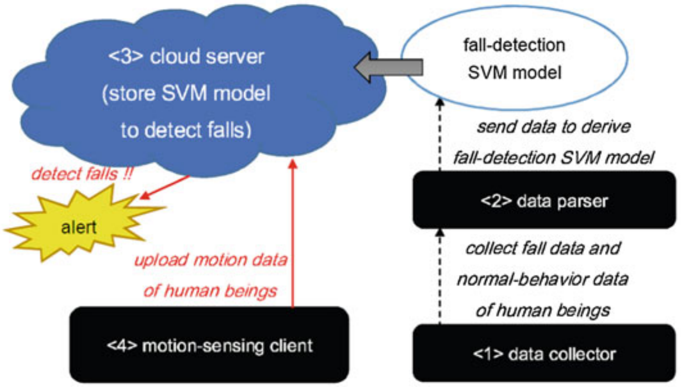


Fig. 5 Overall training framework

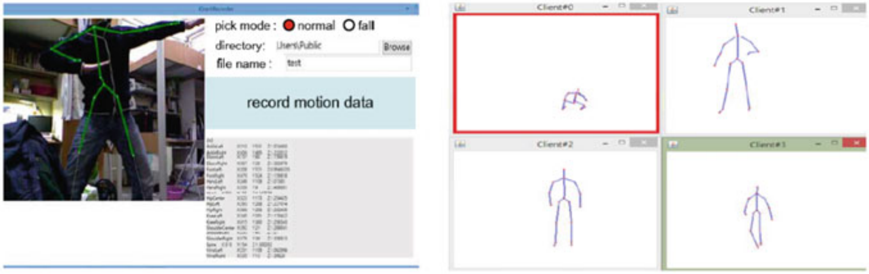
shown in Fig. 4a [16] where Fig. 4b indicates a boundaries with the minimal errors and the maximal margin.

SVM can be applied to classification problems and derives a smooth function that minimizes slacks through three steps: (1) primal-form optimization, (2) dual-form expansion, and (3) kernel-function substitution. The nature of the classification problem is first presented by the primal form. Then, through kernel functions, Lagrange method transforms from the primal form into the dual form. In this paper, fall data and normal-behavior data are both used to train SVM models with four different kernel functions (linear, polynomial, radial and sigmoid). After deriving the SVM models, it can be used to distinguish fall accidents from normal behaviors of human beings.

Data Collection and Fall-Detection Model

To build a real-time SVM-based cloud system with motion sensors, this work proposes a training framework consisting of a data collector, a data parser, cloud severs and motion-sensing clients as shown in Fig. 5.

1. Data collector: a data collector in Fig. 6a is a software implemented to collect motion data (either falls or normal behaviors) from Kincet devices. Then, these data will be sent to the data parser. The time-length of a datum can be defined at user’s discretion. In our experiment, each datum contains motion information of three seconds long.
2. Data parser: this tool reads the skeleton data including fall and normal-behavior ones into the cloud server for the SVM training to derive models.
3. Cloud server: it applies the pre-built SVM models to first distinguish falls from normal behaviors immediately and then return the result to the clients. If a fall


pick mode : ● normal ○ fall

directory: Users\Public

file name : test

(a) data collector

(b) cloud server provide multiple service

Fig. 6 Data collector and cloud server in our training framework

accident is detected, the cloud server will return an alert signal to the client as shown in Fig. 6b.

4. Motion-sensing client: every 0.5/1/1.5 s, it sends out a set of motion data of three seconds long to the cloud server for fall detection.

After using the data collector to collect a set of fall samples and normal-behavior samples, a half of the data is used to train fall-detection SVM models. Then, the other half is used to evaluate the accuracy of the SVM model.

Experimental Results

Our framework was implemented in C# for the client and in Java for the server. It ran on a Linux machine with a Intel Core i5 (3.6 GHz) processor and 2 GB memory. 320 fall data and 320 normal-behavior data were collected through our data collector. Each datum contains three-second long skeleton-points as the motion information. A half of the fall data and normal-behavior data was used to train the SVM model while the other half was used to validate the SVM model. In the experiments, the training time, compression ratio, accuracy and fall-detection time of the SVM models under different kernel functions (linear, polynomial, radial basis and sigmoid) and different update periods (0.5, 1.0 and 1.5 s) of input data are presented. The default parameters provided by the LIBSVM library [17] of the SVM-model kernel functions were used in this paper.

Table 2 compares the SVM-model training time and SVM-model compression ratio under different kernel functions and different update periods of input data. Training SVM models with the linear and polynomial kernel functions are preferable over the other two due to better compression. Particularly, training by the radial basis kernel function results in no compression. In addition, using a longer update period of input data (i.e. 1.5 s) can derive the model faster than using a shorter period (0.5 s) on the same data. However, the compression ratio of a

Table 2 Training-time and compression-ratio comparison on different settings

Kernel function	Training time (s)			Compression ratio (%)		
	Update period of input data			Update period of input data		
	0.5 s	1.0 s	1.5 s	0.5 s	1.0 s	1.5 s
Linear	17.6	6.8	4.7	0.70	1.19	1.79
Polynomial	17.0	6.9	4.8	0.82	1.39	1.85
Radial basis	2620.0	226.0	104.0	100.00	100.00	100.00
Sigmoid	39.0	15.6	10.9	2.88	5.68	8.39

Table 3 Fall-detection accuracy and time comparison on different settings

Kernel function	Fall-detection accuracy (%)			Fall-detection time (10^{-3}s)		
	Update period of input data			Update period of input data		
	0.5 s	1.0 s	1.5 s	0.5 s	1.0 s	1.5 s
Linear	99.89	99.82	99.79	1.06	0.94	0.86
Polynomial	99.90	99.84	99.81	1.00	0.91	0.87
Radial basis	99.27	98.55	97.84	89.52	47.03	32.09
Sigmoid	99.27	98.55	97.84	3.02	2.96	3.06

shorter update period is better than that of a longer update period for most of the cases in Table 2.

Table 3 compares the accuracy and fall-detection time of SVM models under different kernel functions and different update periods. All models can reach more than 97 % accuracy. Particularly, using SVM models with the linear and polynomial kernel functions can achieve higher accuracy (≧99.79 %) for all cases. In addition, the SVM models derived from the linear and polynomial kernel functions only take less than 10^{-3} s to detect falls. Therefore, based on the experimental results, the linear and polynomial kernel functions are recommended to be used for deriving the fall-detection SVM model. To sum up, our experiments demonstrate that the proposed SVM-based cloud system with motion sensors can run at real time and detect fall behaviors accurately.

Conclusion

For enabling a real-time highly-accurate fall-detection cloud system, skeleton-based motion sensors (i.e. Microsoft Kinect) and support-vector-machine (SVM) learning are employed. Using skeleton trace of motions can provide high privacy, high convenience, high application flexibility and effective data compression. Moreover, the SVM models derived from the linear and polynomial kernel functions can result in high accuracy (≧99.79 %) and take less than 10^{-3} s for fall detection. Through the experimental results, the proposed SVM-based cloud system with motion sensors has been successfully demonstrated to be able to detect falls at real time with high accuracy.

References

1. Tomii S, Ohtsuki T (2012) Falling detection using multiple doppler sensors. In: IEEE 14th international conference on e-health networking, applications and services (Healthcom), pp 196–201
2. Popescu M, Hotrabhavananda B, Moore M, Skubic M (2012) VAMPIR- an automatic fall detection system using a vertical PIR sensor array. In: 6th international conference on pervasive computing technologies for healthcare (Pervasive Health), pp 163–166
3. Lustrek M, Kaluza B (2009) Fall detection and activity recognition with machine learning. informatica (Slovenia), pp 197–204
4. Yu M, Rhuma A, Naqvi SM, Wang L, Chambers J (2012) A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment. IEEE Trans Inf Technol Biomed 16(6):1274–1286
5. Cheng J, Chen X, Shen M (2013) A framework for daily activity monitoring and fall detection based on surface electromyography and accelerometer signals. IEEE J Biomed Health Inf 17(1):38–45
6. Tong L, Song Q, Ge Y, Liu M (2013) HMM-based human fall detection and prediction method using tri-axial accelerometer. IEEE Sens J 13(5):1849–1856
7. Lan C-C, Hsueh Y-H, Hu R-Y (2012) Real-time fall detecting system using a tri-axial accelerometer for home care. In: international conference on biomedical engineering and biotechnology (iCBEB), pp 1077–1080
8. Li Q, Stankovic JA, Hanson MA, Barth AT, Lach J, Zhou G (2009) Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In: Sixth international workshop on wearable and implantable body sensor networks, pp 138–143
9. Liu C-H, Hsieh S-L (2011) A fall detection system using accelerometer and gyroscope. Master thesis, Tatung University
10. Ozcan K, Mahabalagiri AK, Casares M, Velipasalar S (2013) Automatic fall detection and activity classification by a wearable embedded smart camera. IEEE J Emerg Selected Topics Circuits Syst, PP(99): 1–12
11. Yu M, Naqvi SM, Rhuma A, Chambers J (2012) One class boundary method classifiers for application in a video-based fall detection system. IET Comput Vision 6(2):90–100
12. Lei C-W, Wan T-P (2011) Using distributed video coding and decoding-friendly encoder design for video in the cloud. Master thesis, National Kaohsiung First University of Science and Technology, Taiwan
13. Microsoft kinect, http://www.microsoft.com/en-us/kinectforwindows/
14. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
15. ASUS xtion, http://www.asus.com/Multimedia/Xtion_PRO/
16. Peng H-K, Huang H-M, Kuo Y-S, Wen H-P (2012) Wen: statistical soft error rate (SSER) analysis for scaled CMOS designs. ACM transactions on design automation of electronic systems (TODAES), 17(1): 9:1–9:24
17. Chang C-C, Lin C-J, LIBSVM- A library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

A Study on Persuasive Effect of Preference of Virtual Agents

Akihito Yoshii and Tatsuo Nakajima

Abstract A virtual agent can have a graphical appearance and give users non-verbal information such as gestures and facial expressions. A computer system can construct intimate relationship with and credibility from users using virtual agents. Although related work have been discussing how to make users feel better about computers, a room for discussing effectiveness of letting users choose their favorite characters still exists. If a user's favorite agent is more believable than not favorite one for him/her, the computer system can construct better relationship with the user through an agent. In this paper, we have examined an effect of making user's favorite agent selectable on his/her behavior by conducting an experiment. We divided participants into four groups according to these conditions and ask them to have a conversation with an agent. As a result, we found a possibility of increasing credibility of an agent from users by letting them choose their favorite one.

Introduction

Persuasion and Agents

Persuasion is an attempt to encourage an individual to change his/her behaviors or attitudes [1]. Researches on persuasive computer systems have accelerated as the prevalence of computers and the Internet.

A. Yoshii (✉) · T. Nakajima
Department of Computer Science and Engineering, Waseda University, Tokyo, Japan
e-mail: a_yoshii@dcl.cs.waseda.ac.jp

T. Nakajima
e-mail: tatsuo@dcl.cs.waseda.ac.jp

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_6, © Springer Science+Business Media Dordrecht 2014

Using a virtual agent as a persuader is one of methods of constructing closer relationship between a computer and a user. In this paper, we use a word "virtual agent" as the meaning of visual entity represented graphically by a computer; specifically a character which can have a conversation with its user. Related work on using conversational agents as a user interface has also been exists. For example, Bickmore et al. have discussed a model of dialogue building trust from users mentioning a *relational agent* which uses verbal and nonverbal conversational strategies same as human uses [2].

Credibility of Computer Systems

Whether a user credits a computer or not also influences on outcomes of persuasion from the computer; for example, Fogg has described credibility in persuasion [1].

Credibility can be used as a clue of whether one can believe computers or other individuals. For example, credibility which is called *surface credibility* is one of four types of credibility described in [1]. This kind of credibility comes from first impression of surface traits such as appearance [1].

A room for discussion of making users' impression of computers better in an aspect of credibility still exists. Attempts of giving users favorable impression have existed and the discussed agents have varied from text-based to graphical ones (for example, [2] has summarized related work). In addition, a computer system which user can choose and interact with his/her favorite agent also exists [3]. However, a comparison between the case an agent can be chosen by a user and an agent is fixed regardless of a user's preference can give a profound aspect. That is, if the agent just meets a user's preference or a character from a user's favorite anime or games, s/he is expected to interact with the agent with eagerly by enhanced persuasive effect.

Our Purpose

In our research, we examine a degree of persuasiveness of conversation with an agent considering two cases. One case is that a user can choose an agent according to his/her preference and the other is that a user can not choose his/her favorite agent.

We constructed an agent system using existing software and conducted an experiment using our system. In the experiment, we asked male participants who belong to one of four groups combining two different conditions to talk with a female agent; after the conversation, we interviewed them about the talk with the agent.

Based on the results of the experiment, we will also discuss an effectiveness of reflecting preferences of a user to selection of an agent; in an aspect of designing credible persuasive computer systems which incorporate with agents.

Related Work

In this section, we introduce related work on an agent and discuss the relationship with our case.

An Appearance of an Agent

Zanbaka et al. have examined effect of gender and appearance of a virtual agent as a persuader [4]. In this research, an experiment has been conducted where three kinds of agents persuade users about the same topic. One agent is a picture of a real person (human), another is a CG based human-like agent (virtual human) and the other is a CG based not-human-like agent (virtual character).

As a result, a user has been more persuaded by an agent with different gender than with same although significant effects of appearance and gender on persuasion did not exist. In addition, although users had positive impression toward the virtual agents, a virtual character can be perceived as bolder than other types of an agent.

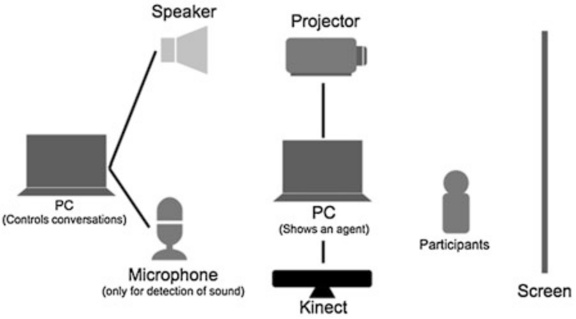
Affiliation Need

Katagiri et al. have mentioned a need to establish and maintain affinitive relationships with others, called as affiliation need. They have examined construction of relationship between an agent and a user based on affiliation need using an exhibition guidance system incorporates with an agent [3]. With this system, a user can receive explanations and recommendations of exhibition transferring an agent from his/her portable device to an information terminal besides the exhibition.

A user can choose his/her favorite agent out of nine kinds of agents. An agent tells the user that it will wait for him/her at the exhibition which has been recommended by the agent after the fourth visit to exhibitions. If the user goes to the recommended exhibition, the agent appreciates him/her; on the other hand, if the user does not, the agent complains about it. The study has confirmed whether these reactions of the agent induce affiliation need from the user and affect later behavior of the user.

Consequently, the result has showed that those who had visited more than four exhibitions had changed their behavior after the fourth visit. In this experiment, participants have been divided into two groups and asked to walk around the exhibitions using the guidance system. Each member of one group finished after four visits of exhibitions while a member who belongs to the other group was able to visit more than four exhibitions. Although they have said that they need more participants in order to obtain statistically clearer results, they concluded that the interaction from the agent had had an effect on behavior of a user.

Fig. 1 An overview of the conversation system



System Design and Materials

We conducted an experiment in order to examine how will a user responds to an agent when we let them choose their favorite agent. Before the study, we constructed a conversation system which incorporates with an agent and interacts with users. We will describe details of this system in this section.

Overview

Figure 1 shows an overview of our conversation system. A user can have an oral conversation with a virtual agent which can speak in synthesized voice. An appearance of an agent and speech were presented in a wizard-of-oz style. More precisely, an agent was controlled by gestures of the researcher using a Kinect and the each sentences of speech were stepped by key input on a shell.

An agent and the speech components were deployed on different computers separately because of the difference of operating systems. An agent was controlled via a PC on which Microsoft Windows 7 was installed and the speech was generated on a virtualized Debian Linux (Squeeze) environment of another computer.

Agent

An agent was a 3D character which has been generated by software which is called as MikuMikuDance (MMD)[1]. The MMD has been originally developed a software tool to enable authors to generate a music video using 3D virtual character with synthesized vocal and music notes. We chose a Kinect to control behaviors of an agent by gestures using a plugin[2] instead of programming numerically.

[1] http://www.geocities.jp/higuchuu4/index_e.htm

[2] http://www.xbox.com/ja-JP/kinect

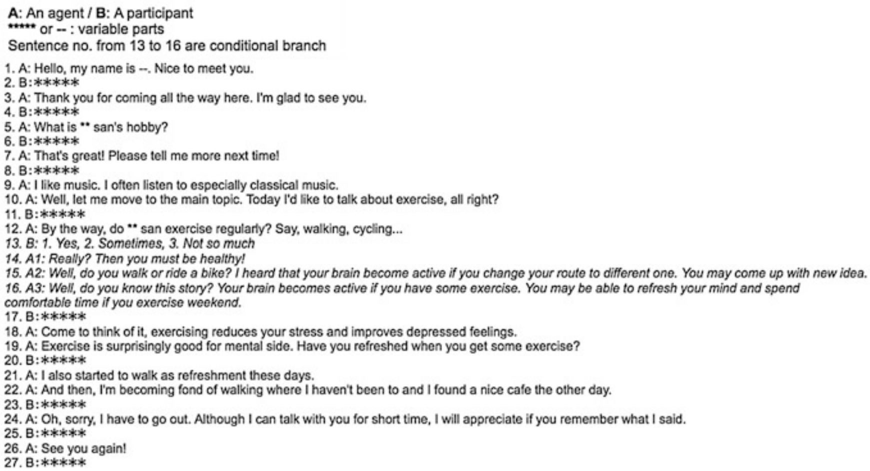
We used 3D character models bundled with books which were published by Shinyusha featuring the MMD. Such models have been provided by many people sometimes based on anime characters or games and they can be loaded to the MMD.

Speech of an Agent

The speech of an agent was synthesized by OpenHRI[3]. This software is a collection of components for Human Robots Interaction including speech synthesis and recognition. We used the speech synthesis feature based on Open JTalk[4], which is supported in the OpenHRI.

We prepared a component which receives text input from a shell and command speech related components in order to generate synthesized speech. The OpenHRI provides each features such as speech recognition or speech synthesis as components. These components can be connected to each other graphically via an input port and an output port using RT System Editor which was installed on Eclipse.

We chose a female voice because all of the character models we had prepared were female and we used the same voice for all characters. The conversations are constructed partially based on social dialogue [5]. For example, we used "empathetic statements" (line 3 in Fig. 2) and "prompting for self-disclosure by the participant" (line 5 in Fig. 2).



A: An agent / **B**: A participant

***** or -- : variable parts

Sentence no. from 13 to 16 are conditional branch

1. A: Hello, my name is --. Nice to meet you.
2. B:*****
3. A: Thank you for coming all the way here. I'm glad to see you.
4. B:*****
5. A: What is ** san's hobby?
6. B:*****
7. A: That's great! Please tell me more next time!
8. B:*****
9. A: I like music. I often listen to especially classical music.
10. A: Well, let me move to the main topic. Today I'd like to talk about exercise, all right?
11. B:*****
12. A: By the way, do ** san exercise regularly? Say, walking, cycling...
13. *B: 1. Yes, 2. Sometimes, 3. Not so much*
14. *A1: Really? Then you must be healthy!*
15. *A2: Well, do you walk or ride a bike? I heard that your brain become active if you change your route to different one. You may come up with new idea.*
16. *A3: Well, do you know this story? Your brain becomes active if you have some exercise. You may be able to refresh your mind and spend comfortable time if you exercise weekend.*
17. B:*****
18. A: Come to think of it, exercising reduces your stress and improves depressed feelings.
19. A: Exercise is surprisingly good for mental side. Have you refreshed when you get some exercise?
20. B:*****
21. A: I also started to walk as refreshment these days.
22. A: And then, I'm becoming fond of walking where I haven't been to and I found a nice cafe the other day.
23. B:*****
24. A: Oh, sorry, I have to go out. Although I can talk with you for short time, I will appreciate if you remember what I said.
25. B:*****
26. A: See you again!
27. B:*****

Fig. 2 The script of conversation

[3] http://openhri.net/?lang=en

[4] http://open-jtalk.sourceforge.net

Table 1 Groups of the participants (the unit is [person(s)])

	(A) person(s)	(B) person(s)
1	3	1
2	1	1

The microphone was used only for indicator of sound levels because the researcher attempted not to hear the speech of a participant in order to let them talk without pressure.

Experiment

An experiment consists of three parts. These parts include a pre-questionnaire, a conversation session with an agent and an interview.

Participants and Tasks

We recruited 6 participants (5 Japanese, 1 Chinese) and divided into four groups according to two conditions (Table 1). One condition is whether a participant can choose his/her favorite agent from the book (a) or not (b); the other is whether s/he has a specific favorite character, game or anime regularly (1) or not (2). As the Table 1 shows, we named each groups as combination of an alphabet and a number; for example, the group whose members do not have specific preference (2) and they can select their favorite agent (a) is group A-2. One participant could not fully understand the content of the conversation and we inquired mainly the impression of the agent and imaginary-based opinion.

Gender of the participants was unified to be male so as to exclude an effect of difference of participants' gender. All of the agents were female and according to Zanbaka et al. female speakers are more persuasive toward male participants than male speakers [4].

In the main study, each participants had conversations related to exercise with an agent. The conversations took place in Japanese and look like Fig. 2 as an English translated form.

During the conversations, the researcher heard music via earphones in order to hide the details of the participants' talk letting them talk more naturally.

Interview

Participants were asked to give an open-style interview with the researcher about the conversation and the agent after the conversation. In the interview, we firstly asked participants to tell us about the entire conversation freely and then we interviewed them according to following topics.

- How did you feel about the conversations with the agent?
- Did you really like the agent? Why?
- Did you have an interest with exercise? Why?
- Do you have a favorite character?

Results and Discussion

In this section, we will show the result of the experiment and discuss them. We will use fragments derived from comments of participants and these are edited for explanation.

Conversation with an Agent

From the results of the experiment, five participants felt the speech of an agent was unnatural. Such perception relates to all of or a part of the timing, the voice quality and the manner of speaking. For example, longer time lags or overlaps between speech of a participant and an agent have occurred. This was mainly because of manual conversation control without hearing the participant's talk. Besides, according to four participants, intonations of the speech was "machine-like".

Synthesized voice can have a negative effect on credibility of an agent with less reduction of persuasive effect. However, if we can increase credibility of agents using characteristics such as a visual appearance or a personality, negative effects of synthesized voice can be reduced. As for perception of synthesized speech, Stern et al. have compared synthesized speech and human speech. Their results have shown that synthesized speech was rated less knowledgeable and truthful. However, in terms of the persuasion, a significant difference between human speech and synthesized speech did not exist [6].

Favorableness of an Agent

In an aspect of the favorableness of an agent, a significant difference between the groups was not found. One of members of the group A-1 said that "Although I am not fun of a specific character, the agent which I have selected was favorable". On the other hand, a member of group A-2 said that "I did not feel especially about the agent". These results suggest that users' regular preference does not significantly affect the immediate favorableness of an agent. We still have to examine other characteristics such as a personality and a voice of an agent in order to reduce unnatural impression from a user.

Other comments from four participants suggested the possibility of positive impression by a selectable character. We asked the participants how they thought that they can choose their favorite agents; as for those who are in group B, we asked them to imagine the situation. In addition, we also asked them what they think if an agent of our conversation system were replaced by their "regular" favorite characters of games, anime or any other media. Comments we received were "I am not feel like talking with the agent if she were not my favorite", "Choosing from many agents was difficult. But if I could not choose an agent, she may make me less impressed" and "I did not have special feelings toward the agent this time, but I may listen to my favorite character more". On the other hand, one of members of group B told us that he will have same feelings even if he could choose his favorite agent.

Persuasiveness of an Agent

Two participants were affected by the agent's persuasion. Both of them had chosen "I do not exercise regularly" on the pre-questionnaire about exercise. One participant said that "I have got to know new facts about exercise and I may remember them and exercise someday." On the other hand, four participants were not persuaded. The comment of such participant was "I will not change my behavior unless I start a conversation with an agent on my own will".

This result suggests relationship between a topic of a conversation and participants' current exercise behavior have to be considered while the experiment with significant number of participants is expected. Specifically, we did not assign the participants to a group according to exercise behavior of them; in addition, the conversation which we have prepared contained same persuasion for all participants except for a conditional branch on lines from 13 to 16 in Fig. 2.

Each option of the questionnaire about exercise can refer to a behavior model. According to Prochaska et al. behavior can be divided into five stages from *precontemplation* level to *maintenance* [7]. Among these stages, *precontemplation* is a stage for those who do not intend to change their behavior and *contemplation* is for those who are seriously considering changing their behavior. In addition, different processes of change are needed in order to move from one stage to another smoothly. For example, giving information about a target behavior to a *precontemplation* or *contemplation* individual is *consciousness-raising*. Informing participants of positive aspect of exercising can be consciousness-raising for those who answered "I do not exercise regularly" to our questionnaire; however, not to those who exercising regularly.

Conclusion

We discussed an effect of letting a user choose his/her favorite agent on behavior of the user. In addition, we also conducted an experiment constructing an agent system with which a user can have a conversation in order to examine the effect of a favorite agent. In the study, we divided participants into four groups according to two conditions. One condition was whether a participant can choose his favorite agent and the other condition was whether a participant has a "regular" favorite characters, game or anime in his/her daily life.

As a result, making an agent selectable by a user according to his/her preference has a possibility of increasing credibility of an agent system. However, we still have to adopt personalities, voices and the contents of conversations to the appearance of an agent in order to reduce unnatural feelings from participants.

Acknowledgments The authors would like to thank those who participated our experiment and authors of components and materials.

References

1. Fogg BJ (2003) Persuasive technology. Morgan Kaufmann Publishers, Burlington
2. Bickmore T, Cassell J (2001) Relational agents: a model and implementation of building user trust. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '01, New York, USA, ACM 396–403
3. Katagiri Y, Takahashi T, Takeuchi Y (2001) Social persuasion in human-agent interaction. In: Second IJCAI (ed) Workshop on knowledge and reasoning in practical dialogue systems, IJCAI-2001. Morgan Kaufman Publishers, Menlo Park, pp 64–69
4. Zanbaka C, Goolkasian P, Hodges L (2006) Can a virtual cat persuade you? the role of gender and realism in speaker persuasiveness. In: proceedings of the SIGCHI conference on human factors in computing systems. CHI '06, New York, USA, ACM, pp 1153–1162
5. Schulman D, Bickmore T (2009) Persuading users through counseling dialogue with a conversational agent. In: persuasive '09: Proceedings of the 4th international conference on persuasive technology, New York, USA, ACM, pp 1–8
6. Stern SE, Mullennix JW, Dyson Cl, Wilson SJ (1999) The persuasiveness of synthetic speech versus human speech. Human factors: the journal of the human factors and ergonomics society 41(4), pp 588–595
7. Prochaska JO, Norcross JC, DiClemente CC (1994) Changing for good. William Morrow, an imprint of Harper Collins Publishers

Human Computing for Business Modelling

Yih-Lang Chen and Yih-Chang Chen

Abstract The importance of linking information technology to business goals and objectives is addressed by the concept of BPR. However, many methodologies of BPR do not address the implementation issues of the information systems development and the combination of this with the implementation of BPR in detail. What we need is an effective method to analyse the dynamic behaviour of the organization and to evaluate the alternative choices of solutions to problems. In this paper, we introduce the concept of participative process modelling where we view system development and BPR as cooperative activities which involve different groups of people with different competencies, viewpoints and requirements. We will also describe a framework, based on the principles of Empirical Modelling, aims to give a comprehensive view of the real-world situations so as to allow all participants to experience what it is, or would be like, to work within such situations and therefore draw on their tacit and non-explicit knowledge and experience. We propose the application of EM to provide a practical way of implementing participative BPR.

Keywords Business Modelling · Empirical Modelling · Experience-based · BPR

Y.-L. Chen
Department of Food and Beverage Management, National Kaohsiung University
of Hospitality and Tourism, Xiaogang, Kaohsiung 81271, Taiwan, Republic of China
e-mail: chenyl@mail.nkuht.edu.tw

Y.-C. Chen (✉)
Department of Information Management, Chang Jung Christian University,
Guei-Ren, Tainan 71101, Taiwan, Republic of China
e-mail: cheny@mail.cjcu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_7, © Springer Science+Business Media Dordrecht 2014

Introduction

The role of information technology as the enabler for organizational rethinking has been emphasized in much business modelling and business process reengineering (BPR) literature. However, when coming to the actual implementation of the BPR project, the implementation issues of IT are usually ignored or addressed by just picking an off-the-shelf application and changing the business processes to fit it. We have pointed out in [1] that businesses and business processes are such complex systems that the developers and users require appropriate models to understand the behaviour of such systems whether in order to design new systems or to improve the existing ones. In defining the requirements for BPR and its support systems, it is essential to have a broad understanding of the organizational environment in order to make appropriate decisions about what changes to make or which parts to retain. In this paper, we will investigate the potential of computer models as a suitable technique for business modelling. We start by introducing of the concepts of participative process modelling. After the presentation of our SPORE framework, we will assess its potential as a medium for participation in business modelling.

Participative Process Modelling

Participative process modelling can be interpreted in two different ways: *modelling of participative processes* and *participative modelling of processes*. When we think about how processes are conceived in the business area, there are two kinds of actions involved: the manual actions by human being and the automated actions by systems. These can be described in terms of two kinds of agency. The first is associated with the *internal* agents which enact the process and which are responsible for the state change. Their interaction can be referred to as a 'participative process'. The interactions of the internal agents may be governed by strict rules and reflect different degrees of autonomy. The second is associated with the *external* agents whose observation and comprehension of the interactions within a process rely on the integration of many agent viewpoints. Their interaction can be referred to 'participative modelling'. The external agents are typically responsible for designing and managing the process, but may also act simply as observers.

The distinction between internal and external agents can help to refine our thinking about the issue of BPR. The basic tenet of BPR is to gain competitive advantages by rethinking business processes and the use of IT for the redesigned processes. Such rethinking can be made from two directions. External rethinking is made from the *external* observers of the business system and its products or services. The aim is to make the business more responsive and effective. *Internal* rethinking is made from the viewpoints of component agents of the process.

The aim is to find the barriers in that process or any difficulty these members may meet, and thus look for new ways to make the process more efficient. The work of BPR thus should combine both kinds of rethinking: to provide the personalized empowered environment for individuals and to develop the supporting technology for the automation of business processes and the collaborative environment.

But in most BPR contexts, the creation or modification of computer systems to support the newly designed business processes is not suitable or even has a negative impact on the business. This is because the complex issues of human factors are not taken sufficiently into consideration during the BPR work. Such rethinking is mainly based on the external viewpoints and the redesigned processes are framed in terms of preconceived interactions to serve particular purposes. So the models are designed for describing 'what' the business process looks like, but cannot express 'why' the process has a certain form, and the motivation and intents behind the activities. Without considering or understanding these factors, it will be difficult to integrate existing systems or to design new systems to meet the changing needs. Although the design of the software system does radically impact the design of the business, two important points are still ignored by conventional BPR approaches:

- The software system is still a software system with all the old problems of versioning, integration, complexity, etc.
- The rules governing good architectural practices in software are not the same as the rules governing good business structures.

The SPORE Framework

SPORE (Situated Process of Requirements Engineering) is a human-cantered framework which was proposed by the second author [2]. It is problem-oriented because the requirements in this framework are viewed as the solutions to the problems identified in the application domain. The requirements are developed in an open-ended and situated manner. It is open-ended because such requirements cannot be completely specified in advance; and it is situated because the context presented in SPORE is closely connected to the referent in the real-world domain. The SPORE framework is depicted in Fig. 1. Three kinds of inputs of SPORE are: *Key problems* of the domain which are identified by the participants in seeking to address the requirements of the proposed system. *Relevant contexts*, act as motives and constraints for the participants in creating the outputs. *Available resources* are used by participants to facilitate the creation of SPORE outputs.

There are also four kinds of outputs from SPORE. The main one is *provisional solutions* to the identified problems which are developed by participants on the basis of the available resources and the current contexts. The other three outputs, which include *new contexts*, *new resources* and *new problems*, combine with their earlier versions and in turn form the new inputs to the model for creating the next

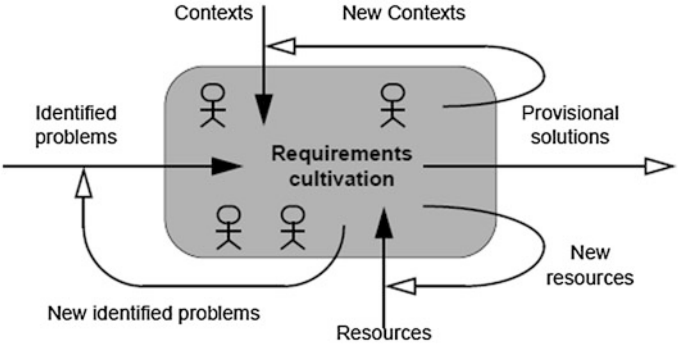


Fig. 1 The SPORE Framework

outputs. That is, all these contexts, resources and problems are modifiable and extensible in SPORE. Thus participants can develop requirements in a situated manner to respond to the rapid change in the contexts, resources as well as the problems themselves.

The models that feature in a requirements specification in conventional software development are usually too abstract for it to be possible to get a detailed understanding of the model by simply reading it. This situation may be even worse if the application domain is a complex system. Therefore having a way for the participants to experience and visualize the behaviours defined by the model is essential for system development. The SPORE framework involves constructing computer-based artefacts to be used to explore and integrate the insights of individual participants in an interactive manner. The artefacts created are ISMs and are based on the principles and tools of Empirical Modelling [3]. In EM the knowledge of participants is constructed in an experimental and not a declarative manner. That is, the insight of the participant is expressed by the coherence between what he expects in his mind and his experiments with the ISMs and the external referents. His insight can also be extended through ‘what-if’ experiments. Any introduction of new definitions, or redefinition of existing definitions, will evoke changes of state in the models and these in turn will affect the state of his mind through the visual interface.

Through our distributed EM tools the interaction of an individual participant can be propagated to other artefacts and thus affect the views and insights of other participants. In this way the participants can collaboratively interact with each other through their artefacts. This distributed EM tool supplies the framework for the collaborative environment in which the shared understanding of the key problems and their solutions (i.e. the requirements) can be established.

One of the most important benefits of interacting with the ISM is that it makes the individual insights and the shared understanding visible and communicable. This overcomes the disadvantage of invisibility and incommunicability of shared understanding based on conventionally test-based models. This experimental interaction can also keep the requirements synchronized with the shared understanding among participants, which evolves faster than textual specifications. That is, the way in which the evolution of computer models and the individual insights are synchronized allows the participants to ‘see’ other participants’ viewpoints and to ‘communicate’ with them by interacting with their own artefact.

EM for Business Modelling

In today’s business environment, information systems are becoming more interconnected to each other and are increasingly involved in complex business processes. That is, the systems used in an organization need to cooperate with human beings to achieve the organizational goals. So the information system development occurs in the context of legacy systems and business processes, which involves the issues of systems comprehension and BPR. In determining the requirements of information systems, it is necessary to understand the organizational environment so that the proposed systems can work well together with human beings. What we need is an open-ended and flexible approach to modelling the organizational environment and the behaviour of the actors and the support system. In EM, the agents, and how they relate to each other, are characterized in terms of observations. Through the construction of the computer model, the understanding and analysis of agency can be made concrete and amenable.

As mentioned earlier, a model should have both the *indicative* properties referring to properties of the existing environment and the *optative* properties referring to properties of the future environment in which the proposed system is to operate. In the developing procedure of EM (cf. Figure 2), the elicitation activity (the arrow of new input) identifies the indicative properties of the existing system (the ISM) and its environment from observation of the real-world situations. It provides the support for producing the conceptual model and generating goals from the interactions and experience of the observed system and model. Through this procedure, the interrelations between the abstract requirements (and goals) and the artefact of real-world referent can be established, and these can be used for the elicitation of system requirements.

Since the conceptual models of participants are too complicated and difficult to represent in notations which all participants can understand, we propose the cooperative validation of the conceptual models within the SPORE framework. This is carried out by participants with different roles who can elaborate different behaviours through ISMs in an interactive experiment. During the experiment, different patterns of behaviour can be explored.

In EM, the construction of ISMs is closely linked to comprehension. Thus the interactions in the real-world domain and the development and validation of ISMs are interdependent. During the validation process (the arrow of test and experiments), the indicative factors identified during the elicitation activity will be extended and adapted to fulfil the additional optative properties. This validation is

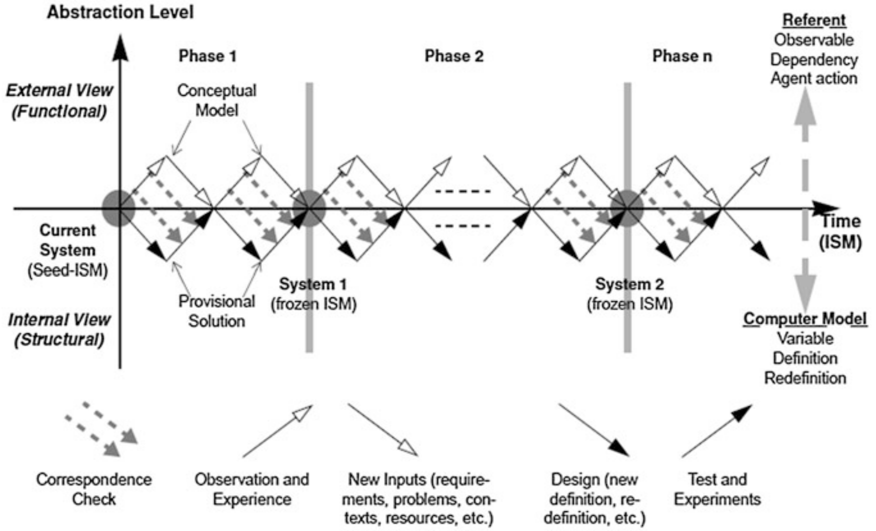


Fig. 2 The unified development procedure in empirical modelling

mainly achieved by correspondence checks. Once a reliable correspondence between the states of ISMs and those observed in the referent is established, the interaction with ISMs can serve as a representation of the understanding of real-world domain. Thus, through the correspondence checks, the modelers will gain insights into both the existing reality (the indicative properties) and the new requirements or new goals (the optative properties) which in turn form the new inputs for the next phase of system development. The open-ended characteristic of ISMs provides a way of integrating the elicitation and validation. Prior to defining and establishing the reliable patterns of state change, the interactions with ISMs, similar to activities in our everyday life, have an experimental character and are the primary means to improve our understanding of the domain.

The Characteristics of EM in Business Modelling

From our previous discussion, we infer that EM has great potential for system development and BPR. The following summarizes the characteristics of experimental interactions with ISMs, as these apply to business modelling:

- EM is an unified activity rather than the traditional development lifecycle. It can potentially address BPR from the perspectives of both the internal and external agents, and empower and allow participants to consider the meaning of the task at hand rather than becoming embroiled in peculiarities of its implementation.

- Interacting with ISMs within the SPORE framework allows participants to experiment with any entity of the business system. This means that the behaviour of both computer systems and human components can be identified and thus incorporated in a natural way. Thus the various enablers of BPR, such as information systems or human resource management strategies, can be investigated in essentially the same experimental manner.
- Under the 'open development' paradigm, participants can set up problematic scenarios, realize them visually and correct them easily and inexpensively using a 'what-if' strategy. Furthermore, the computer-based character of ISMs renders the changes and updates to the existing system apparent and makes the model maintainable and reusable.
- EM helps the participants to communicate their ideas and assess the impact of proposed changes/alternatives immediately. Participants can be guided towards a shared understanding and consensus for decision-making through the continuous evaluation, communication and checks for consistency in the distributed ISMs.
- The visibility and communicability of the interaction within the SPORE framework increases the participants' understanding of roles and relationships amongst others, as well as the effects of individual activities. This lets the participants gain feedback from the results of experiments which is used not only as a basis for comparing the alternative solutions but also as a means of evaluating their validity.
- EM allows the participants to obtain a 'global' view of the effects of 'local' changes made by individual artefacts. This assists the identification of implicit dependencies between parts of the business system.

Conclusions

Many processes in science and engineering are precisely prescribed by theories and equations, but for business processes it is difficult to consider the process and its associated real-world factors (including human factors) in an abstract and unified way. Our best line of attack is to develop the rich and flexible concepts and models to directly involve the potential users and incorporate their activities into the analysis and design of the new system or business. In this context, subject to the modeler having an adequate construal and sufficient understanding of the situation, EM has the advantage of allowing the states of the real world to be modelled in an open-ended fashion so that any new factors considered relevant can be taken into account. This kind of experimental interaction with the computer model seems to be the appropriate technique for the purpose of business process modelling. It can also be used for experimentation purposes and to help decision making during the modelling procedure. By the EM approach singular conditions can be explicitly modelled and human intervention, which is essential when modelling the scenarios in business, is possible throughout the modelling process. To sum up, EM provides flexible and human-cantered support for the design and construction of a wide range of interactive multi-user systems.

References

1. Chen YC (2012) Innovative intelligent computing for software development: an empirical modelling approach. In: international symposium on computer, consumer and control, pp 898–901
2. Chen YC (2001) Empirical modelling for participative business process reengineering. Ph.D. thesis, department of computer science, The University of Warwick, UK
3. Empirical modelling (EM) research group of the University of Warwick website, http://www.dcs.warwick.ac.uk/modelling
4. Jacobson I, Ng PW, McMahon PE, Spence I, Lidman S (2013) The essence of software engineering. Addison-Wesley, Boston
5. Warboys B, Snowdon B, Greenwood RM, Seet W, Robertson I, Morrison R, Balasubramaniam D, Kirby GNC, Mickan K (2005) An active-architecture approach to COTS integration. IEEE Softw 22:20–27

Teaching and Learning Foreign Languages in a Digital Multimedia Environment

Vladimir Kryukov, Alexey Gorin and Dmitry Mordvintsev

Abstract The paper describes the structure and functions of a digital multimedia environment as a means of teaching and studying foreign languages. The authors suggest an LMS-centered environment consisting of language labs, interactive whiteboard rooms, multimedia rooms, a videoconferencing hall, a webinar platform, and multimedia repositories, with all the components being linked together by an e-learning platform. The article is intended for experts in information technology and second language teaching methodology, as well as for all those interested in the problems of computer-assisted language learning.

Keywords Computer-assisted language learning · e-Learning · Blended learning · Multimedia environment · Language lab · Videoconferencing · Webinar

Introduction

The concept of the development and application of information technology at the Vladivostok State University of Economics and Service is based on the model of e-campus. The e-campus is aimed at increasing the efficiency of the business processes at the University by incorporating its information infrastructure and corporate information systems.

V. Kryukov (✉) · A. Gorin · D. Mordvintsev
Vladivostok State University of Economics and Service, Vladivostok, Russia
e-mail: vladimir.kryukov@vvsu.ru

A. Gorin
e-mail: aleksey.gorin@vvsu.ru

D. Mordvintsev
e-mail: dmitriy.mordvintsev@vvsu.ru

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_8, © Springer Science+Business Media Dordrecht 2014

In the era of economic, political and cultural globalisation, the University pays special attention to teaching foreign languages as a means of improving the competitiveness of its graduates in today's global marketplace.

This brought about the idea of elaborating an environment for teaching and learning foreign languages on the basis of digital multimedia technology.

A Digital Multimedia Environment for Teaching and Learning Foreign Languages

The digital multimedia environment at the Vladivostok State University of Economics and Service comprises the following key elements: language labs, interactive whiteboard rooms, multimedia rooms, a videoconferencing hall, a webinar platform, and multimedia repositories. All these elements are linked together by the University's e-learning platform.

The multimedia environment is used to deliver foreign language courses in a variety of ways [1], both in class (language labs, multimedia and interactive whiteboard rooms) and via asynchronous (e-mail, forums) and synchronous (chat rooms, webinars, and video conferences) communication channels. Such a combination of different modes of delivery, known as blended learning, represents a real opportunity to create engaging learning experiences for each and every student [2].

E-Learning Platform

Description. The e-learning platform is based on modular object-oriented dynamic learning environment (MOODLE) software which is a free open-source e-learning solution [3]. Its features are typical of an e-learning platform. Instructors can upload study materials to be viewed and/or downloaded by students, give them assignments for further submission, grade their work, publish news and announcements, organise chats and forums, administer tests and quizzes online, etc.

The main reason for choosing Moodle was its high interoperability, modular structure and open source code, which allowed for easy integration into the University's e-campus of which it is now an integral part. Instructors use their domain accounts to log in; Moodle's database is synchronised with the databases storing information about students, student groups, study subjects, schedule, etc. A special module made it possible to integrate Moodle with the University's webinar platform.

Usage. The e-learning platform is the core of the University's digital multimedia environment. Instructors create multimedia-rich electronic courses which contain training materials in the form of hyperlinks to resources in the University's digital repository and embedded videos from the video portal. Being available on the Internet and on the intranet, the e-learning platform is extensively used both in the language labs, the multimedia rooms, and the interactive whiteboard rooms and as a resource for students' independent work.

Experience and problems. The main inconvenience connected with the usage of the e-learning platform is that, if an instructor wants to include in their course a study material from the repository, they have to copy a link to the material in the repository and then paste it into a Moodle course. Similarly, including a video from the video portal requires copying and pasting the embed code.

It should be mentioned that improving the usability of the e-learning platform is of paramount importance, as one of the University's most significant goals is to further develop computer-assisted learning technologies by encouraging instructors to design interactive multimedia-rich e-learning materials that arouse students' interest and increase their academic performance and motivation.

Future plans. The University's IT personnel are now working on a Moodle module to fully integrate the digital repository and the video portal into Moodle.

Language Labs

Description. The University has three software-based language labs with 15, 25, and 30 workstations. The software Dialog Nibelung 2.0 (produced by the Russian manufacturer Lain Ltd.) installed in the labs uses the local network to link computers in the class [4]. This local network can be used for transmission of audio and video materials, various text documents and other kinds of files, and for full control of students' PCs from the teacher's workplace.

The language instructor can arrange student workplaces by groups or pairs for discussion, assign different tasks to different groups or pairs, listen in or talk to a selected student, demonstrate his/her screen to students, show a student's screen as an example to other students, record audio materials, automatically monitor students' PCs, view the selected/all students' screens, remotely control students' PCs, transmit/receive various documents to/from students, organise chat sessions, etc.

Obviously, the functions of the software are standard for the software of this kind. However, the main advantage of the solution is that it is quite inexpensive (less than $200 per workplace, excluding the cost of PCs), although it has all features necessary for teaching and learning foreign languages through discussion and extensive use of multimedia. At the same time, it allows lessons on many other subjects to be efficiently conducted, and the students' knowledge to be evaluated with the help of the integrated test system.

Usage. The language labs are used for teaching both general English and English for specific purposes (ESP), as well as for training interpreters. The largest lab (30 workstations) is also a venue for students' independent work. Not less than six hours daily are allocated for this purpose.

Experience and problems. Being purely software tools, the language labs cannot always provide the sound quality necessary for teaching and learning foreign languages [5]. Sound is sometimes lost, echoed or distorted, which complicates communication. Besides, the computers are not equipped with webcams, although this would make the language labs suitable for conducting webinars.

Future plans. To solve the problem above, the University is planning to equip the labs with gigabit switches and more perfect headsets. Webcams will also be purchased. It is also planned to replace PCs with zero clients to reduce maintenance costs and improve the efficiency of the IT infrastructure.

Multimedia Repositories

Multimedia repositories are facilities to store multimedia content used in the language labs and multimedia and interactive whiteboard rooms as well as those uploaded to the e-learning platform.

Multimedia content is stored in the University's digital repository which is intended for storing, searching and granting access to digital study materials for teaching and learning foreign languages.

Video materials are stored on the University's video portal which also makes it possible to store and search data. Videos can be viewed on the portal directly or be embedded into Moodle courses.

Multimedia and Interactive Whiteboard Rooms

The university has more than 200 multimedia rooms equipped with projectors and projection screens. Instructors use VGA outlets to connect their laptops to the projector. The multimedia rooms make it possible to enrich foreign language classes with multimedia content, with the only problem being the lack of stationary sound systems. This makes instructors carry portable speakers, which is inconvenient.

More than 20 rooms are equipped with interactive whiteboards, three of which are used for teaching and learning foreign languages. The University has a wide variety of digital multimedia resources produced by the world's leading publishers.

Apart from this, all students are given netbooks and can connect to the e-campus during class, which enables them to access the digital multimedia environment from virtually anywhere on campus.

Videoconference Hall

Description. The videoconferencing hall is intended for conducting local and international videoconferences. It holds 80 delegates and makes it possible to conduct conferences and seminars in 3 different languages with simultaneous interpretation.

Delegates' workplaces are equipped with delegates' consoles having microphones and earphones, if necessary. When a delegate's microphone is activated, one of the six video cameras installed along the perimeter of the room focuses on the delegate and transmits the image to the opposite side.

All videos are projected to the central screen or to the smaller side screens, with the latter option being used if the presenter and their power point are to be demonstrated simultaneously. All videos are stored on the server for further processing and elaboration of video lectures and tuition courses.

Usage. As practice shows, in approximately 80 % of cases the hall is used to deliver remote lectures for the University's branches and representative offices in other cities. Lectures are often delivered to both remote students and local students sitting in the hall.

The remaining 20 % of the time is occupied by various conferences and seminars, including lectures by foreign professors.

Experience and problems. Undoubtedly, the University needs a hall of such capacity. However, the large number of workplaces leads to some difficulties in its usage. The main problem is that the hall is overloaded owing to a wide variety of the functions it performs. Large-scale events are often preferred to lessons in small groups. At the same time, the usage of such powerful and costly equipment is not always justified while working with small groups.

Another problem is the lack of special soundproof booths for simultaneous interpreters whose workplaces are now located in the next room, together with the videoconference communication operator's workplace. This creates additional difficulties for two interpreters and an operator working in a small room.

One of the disadvantages of the hardware is that it is impossible to connect remote participants to different simultaneous interpretation channels. In other words, all remote participants can hear translation to only one language.

Besides, the videoconference hall does not fully meet the requirements of the training of simultaneous interpreters, as it is not intended for practical classes. All operation modes are designed for actual lectures, seminars and conferences only.

Future plans. By the end of next year, it is planned to significantly expand the presence of IT and multimedia in teaching and learning foreign languages. To meet this goal, the University developed projects of a specialist facility for teaching simultaneous interpretation and an additional facility for videoconferencing and webinars. The experts did their best to consider both teachers' and students' requirements and previously gained experience of working with similar technologies.

Webinar Platform

Description. The webinar platform is based on the BigBlueButton web conferencing system [6], and it is fully integrated into the University's IT infrastructure; instructors and students use their domain accounts to log on and access webinars. Instructors can also conduct webinars within the e-learning platform, including record and playback of sessions.

Instructors can arrange the date and time of a webinar, control the number of its participants, transmit audio and video, demonstrate documents of various formats (.ppt(x),.doc(x),.xls(x),.pdf,.jpg,.png,.pdf,.djvu,.gs, etc.), use whiteboard features, and broadcast their desktop for all students to see. It is also possible to communicate with students via the chat.

Usage. The webinar platform is used to give classes to students of the University's remote branches and makes it possible for instructors and experts from other cities and countries to deliver lectures to the University's students.

Experience and problems. The use of webinars for teaching and learning foreign languages is limited by the high cost and relatively low quality of internet connectivity for end-users.

Future plans. To solve the problem, the University is planning to arrange additional webinar facilities and to equip the language labs with webcams, so that students could be able to participate in webinars irrespective of the quality of their home internet service.

Challenges for Developing the Digital Multimedia Environment

The initial development of the LMS-centered multimedia environment faced some integration problems with the University's existing information infrastructure (particularly, with Nginx Web Server). There was also a number of third-party Moodle plugins having interoperability issues with Microsoft SQL Server.

However, the biggest challenge was connected with designing online materials. Due to the lack of financing, instructors used to confine themselves to designing the simple multiple-choice quizzes and fill-in-the-blank exercises instead of multimedia-rich study materials with carefully chosen video, sound, graphics, photographs and animations [1], which led the University to develop a set of financial incentives to address the challenge. Research [7] has shown that the described situation is typical of many educational institutions.

Future Installations

Language Lab for Training Simultaneous Interpreters

A language lab for training simultaneous interpreters was designed on the basis of the hardware and software Sanako Lab 100 STS. The lab includes 21 workplaces (16 delegates, 4 interpreters and the teacher), a multimedia projector, a personal computer, an interactive graphics tablet, a document camera, and equipment for recording digital audio.

The lab can be used both as a regular language lab and a simultaneous interpretation lab. Interpreters' workplaces are separated from the lab by a soundproof glass partition. A similar glass partition forms two booths, each for two interpreters. This makes it possible to simulate real working conditions.

The document camera is connected to the multimedia projector and makes it possible to demonstrate images of printed materials, such as A4 sheets, book leaves, or printed texts. The interactive graphics tablet makes it possible to process any image and save all notes and corrections made into it. All this can be done at the teacher's workplace.

Videoconference and Webinar Room

The room has 34 workplaces (11 videoconference delegates, 1 teacher, and 22 participants), an information display system (an interactive whiteboard, a projector, and 4 plasma TV sets), and an audio system.

The delegates' and the teacher's workplaces are equipped with microphone consoles. When a microphone is activated, of the two video cameras automatically focuses on the speaker and transmits the image to the opposite side and to the internal displays, if necessary. Signal from any device in the room can be transmitted to any display. Displays are located in such a way that each delegate or participant has visual access to all the materials displayed.

Participants' workplaces are equipped with cloud-based terminals, webcams and headsets. This enables the participants to participate in different webinars simultaneously. Besides, the teacher who delivers the webinar can demonstrate it to all participants present in the room. In the future, it is planned to move one of the existing language labs to this room, which will significantly increase its functionality.

The room is planned to be used for conducting classes in small groups, delivering lectures to remote students and recording educational videos. This will make the usage of the existing conference hall more efficient and productive, and the two facilities will complement each other.

Conclusion

In conclusion, it should be said that the digital multimedia environment at the Vladivostok State University of Economics and Service incorporates engagement and effectiveness, allowing for learning through a mixture of technological features possible in online and offline course delivery.

Foreign language teaching and learning in a blended environment promotes the high quality and efficiency of second language acquisition and creates a real opportunity for each and every individual to study a foreign language in the right time and at the right place.

References

1. Clarke A (2001) Designing computer-based learning materials. Gower Publishing, Aldershot
2. Thorne K (2003) Blended learning. Kogan Page, London
3. Moodle.org: Open-source community-based tools for learning, https://moodle.org
4. Dialog nibelung language labs, http://www.dialog.su/en/
5. Davies G, Bangs P, Frisby R, Walton E (2012) Guidance for MFL heads of department and ICT managers on setting up and using digital labs and multimedia ICT suites, http://www.camsoftpartners.co.uk/docs/CILT_Digital_Labs.htm
6. BigBlueButton web conferencing system, http://bigbluebutton.com
7. Beatty Ken (2010) Teaching and researching computer-assisted language learning, Pearson Education, London

An Interactive Web-Based Navigation System for Learning Human Anatomy

Haichao Zhu, Weiming Wang, Jingxian Sun, Qiang Meng, Jinze Yu, Jing Qin and Pheng-Ann Heng

Abstract This paper presents an interactive web-based anatomy navigation system based on the high-resolution Chinese Visible Human (CVH) dataset. Compared with previous anatomy learning software, there are three new features in our navigation system. First, we directly exploit the capabilities of graphics hardware to achieve real-time computation of large medical dataset on the web. In addition, various visualization effects are supplied to enhance the visual perception of human model. Second, to facilitate user interaction, we design a set of user-friendly interface by incorporating the Microsoft Kinect into the system, and the users can navigate the Visible Human with their hand gestures. Third, in order to eliminate the unreliable bottleneck: network transmission, we employ a progressive strategy

H. Zhu (✉) · W. Wang · J. Sun · Q. Meng · J. Yu · J. Qin · P.-A. Heng
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
e-mail: hczhu@cse.cuhk.edu.hk

W. Wang
e-mail: wangwm@cse.cuhk.edu.hk

J. Sun
e-mail: jxsun@cse.cuhk.edu.hk

Q. Meng
e-mail: qmeng@cse.cuhk.edu.hk

J. Yu
e-mail: jzyu@cse.cuhk.edu.hk

J. Qin
e-mail: jqin@cse.cuhk.edu.hk

P.-A. Heng
e-mail: pheng@cse.cuhk.edu.hk

J. Qin · P.-A. Heng
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_9, © Springer Science+Business Media Dordrecht 2014

to transmit the data between the server and the client. Experimental results validate the advantages of the proposed navigation system for learning human anatomy, indicating its great potential in clinical applications.

Keywords Anatomy navigation system · Chinese Visible Human · Graphics hardware · User interaction · Network transmission

Introduction

Due to fast development and new enhancement of network services, such as fast net-work transmission and high security mechanisms, web-based applications have attracted more and more attentions. Medical training and diagnosis systems based on web also emerged rapidly in recent years, where the doctors or experts from different locations can collaborate on specific tasks through the Internet. Major challenges in designing a successful web-based application include how to achieve real-time computation for large dataset on the web, how to distribute the workload between the server and the client so that the overall performance can be maximized, and how to design user-friendly interface to facilitate the user interaction.

Various web-based software and tools for medical applications have been proposed, and their impact on medical education is still growing [1]. Poliakov et al. [2] presented a server-client approach that enables the user to visualize, manipulate, and analyze the brain imaging dataset through the Internet. However, this approach may be degraded under bad network condition or if there are massive client requests at the same time. In [3], Mahmoudi et al. developed several interactive web-based tools for 2D and 3D medical image processing and visualization. A wide range of methods, such as registration and segmentation, are incorporated, and the client can directly use these functionalities without any plug-in installation. As hardware acceleration is not exploited, better performance is still expected for real-time applications. A web-based navigation system for Visible Human is also developed in [4], but important anatomical details are lost due to relatively low resolution of the rendered sections.

This paper extends our previous work [5] to provide an interactive web-based navigation system for users to learn the human anatomy through the Internet. Specifically, we fully exploit the functionalities of graphics hardware to achieve real-time computation on the web by utilizing the WebGL [6] API, which supports direct access to Graphics Processing Units (GPU) from the browser. Various visualization effects, e.g., surface rendering and translucent rendering, are also supported in the navigation sys-tem. Moreover, a set of user-friendly interface is designed to facilitate the user inter-action, where the Microsoft Kinect is employed to track and recognize the user' hand gestures. Lastly, to eliminate the unreliable bottleneck: network transmission, we adopt a progressive strategy to transmit the data between the server and the client. The server first sends a coarse image to the client without any delay, and finer images are later transmitted progressively if the user is interested on certain anatomical structures.

Methods

The high-resolution CVH dataset [7], with a total size about 72 GB, is employed to test the proposed anatomy navigation system. There are 3,640 slices in the dataset (0.5 mm interval) and each slice has a resolution of 3,872 × 2,048. Compared with other imaging modalities, such as CT, MRI and US, major advantage of the CVH dataset lies on its high-resolution cross-sectional images. Hence various anatomical details are preserved, which is of great value for learning human anatomy.

The overview of our anatomy navigation system is shown in Fig. 1 and there are two major parts: the server and the client. We distribute the workload between the server and the client so as to maximize the overall performance. The client visualizes the human model and tracks the user' hand gestures to determine the cross-sectional plane. The server computes the corresponding cross-sectional anatomical image and sends the results to the client progressively. In the following subsections, we detail the processing of major components.

WebGL-Based Rendering

After receiving the human model from the server, the client renders the human body with various visualization effects. However, it is not an easy task to achieve

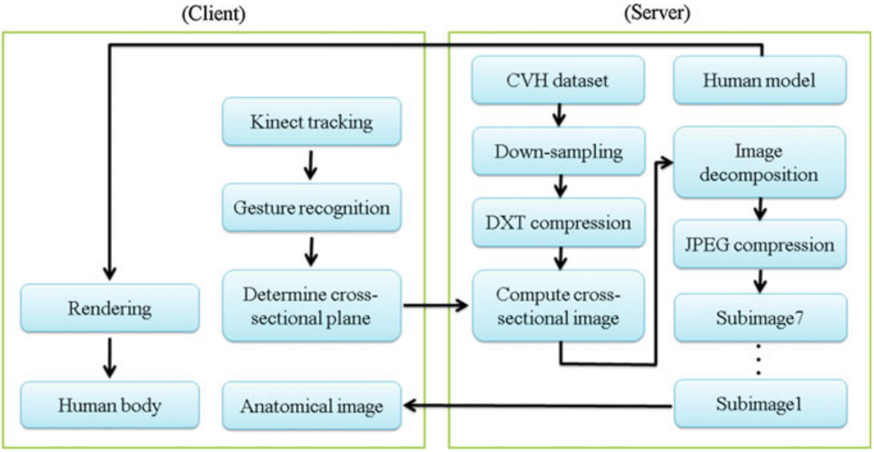


Fig. 1 Overview of the proposed anatomy navigation system

real-time rendering for complex translucent objects on the web. Traditionally, to achieve correct translucent effect with alpha blending, it is required to first sort the geometry primitives according to their depth distances and then render the primitives from front to back(or from back to front) sequentially. However, the sorting operation will be very time-consuming if the models are very complex. In [8], Everitt presented a GPU-accelerated depth peeling algorithm to solve this problem without any sorting operation. In this algorithm, each unique depth in the scene is extracted into one layer, which will be later blended in depth-sorted order to produce the final image. In practice, the algorithm is implemented with a depth test to extract the currently frontmost layer (including both color and depth information) in every rendering pass.

Unfortunately, the original depth peeling algorithm cannot be directly applied on the web due to limited supports of current WebGL standards. Adapting to the features of WebGL, we modify the algorithm to provide translucent visualization effect on the web by utilizing multiple framebuffer and texture objects to store the color and depth information of each layer separately. The major difference between the modified and original depth peeling algorithm is that we need to render the scene twice now in order to extract each layer of the scene, one for color information and the other one for depth information. After all the layers are peeled and blended correctly, we obtain the final translucent image.

User Interaction

Traditionally, the keyboard or mouse is usually employed as input device for human–computer interaction. However, the user may need to learn lots of commands in a complex training system, which requires long time practice. To facilitate the interaction process, we design a set of user-friendly interface based on the Microsoft Kinect, and allow the user to interact with the computer using their body gestures. The Kinect is a real-time video motion detecting device and is employed here to track and recognize the user' hand gestures. The hand gestures are then used to determine the position and direction of the cross-sectional plane, which will be sent to the server to compute the corresponding cross-sectional image. Currently, four hand gestures are defined in our system as follows:

- **Translation**: users can translate the position of the cross-sectional plane with the movement of one hand (Fig. 2a).
- **Rotation**: users can rotate the direction of the cross-sectional plane with a circular motion of two hands (Fig. 2b).
- **Scaling**: users can zoom in or out the cross-sectional anatomical image by adjusting the distance of two hands (Fig. 2c).
- **Screenshot**: users can screen shot the cross-sectional anatomical image by holding the fist (Fig. 2d).

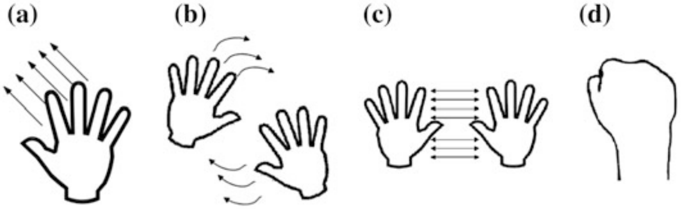


Fig. 2 **a** Translation. **b** Rotation. **c** Scaling. **d** Screenshot

Cross-Sectional Computation

Given the information about the cross-sectional plane, the server can compute the corresponding anatomical image. In order to accelerate the computation, several preprocessing steps are performed to compress the CVH dataset. In details, the dataset is first down-sampled without any visual quality degradation. Then we detect image blocks that contain meaningful contents and discard the other regions. Finally, the image blocks are compressed with DXT algorithm [9] to further reduce the data size. The advantages of DXT algorithm lie on its low quality loss and fast decompression speed. Please refer to [10] for more details.

After the above processing, the dataset is reduced to about 1 GB that is small enough to be loaded into computer memory. Because the calculation for each pixel within the cross-sectional plane is independent, the performance can be accelerated by parallelizing the computation. Moreover, to achieve smooth resultant images, 3D interpolation is employed to sample the pixels that do not lie in the dataset.

Progressive Data Transmission

To reduce the burden of data transmission, we adopt a progressive strategy to transmit the data between the server and the client. The idea is motivated by the Adam7 algorithm [11], where an image is decomposed into several small subimages with different levels of fidelity. Specifically, as shown in Fig. 3, we decompose the original cross-sectional image with resolution 3,872 × 2,048 into seven subimages as: 484 × 256, 484 × 256, 968 × 256, 986 × 512, 1,972 × 512, 1,972 × 1,024 and 3,972 × 1,024. During the transmission, the coarsest subimage is first sent to the client without any delay. If the user is interested on that anatomical image, the server will progressively send the finer subimages to the client to recover the original image.

To decrease the transmission time, we further compress the seven subimages to reduce the image size. There are lots of techniques for image compression, such as Discrete Cosine Transform (DCT) [12], Huffman entropy encoding [13], and

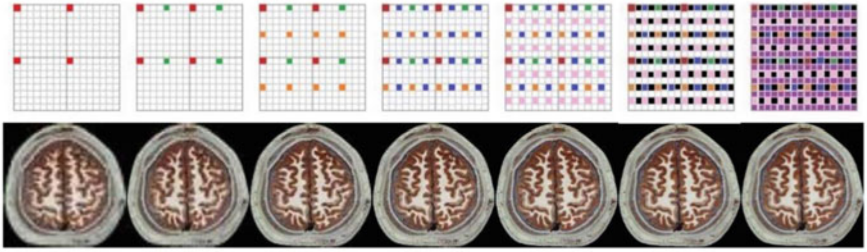


Fig. 3 Representation of image decomposition

wavelet [14, 15]. As JavaScript [16] is still not fast enough for online decompression, here we compress the images with JPEG format that is supported by most web browsers. The performance gain attribute to the JPEG compression is analyzed in the next section.

Experiments

We use the CVH dataset to demonstrate the proposed anatomy navigation system. The output for the user includes two parts. The first part is for the visualization of the human model (left of Fig. 4). The human bone is rendered with high opacity while the skin is visualized with translucent effect. The second part is for the display of the anatomical image (right of Fig. 4). The high-resolution cross-sectional image clearly shows the inner structures of the human body, and it will be very useful for anatomy teaching and learning. Lastly, a photo of our anatomy navigation system is presented in Fig. 5, where the user stands in front of the computer and uses her hand gestures to navigate the Visible Human.

Fig. 4 Screenshot of our anatomy navigation system

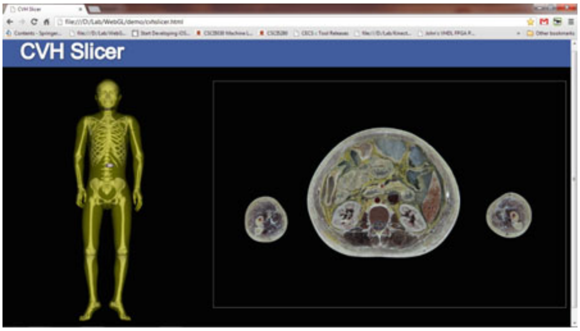
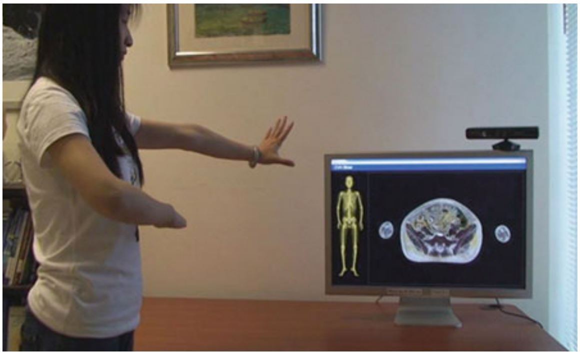


Fig. 5 A photo of our anatomy navigation system



Analysis of Data Transmission

In this section, we analyze the performance for transmitting data between the server and client to demonstrate the advantages of our progressive transmission strategy. The resolution of the cross-sectional image is 3,872 × 2,048, with 32 bits per pixel; hence the image size is about 30.25 MB. Assuming the network bandwidth of the server is unlimited and the impact of Round Trip Time (RTT) can be neglected, it takes about 121 s to transmit the image from the server to the client for bandwidth of 2 Mbit/s.

Even for bandwidth of 100 Mbit/s, it still needs 2.42 s, which is not fast enough for interactive applications.

With our progressive transmission, the resolution of the coarsest image is 484 × 256 and the image size is about 484 KB. After the JPEG compression, the data size is further reduced to about 25 KB and it only needs 0.1 s to transmit the coarsest image for bandwidth of 2 Mbits/s. The transmitting time is even shortened to 0.02 and 0.002 s for bandwidth of 10 Mbits/s and 100 Mbits/s, respectively.

In Table 1, we compare the data transmitting time among three different networks bandwidth: 2 Mbit/s, 10 Mbit/s and 100 Mbit/s. Both the average and maximum time for transmitting the coarsest subimage, the total seven subimages and the original uncompressed image are measured in the table. It can be found that the transmitting time of our progressive transmission is much lower than that for transmitting the original uncompressed image. It is also observed that the actual transmitting time is a bit larger than the above theoretical analysis. This is because that the data transmission is not only influenced by the bandwidth of the server but also the routers in the Internet. However, we still achieve real-time performance even with a moderate network bandwidth.

Table 1 Data transmitting time of three network bandwidth

Bandwidth	Coarsest subimage		Seven subimages		Original image	
	Average (s)	Maximum (s)	Average (s)	Maximum (s)	Average (s)	Maximum (s)
2 Mbits/s	0.11537	0.15816	6.8	10.7	136	156
10 Mbits/s	0.02114	0.02649	1.43	5.3	29	70
100 Mbits/s	0.00232	0.00292	0.124	0.157	4	10

User Evaluation

To validate the advantages of the Kinect-based user interface, we invite twenty students to experience and evaluate the navigation system. These students all do not have much knowledge about human anatomy, with age from 18 to 25. The experiment is conducted in the following way. First, the students are given ten anatomical images showing some common organs and structures, and are taught how to identify these organs and structures. Then these students are asked to located these organs and structures using the proposed anatomy navigation system with different interacting devices. Specifically, the twenty students are divided into two groups, ten for each group. The first group use the keyboard and mouse to navigate the Visible Human while the second group use the Kinect for interaction. We count the total time for each person to locate all the organs and structures. Lastly, we average the time for each group since to provide a statistical measure in accomplishing the task. The average time is 60.8 and 28.4 s for the first and second groups, respectively. The experiment results indicate that the Kinect-based navigation system improves the learning and training process as compared with traditional interacting devices. The reason is that the user can freely control their hand gestures using the Kinect, which facilitates the human–computer interaction.

Conclusion

An interactive web-based navigation system for learning human anatomy is presented in this paper. To achieve real-time performance, we exploit the WebGL API to directly invoke the functionalities of GPU from the browser. To provide user-friendly interface, we allow the user to interact with the computer using their hand gestures. To improve the data transmission, we employ a progressive strategy to transmit the data between the server and the client. However, the functionalities of graphics hardware have not been fully exploited yet due to limited supports of current WebGL standards. In future work, we will continue to design new web-based algorithms for online medical applications, and incorporate them into the proposed anatomy navigation system.

References

1. John NW (2007) The impact of Web3D technologies on medical education and training. Comput Educ 49:19–31
2. Poliakov AV, Albright E, Hinshaw KP, Corina DP, Ojemann G, Martin RF, Brink-ley JF (2005) Server-based approach to web visualization of integrated three-dimensional brain imaging data. J Am Med Inform Assoc 12:140–151
3. Mahmoudi SE, Asl AA, Rahmani R, Faghih-Roohi S, Taimouri V, Sabouri A, Soltanian-Zadeh H (2010) Web-based interactive 2D/3D medical image processing and visualization software. Comput Methods Programs Biomed 98:172–182
4. Hersch RD, Gennart B, Figueiredo O, Mazzariol M, Tarraga J, Vetsch S, Messerli V, Welz R, Bidaut L (2000) The visible human slice web server: a first assessment. In: Proceedings IS&T/SPIE conference on internet imaging, vol 3964, pp 253–258
5. Wang W, Meng Q, Qin J, Wei M, Chui YP, Heng PA (2013) An interactive web-based anatomy navigation system via WebGL and Kinect NextMed/MMVR20, poster
6. Marrin C (2011) WebGL Specification, Khronos WebGL Working Group
7. Zhang SX, Heng PA (2004) The Chinese visible human (CVH) datasets incorporate technical and imaging advances on earlier digital humans. J Anat 204:165–173
8. Everitt C (2001) Interactive order-independent transparency. Technical report, NVIDIA Corporation
9. Renambot L, Jeong B, Leigh J (2007) Real-time compression for high-resolution content In: Proceedings of the access grid retreat
10. Meng Q, Chui YP, Qin J, Kwok WH, Karmakar M, Heng PA (2011) CvhSlicer: an interactive cross-sectional anatomy navigation system based on high-resolution chinese visible human data. Stud Health Technol Inform 163:354–358
11. Greg R (2011) PNG: the definitive guide, O'Reilly and Associates, Inc
12. Rao KR (1976) Orthogonal transforms for digital signal processing. In: IEEE international conference on ICASSP, pp 136–140
13. Van LJ (1976) On the construction of Huffman trees. In: Proceedings of the 3rd international colloquium on automata, languages and programming, pp 382–410
14. Cohen A, Daubechies I, Feauveau JC (1992) Biorthogonal bases of compactly supported wavelets. Information Technology
15. Davis GM, Nosratinia A (1998) Wavelet-based image coding: an overview. Appl Comput Control Sig Circ 1:205–269
16. Di BM, Ponchio F, Ganovelli F, Scopigno R (2010) Spidergl: a Javascript 3d graphics library for next-generation WWW. In Web3D, pp 165–174

Community Topical "Fingerprint" Analysis Based on Social Semantic Networks

Dongsheng Wang, Kyunglag Kwon, Jongsoo Sohn, Bok-Gyu Joo and In-Jeong Chung

Abstract Community analysis of social networks is a widely used technique in many fields. There have been many studies on community detection where the detected communities are attached to a single topic. However, an overall topical analysis for a community is required since community members are often concerned with multiple topics. In this paper, we propose a semantic method to analyze the topical community "fingerprint" in a social network. We represent the social network data as an ontology, and integrate with two other ontologies, creating a Social Semantic Network (SSN) context. Then, we take advantage of previous topological algorithms to detect the communities and retrieve the topical "fingerprint" using SPARQL. We extract about 210,000 Twitter profiles, detect the communities, and demonstrate the topical "fingerprint". It shows human-friendly as well as machine-readable results, which can benefit us when retrieving and analyzing communities according to their interest degrees in various domains.

Keywords Community · Topical fingerprint · Community detection · Social semantic network (SSN) · Semantic network (SN)

D. Wang (✉) · K. Kwon · I.-J. Chung
Department of Computer and Information Science, Korea University, Seoul, Korea
e-mail: dswang2011@korea.ac.kr

K. Kwon
e-mail: helpnara@korea.ac.kr

I.-J. Chung
e-mail: chung@korea.ac.kr

J. Sohn
Service Strategy Team, Visual Display, Samsung Electronics, Suwon, Korea
e-mail: jongsoo.sohn@samsung.com

B.-G. Joo
Department of Computer and Information Communications, Hong-Ik University, Seoul, Korea
e-mail: bkjoo@hongik.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_10, © Springer Science+Business Media Dordrecht 2014

Introduction

Community analysis is widely used in many areas such as personal services, politics, commercial advertising and marketing. Many methods are investigated to detect communities through topological information such as density based algorithms [1] and modulatory algorithms [2]. Also, some topic-oriented methods utilize the profiles of actors as well as the topological information to attach a topic for the detected communities. However, it is unreasonable for a community to be explained by a single topic because the community members are generally concerned with many distinguishable interests or topics in various domains.

In this paper, we analyze the topical "fingerprint" of social network communities. First of all, we crawl a part of the Twitter, and represent the user profiles and their relationships in an ontology. For the crawled graph, we map the influential nodes, specifically celebrities, to the *WordNet Domain* [3] by the Spreading Activation (SA) mechanism. Subsequently, we detect communities through the "*Louvain* method" [2]. Finally, for the detected communities, we can retrieve the celebrities associated with each community, aggregate their domains and average them into degrees of interest. In this way, a topical "fingerprint" for each community can be retrieved and analyzed.

We extract about 210,000 Twitter users' profiles, transform them into an ontology, and then merge them with *YAGO* [4] and the *WordNet Domain* ontology. The experiment shows a topical "fingerprint" that is human-understandable. Moreover, since the domains are mapped to the *WordNet Domain* ontology, the topical "fingerprint" is machine-readable and supports semantic searching of communities. We can retrieve and analyze communities according to their degrees of interest for various domains.

Related Works

Community Detection and Analysis

Recently there has been a lot of research into community detection. A density-based clustering approach in [1] employs two distance functions to validate the advantage and limitation of them, respectively. Newman propose a significant algorithm to partition social network graphs of links and nodes into sub graphs, and an associated concept, modularity, which has also attracted a large amount of attention for development in the study of community detection [5]. Since the main drawback is that this algorithm is time-consuming, Vincent suggests the modified version of the algorithm to make it faster, giving rise to what is known as the "*Louvain* method" [2]. The method iteratively optimizes the modularity in a local way, and aggregates nodes of the same community. The modularity gain

ΔQ obtained by moving an isolated node i into a community C can be computed by the following Eq. (1) [2].

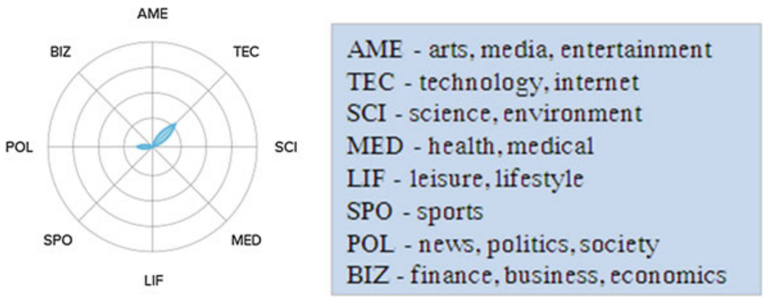
$$\Delta Q = \left[\frac{\sum_{\text{in}} + 2k_{i,\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{1}$$

In Eq. (1), $\sum_{\text{in}}$ is the sum of the weights of the links inside C, $\sum_{\text{tot}}$ is the sum of the weights of the links incident to nodes in C, k_i is the sum of the weights of the links incident to node i, $k_{i,\text{in}}$ is the sum of the weights of the links from i to nodes in C, and m is the sum of the weights of all the links in the network.

The topic-oriented method in [6] combines both social object clustering and link analysis. The entropy based method in [7] combines topological and semantic information to detect communities more accurately. The authors of [8] build community detection on a Semantic Network (SN), and prove that this method is faster, more effective, and has robust benefits. Another semantic web technology based method, called *SemTagP* [9], labels each community as tags used by people as well as relationships inferred between tags.

“Fingerprints” in Social Networks

There are some studies about a personal “fingerprint” in social networks. An interesting commercial one is called as *PeerIndex,*[1] which divides all topics into eight different areas. Figure 1 shows the topical “fingerprint” in these areas. Gunther argues that a user ID can be forged but his or her habits cannot be detected. Therefore he attempts to depict the users’ behavior “fingerprints” through tracking users’ internet behavior and express the detected features through a semantic pattern [10].



AME - arts, media, entertainment
TEC - technology, internet
SCI - science, environment
MED - health, medical
LIF - leisure, lifestyle
SPO - sports
POL - news, politics, society
BIZ - finance, business, economics

Fig. 1 *PeerIndex* personal fingerprint

[1] http://www.peerindex.com

Community Topical "Fingerprints" Based on SSN

The Social Semantic Network (SSN) we work on includes: *YAGO*, the *WordNet Domain* ontology, and the Twitter graph generated by us (i.e. nodes and relations between users). The *WordNet Domain* ontology is manually generated from the *WordNet Domain* Hierarchy[2] which is a lexical resource. In section Domain Mapping for Celebrities (or Hubs), we discuss how celebrities are classified and mapped into the *WordNet Domain* ontology. In section Community Topical "Fingerprint", we show how semantic search works. We can formalize the progress into three steps:

- Step 1: First of all, we store social network data in an ontology, and integrate this ontology with *YAGO* and the *WordNet Domain* ontologies. We then map celebrities into the *WordNet Domain* ontology by performing SA on the knowledge base, creating a huge SSN context.
- Step 2: Secondly, we detect the communities using the existing "*Louvain* method".
- Step 3: Finally, we analyze the community's topical "fingerprint" by retrieving celebrities followed by the community members, aggregating their domains and averaging by community member amount.

User profiles are presented as instances of Friend-Of-A-Friend (FOAF) concepts. The knowledge base is integrated with the social network. The social network within the Semantic Network (SN) is called the Social Semantic Network (SSN) [11]. We demonstrate the overall graph as a three layer architecture, as shown in Fig. 2. At the bottom are common actors who are following celebrities in the second layer (influential nodes). The celebrities are mapped to the *WordNet Domain* ontology in the third layer.

Definition (Social Semantic Network, *SSN*): A set *SSN* means the overall three layer Social Semantic Network, and we express it as $SSN = (SN, A, H, D, R_k, R_m)$ where:

- Set *SN* indicates the first two layers ($L1$ and $L2$); the set $L1$ indicates common actors and $L2$ expresses celebrities or influential nodes ($L1 \subseteq SN$ and $L2 \subseteq SN$). The set $L3$ expresses the domains in the third layer.
- Common actors in the first layer can be expressed as $L1 = (A, R_k)$ where $A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$ where $|A| = n$ is called the node set, and R_k is called the edge set with 'the *knows*' relationship on A. For example, if an actor a_i has 'the *knows*' relationship with a_{i+1}, then $(a_i, a_{i+1}) \in R_k$.
- Celebrities in the second layer can be expressed as the set H, and are denoted by $H = \{h_1, h_2, \ldots, h_j, \ldots, h_m\}$ with $|H| = m$.

[2] http://wndomains.fbk.eu

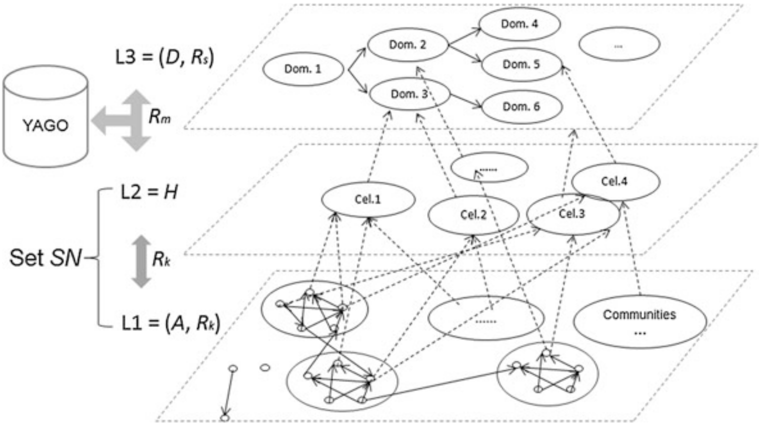


Fig. 2 Three layer architecture of SSN

- Domain hierarchy in layer three can be expressed as $L3 = (D, R_s)$, where $D = \{d_1, d_2, \ldots, d_k, \ldots, d_o\}$ indicates the set of domains with $|D| = o$, and R_s expresses the set of *subdomain* relations on D. For example, if the d_k has a *subdomain* relation of d_{k+1}, then $(d_k, d_{k+1}) \in R_s$.
- It is noted that the mapping edge set from $L1$ to $L2$ also employs the set R_k, namely, the relationship of *knows*. For example, if an actor $a_i(a_i \in A)$ follows a celebrity $h_j\left(h_j \in H\right)$, then $\left(a_i, h_j\right) \in R_k$.
- The mapping edge set from $L2$ to $L3$ employs the set R_m. If a celebrity h_j is mapped to a domain d_k, then $\left(h_j, d_k\right) \in R_m$.

Definition (A set of communities, C): A set C of communities indicates the partitions or subsets of the set $L1$ where $C = \left\{c_1, c_2, \ldots, c_p, \ldots, c_q\right\}$ with $|C| = q$ and $C_p = (A', R_k')$. For a specified community, the number of members within this community is expressed as $\left|C_p\right|$, which is also called the community size.

Domain Mapping for Celebrities (or Hubs)

This part of our work focuses on the mapping from $L2$ to $L3$, as shown in Fig. 2. The knowledge base is a SN where the knowledge is represented in patterns of interconnected nodes and arcs [13]. Celebrities are indicated by those nodes that are very influential and often have a large amount of followers. They are also significant since their trending or words affect their numerous followers and the whole social network. We regard the actors who have more than ten thousand followers as celebrities [14]. It is meaningful to map them into their corresponding domains so that we can know what kind of celebrities normal users are interested

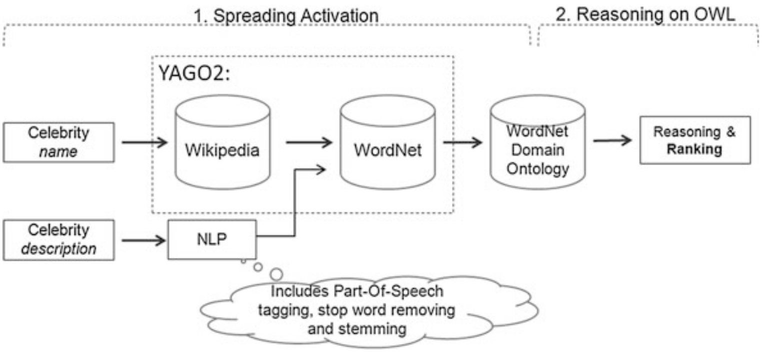


Fig. 3 Spreading activation of celebrity's profiles [12]

in and are following. The automatic mapping method is based on SA on the SN. SA is a model that simulates the memory mechanism of human brain based on a knowledge graph [15]. We apply SA technology on our knowledge base consisting of *YAGO* and the *WordNet Domain* ontology.

As shown in Fig. 3, for a Twitter celebrity, we take the celebrity's *name* and *description* as the input and perform a series of SA from nodes to their neighboring nodes along a constrained direction on the SN. The celebrity's domain can be mapped to the *WordNet Domain* ontology by reasoning the SA results based on the ontology.

Community Topical "Fingerprint"

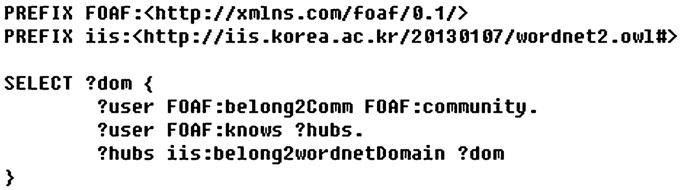
We are left with a huge interconnected SSN after the mapping process discussed in section Domain Mapping for Celebrities (or Hubs). We set a series of Boolean values based on the three layer architecture, which is shown in Fig. 2.

For the mapping from $L1$ to $L2$, We set a Boolean value of $b_{i,j}$, if an actor $a_i(a_i \in A)$ follows a celebrity $h_j(h_j \in H)$, that is, $(a_i, h_j) \in R_k$, then $b_{i,j} = 1$, otherwise, $b_{i,j} = 0$.

For the mapping from $L2$ to $L3$, we set a Boolean value of $b_{j,k}$, if $(h_j, d_k) \in R_m$, then $b_{j,k} = 1$, otherwise, $b_{j,k} = 0$.

We set a Boolean value of $b_{i,p}$ to judge whether an actor belongs to a community. If $a_i \in C_p$, then $b_{i,p} = 1$, otherwise, $b_{i,p} = 0$.

Definition (The Topical Fingerprint, l_k): For a community, we calculate the interest degrees of each basic domain by aggregating the mapping amount from $L1$ to $L3$ and averaging by the community members, called as community size. For a domain d_k, we calculate the interest degree of a community C_p as follows:



```
PREFIX FOAF:<http://xmlns.com/foaf/0.1/>
PREFIX iis:<http://iis.korea.ac.kr/20130107/wordnet2.owl#>

SELECT ?dom {
        ?user FOAF:belong2Comm FOAF:community.
        ?user FOAF:knows ?hubs.
        ?hubs iis:belong2wordnetDomain ?dom
}
```

Fig. 4 Semantic search

$$l_k = \frac{\text{aggregate amount of } d_k}{\text{size of } C_p} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} b_{i,j} \times b_{j,k} \times b_{i,p}}{|C_p|} \tag{2}$$

In Eq. (2), l_k is the interest level of domain d_k for the specified community C_p. The parameters such as i, j, k and p are explained in former definitions. In this way, we calculate the interest degrees of each basic domain for a specified community.

In this equation, l_k represents the aggregate mapping number associated with a specified community for each domain from $L1$ to $L3$. The equation can be easily computed by a single SPARQL as:

As shown in Fig. 4, the parameter *?dom* indicates total domain aggregation. *?user FOAF:belong2Comm FOAF:community* expresses the retrieval of users who belong to a specified community. *?user FOAF:knows ?hubs* retrieves the celebrities (or hubs) the user is following. *?hubs iis:belong2wordnetDomain ?dom* searches the various domains the celebrities are mapped to. In this way, we get the aggregate of the various domains. Then we aggregate each domain amount, average them by community size $|C_p|$, and get the “fingerprint” l_k.

Experiment

For the evaluation of the suggested method, we conduct an experiment with the aid of the Twitter Application Programming Interface (API). The SN (a large ontological data set) is merged and manipulated with the aid of the Jena framework.[3] We randomly fetch 210,000 profiles from Twitter and transform them into a RDF triple. From this data, we filter 15,879 celebrities and map them to the corresponding *WordNet Domain* ontology.

We adopt the “*Louvain* method”, and the modularity we calculated for this social graph is 0.57, as we know that Twitter does not have a high degree of “Socialization” [16]. The distribution of community sizes is listed in Table 1 (1,719 communities in total).

[3] http://jena.apache.org/

Table 1 Distribution of the number of community members

Members	0–100	100–200	200–300	300–400	400–500	500–600	600–700	700–800	>800
Number of community	1,649	16	9	6	4	7	2	1	25

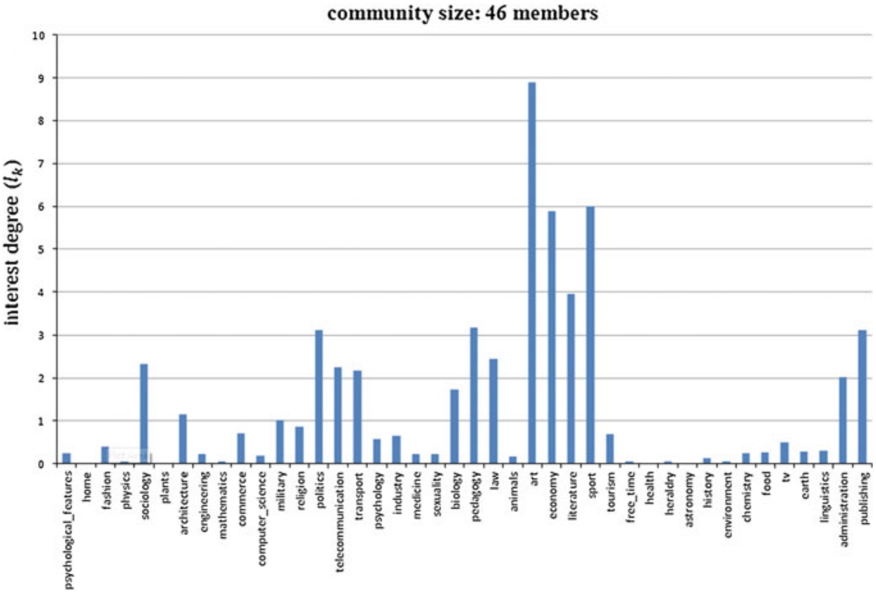


Fig. 5 Community topical “fingerprint”

We implement our method and demonstrate one of the running results. One of the communities has 46 members, and its topical “fingerprints” are illustrated in Fig. 5. As shown in Fig. 5, the average following of the *art* domain celebrities is the highest, followed by the *sport* and *economy* ones.

Conclusion

We presented a topical “fingerprint” analysis method for social network communities that takes into account the fact that community members are generally concerned with various topics in different domains. The method performs semantic search on the SSN to analyze the “fingerprint” of the community. First of all, we create a SSN context where we map the celebrities or hubs to the *WordNet Domain* ontology. Secondly, we take advantage of the previous “*Louvain* method” to detect communities. Finally, we retrieve “celebrities” or “hubs” the community is concerned with, aggregate their domains, and average by the number of community members. We extracted about 210,000 Twitter profiles, detected the communities, and analyzed their "fingerprints". It demonstrates human-friendly as well as machine-readable results. In this way, we can retrieve and analyze communities according to their degree of interest in various domains.

Acknowledgment This research was partially supported by Korea University.

References

1. Subramani K, Velkov A, Ntoutsi I, Kroger P, Kriegel HP (2011) Density-based community detection in social networks. In: 5th IEEE international conference on internet multimedia systems architecture and application, pp 1–8
2. Blondel V, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008:P10008
3. Bentivogli L, Forner P, Magnini B, Pianta E (2004) Revising the wordnet domains hierarchy: semantics, coverage and balancing. In: Proceedings of the workshop on multilingual linguistic resources
4. Suchanek FM, Kasneci G, Weikum G (2008) YAGO: a large ontology from Wikipedia and WordNet. J Web Semant 6:203–217
5. Newman MEJ (2004) Analysis of weighted networks. Phys Rev E 70:056131
6. Zhao Z, Feng S, Wang Q, Huang JZ, Williams GJ, Fan J (2012) Topic oriented community detection through social objects and link analysis in social networks. Knowl Based Syst 26:164–173
7. Cruz JD, Bothorel C, Poulet F (2011) Entropy based community detection in augmented social networks. In: International conference on computational aspects of social networks, pp 163–168
8. Xia Z, Bu Z (2012) Community detection based on a semantic network. Know Based Syst 26:30–39
9. Ereteo G, Gandon F, Buffa M (2011) SemTagP: semantic community detection in Folksonomies. In: IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, pp 324–331
10. Lackner G, Teufl P, Weinberger R (2010) User tracking based on behavioral fingerprints. In: Heng SH, Wright R, Goi BM (eds) Cryptology and network security, vol 6467. Springer, Heidelberg, pp 76–95
11. Mika P (2004) Social networks and the semantic web. In: Proceedings of IEEE/WIC/ACM international conference on web intelligence, pp 285–291
12. Wang D, Kwon K, Chung I (2013) Domain classification for celebrities using spreading activation and reasoning on semantic network. In: 5th international conference on ubiquitous and future networks
13. Sowa JF (2006) Semantic networks. Encyclopedia of Cognitive Science. Wiley, New Jersey
14. Lim KH, Datta A (2012) Following the follower: detecting communities with common interests on twitter. In: Proceedings of the 23rd ACM conference on hypertext and social media
15. Anderson JR (1983) A spreading activation theory of memory. J Verbal Learn Verbal Behav 22:261–295
16. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World Wide Web

Content Recommendation Method Using FOAF and SNA

Daehyun Kang, Kyunglag Kwon, Jongsoo Sohn, Bok-Gyu Joo and In-Jeong Chung

Abstract With the rapid growth of user-created contents and wide use of community-based websites, content recommendation systems have attracted the attention of users. However, most recommendation systems have limitations in properly reflecting each user's characteristics, and difficulty in recommending appropriate contents to users. Therefore, we propose a content recommendation method using Friend-Of-A-Friend (FOAF) and Social Network Analysis (SNA). First, we extract user tags and characteristics using FOAF, and generate graphs with the collected data, with the method. Next, we extract common characteristics from the contents, and hot tags using SNA, and recommend the appropriate contents for users. For verification of the method, we analyzed an experimental social network with the method. From the experiments, we verified that the more users that are added into the social network, the higher the quality of recommendation increases, with comparison to an item-based method. Additionally, we can provide users with more relevant recommendation of contents.

Keywords Social network analysis (SNA) · Content recommendation · Friend-of-a-friend (FOAF)

D. Kang (✉) · K. Kwon · I.-J. Chung
Department of Computer and Information Science, Korea University, Seoul, Korea
e-mail: internetkbs@korea.ac.kr

K. Kwon
e-mail: helpnara@korea.ac.kr

I.-J. Chung
e-mail: chung@korea.ac.kr

J. Sohn
Service Strategy Team, Visual Display, Samsung Electronics, Suwon, Korea
e-mail: jongsoo.sohn@samsung.com

B.-G. Joo
Department of Computer and Information Communications, Hong-Ik University, Seoul, Korea
e-mail: bkjoo@hongik.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_11, © Springer Science+Business Media Dordrecht 2014

Introduction

With the proliferation of users through Web 2.0, it is important for users to participate in a community, and create and share their contents with each other. Community-based websites such as Facebook,[1] Twitter,[2] Del.icio.us,[3] Pinterest,[4] etc. are propagating dramatically, and users in the websites share their favorites, blogs, photos, music, videos, and other content with each other. A report in 2013 described that about 83 % of people of age 18–29, and 77 % of people of age 30–49, use online social networks [1]. In spite of the great success of many community-based websites, few of them consider the characteristics of each user for content recommendation, thus making it difficult to provide users with more relevant contents.

Accordingly, we propose a content recommendation method using Friend-Of-A-Friend (FOAF)[5] ontology and Social Network Analysis (SNA) [2], to reflect the content and characteristics of the user. The method is based on Web Ontology Language (OWL) for the representation of each user profile, and employs SNA for content recommendation. The method consists of three main phases: FOAF data collection, data integration and graph generation, and SNA and content recommendation.

For the validation of the method, we collected the experimental dataset over 2,676 nodes from three different websites, and then evaluated the performance of the method, using the hit and recall ratio. From the result, we showed the method has a higher recall ratio of 0.103 than that of the item-based one, 0.02, and the quality of contents rises, as the number of users increases.

Related Works

Content recommendation in fields such as e-commerce, and community-based websites plays a significant role in providing relevant items for user purchase, and helping users to make a proper decision on the selection of good contents such as music and news [3]. Nowadays, the amount of Internet content has increased drastically, as users of online networks are able to directly create and easily share their contents. Accordingly, user selection of pertinent contents on the websites is of great importance, and research on recommendation methods has been continuously conducted. Representative recommendation methods of contents are

[1] http://www.facebook.com

[2] http://www.twitter.com

[3] http://del.icio.us

[4] http://pinterest.com/

[5] http://www.foaf-project.org/

categorized into three types, namely collaborative methods, content-based methods, and hybrid methods.

Collaborative recommendation methods are widely used in Social Network Services (SNSs) [4]. The methods have advantages, such as low computation complexity on the web server, and easier implementation, since a large number of users directly create and share their contents. Therefore, much research on the utilization of collaborative methods has been conducted in a variety of areas [5, 6]. In [5], authors used the iterative polling method for recommendation. Authors in [6] took advantage of linked-data, and proposed the open recommendation method. Although these methods are suitable for the recent trend of Web 2.0, they have difficulties in considering a diversity of relations between users, or a user and contents for recommendation, to gather each user's characteristics dynamically, and to manage their profiles efficiently. Furthermore, the quality of content for recommendation decreases significantly, when the number of users or participants to evaluate contents is not large enough [7].

Content-based recommendation methods are divided into two categories: item-based [8, 9], and user-based [10–13] recommendation methods. The item-based method [8, 9] takes advantage of contents that users like or liked in the past, and then recommends items to users. In [9], authors suggested a content recommendation method, by measuring each similarity between items. Since the method did not consider user's interests and their relations, it was not able to provide relevant recommendations according to each user's concerns. In addition, the authors in [8] extracted tags of user interests; however, they did not consider user relations in social networks. As most item-based recommendation methods mainly focus on contents to be recommended for users, they have drawbacks, in that they make it difficult to consider each user's interests, and the contents themselves.

In contrast, user-based recommendation [10, 11, 13] methods mainly concentrate on finding users, their relations, and socially common interests. In [12], authors researched the method to score items, and predict another item's score. However, most user-based methods, such as [10–13], have a common disadvantage, in that they do not make full use of characteristics and details for the recommendation of contents. In addition, the quality of recommended contents could be potentially lower, if each user profile does not have enough information [7].

Recently, research on hybrid recommendation methods, which combine both content-based and collaborative methods, has been conducted, to complement each individual method's shortcomings. The authors in [14] suggested a recommendation system for online research papers, based on research topic ontologies, Quickstep and Foxtrot. The author in [15] demonstrated a hybrid news recommendation method, using an aggregated user profile. Moreover, in [16], the authors proposed a system for the recommendation of relevant restaurants, by using domain knowledge that is related to restaurants. Their proposed methods [14, 16] are not easy to comprehensively apply to other fields, since their research has been conducted to recommend contents in a domain-specific area.

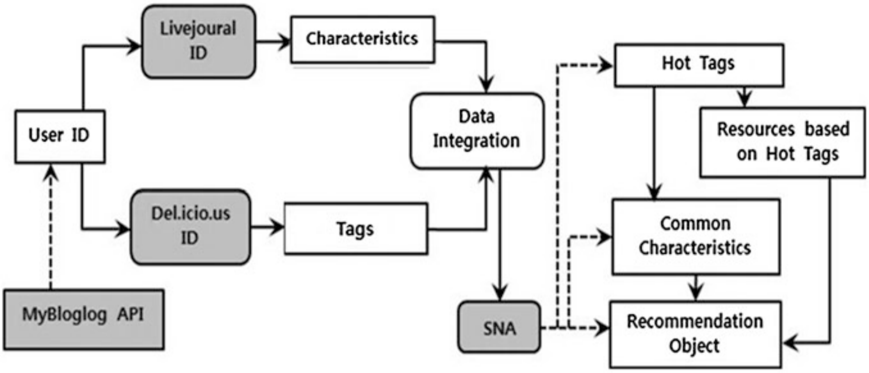


Fig. 1 Overall architecture of the suggested recommendation method

Content Recommendation Using FOAF and SNA

The proposed method can be divided into three main steps: (1) data collection, (2) data integration and graph generation, and (3) SNA and content recommendation. Figure 1 demonstrates the overall architecture of the suggested recommendation method.

Data Collection

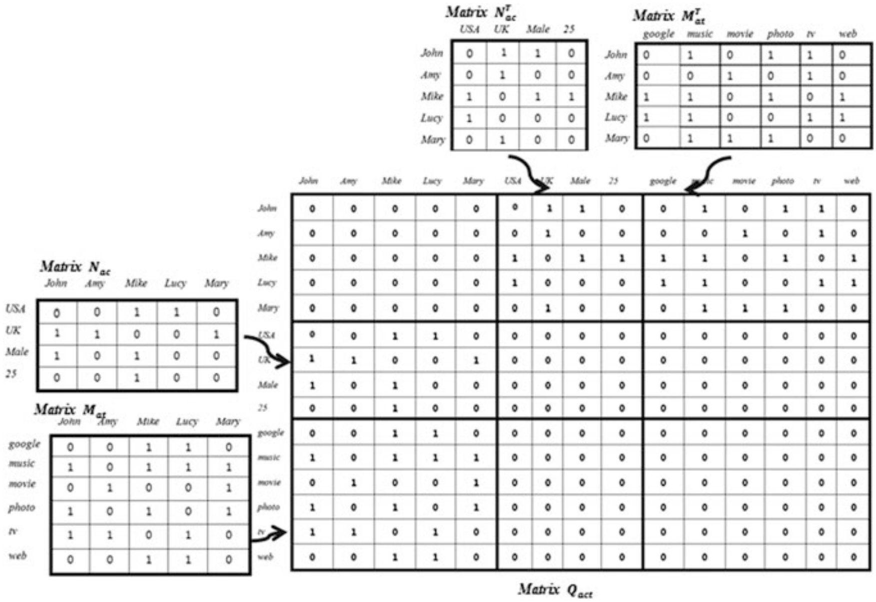
Most community-based websites provide open Application Programming Interfaces (APIs). A user FOAF profile can be obtained using the API, and we can obtain abundant information of each user. A dataset is divided into two types: characteristic data, such as user interest or taste, and tag data, as keywords for user content. We firstly provide three definitions for the proposed method, as follows.

Definition 1 (*A set A of actors*) An actor is a user who uses a web or social network service. A set of actors, A is defined as users who use the website. A set A is denoted by $A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$, where $|A| = n$, and a_i is the i-th element in the set A.

Definition 2 (*A set C of characteristics*) A characteristic of an actor includes nationality, age, and interest, etc. from FOAF ontology. A set C is represented as $C = \{c_1, c_2, \ldots, c_j, \ldots, c_p\}$, where $|C| = p$, and c_j is the j-th element in the set C.

Definition 3 (*A set T of tags*) Each tag in web documents or contents stands for keywords for them. The tags are crawled by the websites, and are defined as a set $T = \{t_1, t_2, \ldots, t_k, \ldots, t_q\}$, where $|T| = q$, and t_k is the k-th tag in the set T.

The suggested method takes advantage of the MyBlogLog API to crawl each user's ID (A). Afterwards, with the collected FOAF data of users, we extract a set of tags (T) from Del.icio.us website, and a set of characteristics (C) from LiveJournal.



Matrix N_{ac}^T

	USA	UK	Male	25
John	0	1	1	0
Amy	0	1	0	0
Mike	1	0	1	1
Lucy	1	0	0	0
Mary	0	1	0	0

Matrix M_{at}^T

	google	music	movie	photo	tv	web
John	0	1	0	1	1	0
Amy	0	0	1	0	1	0
Mike	1	1	0	1	0	1
Lucy	1	1	0	0	1	1
Mary	0	1	1	1	0	0

Matrix N_{ac}

	John	Amy	Mike	Lucy	Mary
USA	0	0	1	1	0
UK	1	1	0	0	1
Male	1	0	1	0	0
25	0	0	1	0	0

Matrix M_{at}

	John	Amy	Mike	Lucy	Mary
google	0	0	1	1	0
music	1	0	1	1	1
movie	0	1	0	0	1
photo	1	0	1	0	1
tv	1	1	0	1	0
web	0	0	1	1	0

	John	Amy	Mike	Lucy	Mary	USA	UK	Male	25	google	music	movie	photo	tv	web
John	0	0	0	0	0	0	1	1	0	0	1	0	1	1	0
Amy	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0
Mike	0	0	0	0	0	1	0	1	1	1	1	0	1	0	1
Lucy	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1
Mary	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0
USA	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
UK	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
Male	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
google	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
music	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0
movie	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
photo	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
tv	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
web	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0

Matrix Q_{act}

Fig. 2 Two-dimensional matrices, N_{ac}, M_{at}, Q_{act}

Data Integration and Graph Generation

From three collected datasets, A, T and C, the proposed method creates two-dimensional matrices, N_{ac}, M_{at}, Q_{act}, as shown in Fig. 2.

Definition 4 A matrix N_{ac} is a two-dimensional matrix, in which each dimension represents a user (a) and a characteristic (c), respectively. Each element of a matrix N_{ac} is represented by n_{xy}, with a relation of the x-th user and the y-th characteristic. If there is a relation between the x-th user and the y-th characteristic, n_{xy} is 1, and otherwise 0.

Definition 5 A matrix M_{at} is a matrix whose dimensions describe a user (a) and a tag (t). Each element in M_{at} is denoted by m_{ij} with the i-th user and the j-th tag. m_{ij} is 1 when there is a relation between the user and tag, and otherwise 0.

Definition 6 A matrix Q_{act} is a matrix which combines two matrices, N_{ac} and M_{at}, and is defined as follows (see the middle figure in Fig. 2).

$$\text{A Matrix } Q_{act} = \left\{ \begin{array}{ccc} 0 & N_{ac}^T & M_{at}^T \\ N_{ac} & 0 & 0 \\ M_{at} & 0 & 0 \end{array} \right\}$$

A matrix Q_{act} is mainly used for the generation of a social network graph. According to Definition 6, the matrix Q_{act} consists of $N_{ac}, M_{at}, N_{ac}^{T}, M_{at}^{T}, 0$. Since the matrix is symmetric, the method computes only two matrices, N_{ac}, M_{at}, thus preventing exponential increase of the computation time. M_{at}, N_{ac} and Q_{act} are converted into graphs, since they represent relations between two dimensions, such as relations between a user and a tag, and a user and a characteristic. Accordingly, three matrices are considered as graphs, which represent the relations of users, characteristics, and tags. Two constructed graphs from M_{at} and N_{ac} are represented by two sub-graphs of the generated graph from a matrix Q_{act}, since a matrix Q_{act} is the combined matrix of M_{at} and N_{ac}. We formalize the three graphs as follows.

Definition 7 (A graph *AT*) A graph *AT* is composed of a set V_{AT} of nodes and a set E_{AT} of edges, such that each edge $e \in E_{AT}$ is associated with an unordered pair of nodes. V_{AT} consists of a set *A* of users, and a set *T* of tags. A graph *AT* is denoted by $AT = (V_{AT}, E_{AT})$, where $V_{AT} = A \cup T$, and $E_{AT} = \{(a, t) \in A \times T| (a \in A, t \in T)\}$.

Definition 8 (*A graph AC*) A graph *AC* consists of a set V_{AC} of nodes and a set E_{AC} of edges, such that each edge $e \in E_{AC}$ is connected with an unordered pair of nodes. V_{AC} consists of a set *A* of users, and a set *C* of characteristics. A graph *AC* is denoted by $AC = (V_{AC}, E_{AC})$, where $V_{AC} = A \cup C$, and $E_{AC} = \{(a, c) \in A \times C| (a \in A, c \in C)\}$.

Definition 9 (*A graph ACT*) A graph *ACT* consists of a set V_{ACT} of nodes and a set E_{ACT} of edges, such that each edge $e_{ij} \in E_{ACT}$ is connected with an unordered pair of nodes, e_i and e_j. V_{ACT} is composed of a set *A* of users, a set *C* of characteristics, and a set *T* of tags. A graph *ACT* is the combination of two graphs *AC* and *AT*, and is denoted by $ACT = (V_{ACT}, E_{ACT})$, where $V_{ACT} = V_{AT} \cup V_{AC}$, and $E_{ACT} = E_{AT} \cup E_{AC}$.

Figure 3 demonstrates an example of a social network graph, *ACT* which is the combination of two graphs *AT* and *AC*. In the Fig. 3, the left graph is graph *AC*, the right graph is graph *AT*, and the middle graph is the combined graph *ACT*. The graph *ACT* has nodes that represent all users, tags, and characteristics, and edges that represent their relations. We denote users as ○, tags as △, and characteristics as □, respectively.

Social Network Analysis and Content Recommendation

Step 1 (**Measurement of degree of centrality**). The degree of centrality [17] is a SNA method used for computation of each centrality of a tag in the suggested method. The larger the value of degree centrality of a node is, the more influential the node is in the social network graph. A node that has a high value of degree centrality usually plays an important role as a hub for other nodes in the graph [2].

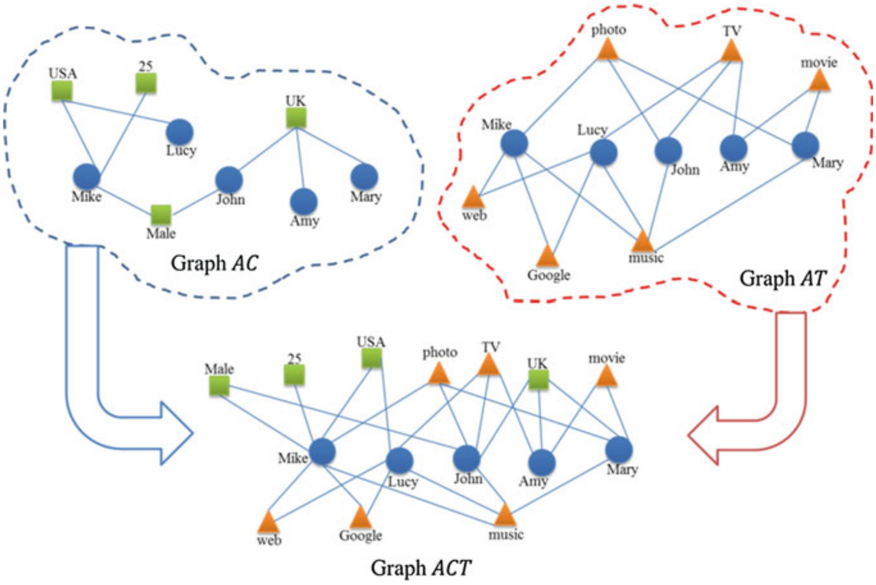


Fig. 3 An example of a social network graph, *ACT*

The degree of centrality C_D for a node is computed with Eq. (1). First, we get the number of nodes that are connected with node i in a set V_{ACT} of a graph ACT. We then calculate each degree of centrality for each node in the graph ACT, with Eq. (1).

$$C_D(i) = \frac{\sum_{j \in V_{ACT}} e_{ij}}{|V_{ACT}| - 1} \quad (0 \le C_D(i) \le 1) \tag{1}$$

Step 2 (**Selection of hot tags**). A tag that a lot of users use is defined as a hot tag. As each hot tag is dependent on the size of the social network, the reflection rate β_t of a tag t is used for a more reliable experiment. As β_t approaches 0, more hot tags are generated in the graph, and they have a lower degree of centrality, and vice versa. After computing the β_t of the tag in the graph, the threshold θ_t is calculated as Eq. (2). In the method, a tag that is higher than the threshold θ_t is selected as a hot tag.

$$\theta_t = \frac{|T| + |A|}{|V_{ACT}|} \times \beta_t \quad (0 \le \beta_t \le 1) \tag{2}$$

Step 3 (**Extraction of common characteristics**). A common characteristic is defined as a feature that the majority of users use properties of, such as *foaf:interest* of FOAF ontology. The common characteristic depends on the size of the social network, so that a reflection rate β_c for a characteristic c is defined. A reflection rate β_c for a characteristic c has the same properties as a reflection rate β_t. After computation of a reflection rate β_c for each characteristic c, a threshold θ_c for c is computed as Eq. (3).

$$\theta_c = \frac{|C| + |A|}{|V_{ACT}|} \times \beta_c \quad (0 \leq \beta_c \leq 1) \tag{3}$$

Step 4 (**Recommendation of selected content for users**). Finally, we find relevant users and contents to recommend. A user to recommend refers to one who has the same characteristics, but who does not have hot tags. A content to be recommended indicates one that will be recommended to users. After a user who has the same common characteristics chooses hot tags, contents connected with selected hot tags are recommended for the user using the proposed method.

Performance Study

Experimental Dataset

For the evaluation of the proposed method, we selected 18 users who use three SNS at the same time, and then collected user IDs from MyBlogLog, their tag data from Del.icio.us, and their characteristics from FOAF ontology in LiveJournal. We arbitrarily selected 18 users, and extracted 98 characteristics from LiveJournal, and 2,560 tags from Del.icio.us. We then conducted experiments using the tools, *Pajek* and *UCINET6*. On the collected dataset, we measured the degree of centrality for each tag, and obtained 30 hot tags (Table 1) with a reflection rate $\beta_t = 0.2$, and 36 common characteristics (Table 2) for each user with a reflection rate $\beta_t = 0.2$.

Table 1 List of selected hot tags

Business	Technology	Blog	Photo	Work	Politics
Download	Tool	Online	News	Twitter	Cards
Music	Health	E-book	Job	Learning	Free
Google	Marketing	Education	Money	Web	Networking
Software	Video	Art	Books	Advertising	Funny

Table 2 List of extracted common characteristics

Writing	80's	Female	UK	Apple	US
Reading	70's	Jazz	Discovery	Novel	50's
Art	Travel	Nature	Google	Literature	Love
Food	Advertising	Photography	Movies	Music	Prayers
Religion	CA	Painting	Singing	Cooking	Education
Charities	Internet	Male	Business	Religion	Education

Evaluation

For the verification of the proposed method, comparison with an item-based recommendation method [18] is conducted. The item-based recommendation method is an approach in which contents are assessed by users' recommendations and votes. Those that have a high score are then recommended to users [19]. For a more precise comparison of the two methods, a hit-ratio and a recall ratio in [18, 20] are adopted, as follows.

$$\text{(Proposed method)} \quad \text{Hit Ratio}(a_i) = |T_{a_i} \cap ht_{a_i}| / |T_{a_i}| \tag{4}$$

$$\text{(Item-based method)} \quad \text{Hit Ratio}(a_i) = |I_{a_i} \cap i_{a_i}| / |I_{a_i}| \tag{5}$$

Equation (4) shows the way to calculate the hit-ratio of the recommended hot tag ht_u of a set T_u of tags in a social network for an i-th actor a_i, where Eq. (5) computes the hit-ratio of the item i_u of a set I_u of items in the social network for an i-th actor a_i. Both equations range from 0 to 1 as a real number, and indicate the number of tags or items that are mapped on all tags or items in a social network.

$$\text{Recall Ratio} = \frac{\sum_{i=1}^{n} \text{Hit Ratio}(a_i)}{n} \times 100\,\% \tag{6}$$

Equation (6) demonstrates a recall ratio for n users, and represents the percentage of the summation of hit-ratios of recommended hot tags and items in a set of documents. For a comparison of the two methods, we calculated each recall ratio of tags or items, using two methods.

In Table 3, the item-based recommendation method does not grasp relevant tags, since the method suggests contents based only on tags in each user document. In contrast, the proposed method shows a higher recall ratio than the existing one, since the method not only extracts tags in each user's document, but also takes advantage of selected hot tags, using SNA and FOAF ontology.

Figure 4 demonstrates the comparison result of two methods as the number of users increases, based on Table 3. The recall ratio is an indicator that shows the number of hit results on a query for information retrieval, and points out that the higher the recall ratio, the more precise the ratio of relevant results. Thus the suggested recommendation method has a better performance than the conventional item-based one. Furthermore, as the number of users increases, the recall ratio also increases proportionally, as shown in Fig. 4. As a result, we are able to expect a much higher recall ratio, if the proposed content recommendation method is applied to a real social network.

Table 3 Comparison of the two recall ratios for the two recommendation methods

No	User	Tags	Item-based method		Proposed method	
			Item	Recall	Hot tags	Recall
1	Brajeshwar	864	4	0.005	26	0.030
2	Tuluum	246	9	0.037	24	0.098
3	Riverred	118	1	0.009	21	0.178
4	aries_hu	270	4	0.015	16	0.059
5	jamieolender	118	4	0.034	20	0.169
6	askjimcobb	182	3	0.016	12	0.065
7	valeriovillari	70	1	0.014	18	0.257
8	smoky8	280	1	0.004	19	0.068
9	NOLArising	26	0	0.000	2	0.077
10	Tucats	83	3	0.036	7	0.084
11	Fakonig	268	2	0.007	25	0.093
12	Winehiker	677	10	0.015	25	0.037
13	Jtfmulder	73	2	0.027	8	0.110
14	Blogindonesia	78	2	0.026	12	0.154
15	Leonbasin	485	5	0.010	29	0.060
16	Xian	563	4	0.007	29	0.051
17	Donssite	119	1	0.008	8	0.067
18	Clixpert	10	1	0.100	2	0.200
Average		251	3	**0.020**	16	**0.103**

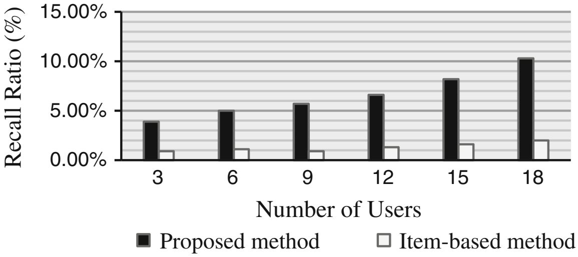


Fig. 4 Comparison of the two methods according to the number of users

Conclusions

Recently, community-based websites have dramatically developed. However, there are still several challenges in recommending pertinent contents to users in SNS, because existing methods mainly focus on the utilization of user interests that are in user account information, and not user characteristics. Therefore, we proposed an enhanced content recommendation method in online social network sites, using FOAF and SNA. The method extracts personal profiles and tags from FOAF in community-based websites, analyzes them, and then provides users with recommendations of contents. Furthermore, the method makes it easier to share contents with much enriched personal information, such as user characteristics, and tags in a social network.

For the verification of the proposed method, we collected the data over 2,676 nodes from three different websites, and then evaluated the performance of the method, using the hit ratio and recall ratio. From the experimental result, we showed that the suggested method has a better performance in recall ratio, than the item-based recommendation method. Moreover, we determined that the quality of contents rises proportionally, as the number of users increase. Therefore, the proposed method increases the reliability of recommended services, and the service providers are able to suggest more relevant contents for users. Additionally, the method is directly applicable to real social network services, blogs, etc., and feasible for sharing contents based on personalized recommendation.

For future research directions, we will conduct more experiments with a much larger size of social network, extract more characteristics of each user, and utilize them. We believe that these efforts can contribute to social network service providers, as well as a variety of other fields, such as e-commerce, e-business, and online marketing.

References

1. Duggan M, Brenner J (2013) The demographics of social media users: 2012, PewResearchCenter
2. Scott J (1991) Social network analysis. Sociology 22(1):109–127
3. Musiał K, Kazienko P, Kajdanowicz T (2008) Social recommendations within the multimedia sharing systems. WSKS'08, LNAI 5288:364–372
4. Zhou J, Luo T (2010) A novel approach to solve the sparsity problem in collaborative filtering. In: International conference on networking, sensing and control, pp 165–170
5. Jeong B, Lee J, Cho H (2009) An iterative semi-explicit rating method for building collaborative recommender systems. Expert Syst Appl 36(3):6181–6186
6. Heitmann B, Hayes C (2010) Using linked data to build open, collaborative recommender systems. Association for the advancement of artificial intelligence
7. Hwang S, Wei C, Huang Y, Tang Y (2010) Combining co-authorship network and content for literature recommendation. In: Pacific Asia conference on information systems
8. Li X, Guo L, Zhao Y (2008) Tag-based social interest discovery. In: International World Wide Web conference committee
9. Deshpande M, Karypis G (2004) Item-based top-n recommendation algorithms. ACM Trans Inf Syst 22:143–177
10. Ali-Hasan N, Adamic LA (2007) Expressing social relationships on the blog through links and comments. In: Proceedings of international conference on Weblogs and Social Media
11. Schwartz MF, Wood DCM (1993) Discovering shared interests using graph analysis. Commun ACM 36(8):78–89
12. Wang P, Ye H (2009) A personalized recommendation algorithm combining slope one scheme and user based collaborative filtering. In: International conference on industrial and information systems
13. Hassan A, Radev D, Cho J, Joshi A (2009) Content based recommendation and summarization in the blogosphere. In: 3rd international ICWSM conference

14. Campos LMD, Fernández-Luna JM, Huete JF, Rueda-Morales MA (2010) Combining content-based and collaborative recommendations: a hybrid approach based on Bayesian networks. J Approximate Reasoning 51(7):785–799
15. Mannens E, Coppens S, Pessemier T, Dacquin H, Deursen D, Sutter R, Walle R (2013) Automatic news recommendations via aggregated profiling. Multimedia Tools Appl 63:407–425
16. Bogers, T, Bosch AVD (2009) Collaborative and content-based filtering for item recommendation on social bookmarking websites. In: ACM workshop on recommender systems and the social web
17. Casciaro T (1998) Seeing things clearly: social structure, personality, and accuracy in social network perception. Soc Networks 20:331–351
18. Karypis G (2001) Evaluation of item-based top-N recommendation algorithms. In: 10th international conference on information and knowledge management
19. Lenhart A, Purcell K, Smith A, Zickuhr K (2010) Social media and mobile internet use among teens and young adults. Pew Internet and American Life Project
20. Ji A, Yeon C, Kim H, Jo G (2007) Collaborative tagging in recommender systems. AI 2007, LNAI, vol 4830, pp 377–386

The Design of an Information Monitoring Platform Based on a WSN for a Metro Station

Dong Chen, Zhen-Jiang Zhang and Yun Liu

Abstract Due to the safety and comfort problems of subway stations, we propose a system that can monitor environmental information by constructing a fundamental information monitoring platform based on a WSN. The basic idea is to deploy a large number of sensor nodes with the ability of sensing, computing and communicating in the station to form a wireless network and the information collected by different nodes will be sent in real time to the backend database. Then, we can acquire all kinds of monitoring information from the platform after a series of processing steps conducted by the platform. If any abnormal data ware found, the corrective measures will be taken immediately. The main features of our proposed platform are the design of the system's functional structure and the design of the software and hardware modules. Also, we propose to incorporate a dormancy mechanism to save energy.

Keywords Monitoring platform · Sensor nodes · WSN

D. Chen · Z.-J. Zhang (✉) · Y. Liu
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: zhjzhang1@bjtu.edu.cn

D. Chen
e-mail: 12120202@bjtu.edu.cn

Y. Liu
e-mail: liuyun@bjtu.edu.cn

D. Chen · Z.-J. Zhang · Y. Liu
Key Laboratory of Communication and Information Systems Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_12, © Springer Science+Business Media Dordrecht 2014

Introduction

The urbanization process has resulted in the rapid development and construction of urban rail transit systems. However, due to the low level of the information and integration, traditional management approaches of construction and operation have not adapted to the requirements of such systems. With more and more rail operation rely on the network, informational and intelligent collection methods are urgently needed to acquire environmental information. Railway stations have encountered many difficulties in the process of constructing the infrastructure of railway stations. The problems associated with the aging of the structure have become increasingly prominent as the service life has been extended. The underground structure of urban subway tunnels is subject to ground water erosion and vibration loading imposed by the trains. So, safety monitoring work, which is difficult to accomplish, has become very important. Because the rail networks are located underground, most stations may have many limits when they are compared with other types of space, Such as insufficient daylight, the lack of fresh air, air pollution, and high temperature [1]. All of these factors have a significant effect on the comfort of the people who use the stations. Thus, work designed to monitor the comfort levels is also important [2].

In order to improve the safety and comfort index of subway stations, we have established an omni-directional, real-time, intelligent digital monitoring platform, based on a wireless sensor network, Java programming, and database technologies, to monitor basic information related to the safety and comfort of rail stations and networks [3]. The establishment of such a platform is important because it will help provide the required information to monitor the rapidly-expanding development of rail systems, enhance the efficiency of the systems, provide a scientific approach that will enhance the management of the systems, and improve the ability to deal with emergences. In addition, the proposed platform also will increase operational safety, improve comfort levels, guarantee the safety of people's lives and property, and promote economic development [4].

Function Design

The structure of the proposed system's function menu is shown in Fig. 1:

Collection of Related Information

The information collected by different monitoring nodes will be sent in real time to the database, and management personnel can log on the screen to check the collected information. With the ability to scroll through updated fundamental

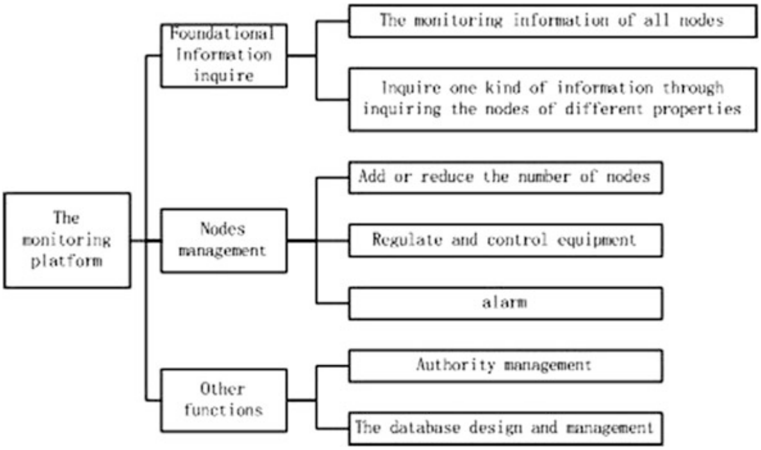

The monitoring platform
Foundational Information inquire
The monitoring information of all nodes
Inquire one kind of information through inquiring the nodes of different properties
Nodes management
Add or reduce the number of nodes
Regulate and control equipment
alarm
Other functions
Authority management
The database design and management

Fig. 1 The structure of the proposed system's function menu

information, management personnel can check and monitor information at any time [5]. There are two ways to make inquiries of the information, i.e., make an overall inquiry, which does not consider the properties of the nodes according to the time sequence of the data that are transferred and make an inquiry for only one kind of information, which focuses on one specific kind of information from the nodes that are being monitored, according to the different properties of the nodes [6]. For example, if users wish to focus on information concerning the temperatures in the system, they can check only the node's temperature information.

Node Management

The requirement of data accuracy is not always constant for the monitoring platform. For example, highly accurate data will be required when the metro is operating in one of its peak period. Conversely, the accuracy of the data can be somewhat less when there are fewer passengers using the metro. Based on the above analyses, the number of working nodes can be adjusted flexibly according to the different accuracy requirements of the data. This approach can reduce costs and save the node's energy. In addition, it can prolong the service life of the node. Control nodes can regulate equipment when they receive orders from the local control center. If monitoring nodes collect abnormal data or the administrator does not react to such data in a reasonable amount of time, the alarm function will be triggered when the values of the abnormal data value exceed the threshold that has been set.

Other Functions

Authority management

In this system, in order to prevent illegal users from getting into the system and to ensure that only lawful users can access system, we propose the authority management function to ensure that the system can be operated safely. The core mission of authority management is to give users relevant privileges according to different levels of access. User will be authorized for different levels of access, ranging from just visiting to revising data to deleting data. Also, security verification and authority distribution are the major divisions of safety management that can achieve operation that distributes the appropriate levels of authority to users who has entered the management system.

System Architecture

The system consists mainly of some sensor nodes and control nodes, a gateway node, a local monitoring center and a remote control center [7].

Sensor nodes are responsible for collecting the temperature and humidity of the air, the temperature of the operating equipment, light intensity, and other important environmental information. All of these sensor nodes must be deployed appropriately. Some of them will be deployed in a key position in the train, and some will be deployed in the tunnel. Due to limited lifespan of the batteries, the lifespan of the nodes and the communication distances are limited to some extent.

Control nodes are used to receive the control information from the local monitoring center or remote control center. They can control the ventilation and heating systems thereby regulating and controlling the environmental parameters of operational equipment. Control nodes use DC power as the power supply.

Gateway nodes serve as important information transfer stations. They are responsible for transferring the information collected by the sensors to the local monitoring center and for transferring commands to the control nodes. The data acquired by the sensor nodes will be transmitted by multi-hop routing through the other adjacent nodes. In the transmission, the monitoring data may be processed by many nodes, and the information will reach the gateway nodes through multi-hop routing. The gateway nodes will transmit the received data directly to the local monitoring center. Therefore, the computers in the monitoring center will make a series of operations including processing, storage, and analysis. Also, they will execute various kinds of control algorithms and send out corresponding control instructions to enhance the monitoring and regulation of the equipment. The local monitoring center can interact with the remote control center through the Internet, which will facilitate remote monitoring. The diagram of system architecture is shown in Fig. 2:

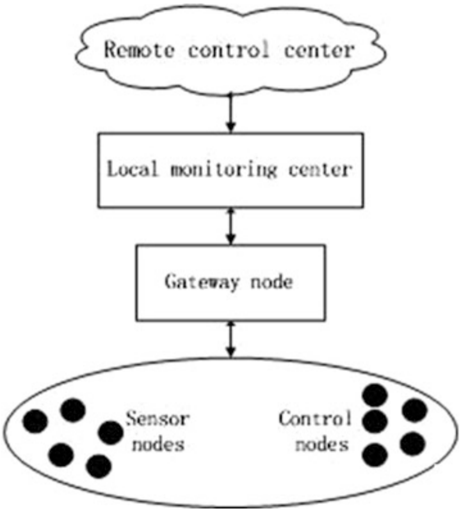


Fig. 2 The system architecture

Software Design

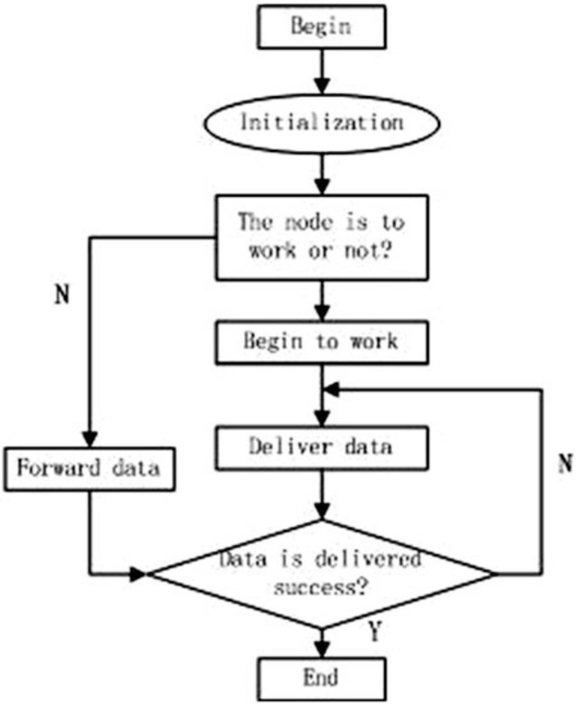
Design of Software for the Sensor Node

The software of the sensor node is used to collect the environmental information and process it simply. Then, Radio Frequency (RF) module transmits the processed data to the gateway nodes through a wireless system. In order to improve the energy efficiency of the nodes, the sensor node's working model is divided into three working states, i.e., dormant, awakened, and normal. In the dormant state, the processor stops working, and the RF module is in a low current-receiving state. After information has been received from the gateway node or neighboring nodes, the sensor node will judge whether it is the destination node or not. If it is, the sensor node will convert into the working state; if it is not, the sensor node will forward the information and return to the dormant state again. The flow diagram for the software that is used to control the sensor nodes is shown in Fig. 3.

Design of the Software for the Gateway Node

Gateway nodes have many functions, Such as creating a wireless network, receiving the data collected by the sensor nodes, setting the properties of the network nodes, transmitting data to the local monitoring center, receiving commands from the monitoring center and forwarding them the commands to control nodes. Figure 4 shows the design of the software for the gateway nodes.

Fig. 3 Flow diagram for the software



Hardware Module

Selection and Management of the Sensor Nodes

The hardware part of a sensor node includes the sensor module, a microprocessor, and a wireless communication module, the power module, and storage and display equipment.

For the microprocessor and the wireless communication module, the TI's system-on-a-chip: CC2430 WAS selected, because it can improve performance and meet the requirements of low cost and low power consumption when using the 2.4 GHz ISM band based on ZigBee. CC2430 has a structure in which ZigBee RF, memory and the microcontroller are integrated into a signal chip. It uses a MCU(8051) of eight bits, 32/64/128-kb programmable flash memory, 8-kb RAM, and it also contains an ADC, several Timers, AES128 collaborative processer, a 32-kHz timer of crystal oscillator of the sleep mode, Power on reset, Brown out detection and 21 programmable I/O pins. But the sensor nodes are made of various kinds of sensor chips. Sensor modules can sense temperature, humidity, and the concentration of carbon dioxide. Temperature and humidity sensors use the I2C bus digital sensor, SHT75. It has two major advantages, i.e., small size and low energy consumption, and its temperature range and accuracy are −40T $\sim$ 123T

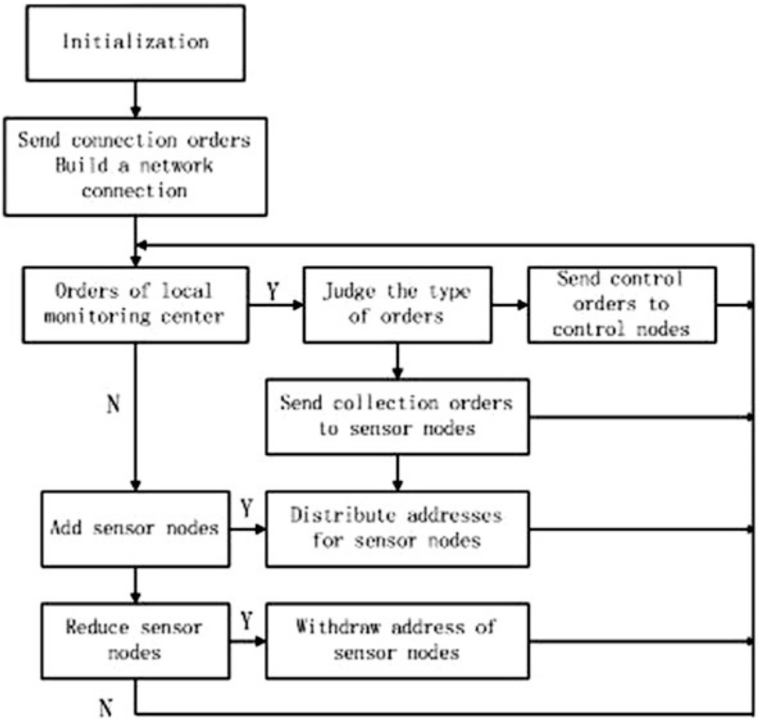

Initialization
Send connection orders Build a network connection
Orders of local monitoring center
Y
Judge the type of orders
Send control orders to control nodes
N
Send collection orders to sensor nodes
Add sensor nodes
Y
Distribute addresses for sensor nodes
Reduce sensor nodes
Y
Withdraw address of sensor nodes
N

Fig. 4 Software flow diagram of the gateway nodes

centigrade and ±0.3T centigrade, respectively, whereas the relative humidity range and accuracy are 0–100 % and ±1.8 TRH respectively.

The carbon dioxide sensor uses a metal oxide semiconductor CO_2 sensor, i.e., SB-AQ6A (The FIS's product). The analytical range of the SB-AQ6A for CO_2 is 400–3,000 ppm and it is used to control concentrations in the air conditioning and ventilation systems. Furthermore, it has the advantages of low cost, no maintenance, and a longer service life than optical analyzers.

Selection and Management of the Control Node

The hardware in the control node consists mainly of a microprocessor, a wireless communication module, the power module, and the driving module. The control node has the same type of micro-processer and wireless communication module as the sensor node. The power module uses DC power. The driver module consists mainly of a relay and a photo-electric coupler, and it controls the dehumidifier, air heater, and air cooler through the relay module when it receives orders from the monitoring center.

Selection and Management of the Gateway Node

The gateway node includes mainly a microprocessor, a wireless communication module, a power module, a storage module, RJ45 ethernet interface module, an RS232 string-line interface module, and a USB interface module.

Management of the Energy of the Nodes

In the research described in this paper, the sensor nodes used 3.3-V batteries to supply power. The service life of the sensor and the transmission distance are restricted by the capacity of the batteries, so power consumption is an important consideration [8]. Thus, a dormancy mechanism can be used to reduce the power requirements; this mechanism closes the wireless communication module and the data acquisition module to save energy when the sensor nodes have no information collection tasks and are not forwarding data for other nodes. In this dormancy mechanism, only a neighboring area of the sensor nodes is active when a sensor task occurs, and the active area will move along with the data transmitted to the gateway node, as this occurs, the active nodes in the last active area can return to the dormancy mode to save energy.

Application Prospects

Wireless sensor networks, combined with a variety of advanced technologies—provides a new approach for obtaining process information. Also, they can provide data that researchers need, automatically and in real time. In addition, they have no adverse impacts on the normal operation of the train and the convenience of this mode of travel. Furthermore, the monitoring data are relatively accurate.

So, wireless sensor networks are a feasible way to increase the use of digital processes and equipment to assess the conditions on a subway train, thereby improving the operational safety of the train as well as the comfort level of the passengers.

In this paper, we reported the development of an information monitoring platform based on a wireless sensor network to significantly improve the conditions associated with rail traffic. Thus, the proposed system could provide reference for the use of WSNs to enhance the comfort and safety of rail traffic in China.

Acknowledgments This research is supported by National Natural Science Foundation of China under Grant 61071076, Beijing Science and Technology Program under Grant Z121100007612003, the Beijing Municipal Natural Science Foundation under Grant 4132057.

References

1. Mo L, Yunhao L (2007) Underground structure monitoring with wireless sensor networks. In: Proceedings of 6th international symposium on information processing in sensor networks, 2007. IPSN 2007
2. Akyildiz IF et al (2002) A survey on sensor networks. Commun Mag IEEE 40(8):102–114
3. Ulema M (2004) Wireless sensor networks: architectures, protocols, and management. In: Network operations and management symposium, 2004. NOMS 2004. IEEE/IFIP
4. Chen CW, Wang Y (2008) Chain-type wireless sensor network for monitoring long range infrastructures: Architecture and protocols. Int J Distrib Sens Netw 4(4):287–314
5. Pereira V et al (2011) A taxonomy of wireless sensor networks with QoS. In: New technologies, mobility and security (NTMS), 2011 4th IFIP international conference
6. Sharma P, Bhadana P (2010) An effective approach for providing anonymity in wireless sensor network: detecting attacks and security measures. Int J Comput Sci Eng 1(5):1830–1835
7. Sikka P, Corke P, Overs L (2004) Wireless sensor devices for animal tracking and control. In: Local computer networks, 2004. 29th annual IEEE international conference
8. Lee DS, Liu YH, Lin CR (2012) A wireless sensor enabled by wireless power. Sensors 12(12):16116–16143

An Improved Social Network Analysis Method for Social Networks

Jongsoo Sohn, Daehyun Kang, Hansaem Park, Bok-Gyu Joo and In-Jeong Chung

Abstract Recently, Social Network Service (SNS) users are rapidly increasing, and Social Network Analysis (SNA) methods are used to analyze the structure of user relationship or messages in many fields. However, the SNA methods based on the shortest distance among nodes is time-consuming in measuring computation time. In order to solve this problem, we present a heuristic method for the shortest path search using SNS user graphs. Our proposed method consists of three steps. First, it sets a start node and a goal node in the Social Network (SN), which is represented by trees. Second, the goal node sets a temporary node starting from a skewed tree, if there is a goal node on a leaf node of the skewed tree. Finally, the betweenness and closeness centralities are computed with the heuristic shortest path search. For verification of the proposed method, we demonstrate an experimental analysis of betweenness centrality and closeness centrality, with 164,910 real data in an SNS. In the experimental results, the method shows that the computation time of betweenness centrality and closeness centrality is faster than the traditional method. This heuristic method can be used to analyze social phenomena and trends in many fields.

J. Sohn (✉)
Service Strategy Team, Visual Display, Samsung Electronics, Suwon, South Korea
e-mail: jongsoo.sohn@samsung.com

D. Kang · H. Park · I.-J. Chung
Department of Computer and Information Science, Korea University, Seoul, South Korea
e-mail: internetkbs@korea.ac.kr

H. Park
e-mail: park11232000@korea.ac.kr

I.-J. Chung
e-mail: chung@korea.ac.kr

B.-G. Joo
Department of Computer and Information Communications, Hongik University, Seoul, South Korea
e-mail: bkjoo@hongik.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_13, © Springer Science+Business Media Dordrecht 2014

Keywords Social network (SN) · Social network analysis (SNA) · Betweenness centrality · Closeness centrality · Heuristic approach

Introduction

Recently, online social network services are becoming popular with users, along with the expansion of Web 2.0-based services and the widespread use of smart devices. Online SNSs are online community services which enables users to communicate with each other, share information, and expand their human relationships [1]. In an SNS, each relation between users is represented by a simple graph, which consists of nodes and edges. As online SNS users are increasing rapidly, SNSs are actively utilized in enterprise marketing, analysis of social phenomena, trends, and so forth [2, 3].

Meanwhile, SNA is a way of analyzing social relationships among users in an SN. Through the SNA, it is possible to measure relationships between members, degree of intimacy, and intensity of connection, and to detect communities. The following are conventional SNA methods: degree centrality, betweenness centrality, and closeness centrality [4]. In the degree centrality analysis, the shortest path is not considered; however, it is used as a crucial factor in betweenness centrality, closeness centrality, and other SNA methods [4]. In previous works, the computation time was not time-consuming, due to the small size of the SN [5]. But, finding the path needs significant time to process data, since the number of nodes consists of online SNSs. For instance, if the number of nodes in an online SNS is n, the maximum number of its link is $n(n-1)/2$. This indicates that it is too expensive to analyze an SN; for example, if the number of nodes is 10,000, the number of links is 49,995,000.

Therefore, we propose a heuristic method for searching the shortest path among users in an SN graph. Moreover, we devise an enhanced method with addition of the best-first search, to reduce the computation time, and search the path rapidly, in an online SNS of huge size.

To verify the proposed method, we crawled 160,000 user IDs from online SNSs and constructed a graph out of them. Then, we compared this with previous methods, which are the best-first search and breadth-first search, in the time taken to search nodes and analyze SNSs. The suggested method took 240 s (7.4 times faster) to search nodes, where the breadth-first search took 1,781 s. Moreover, the method for SNA is 6.8 times and 1.8 times faster than the betweenness centrality analysis and closeness centrality analysis, respectively.

The suggested method shows the possibility of analyzing a large sized SN with better performance in time. Consequently, the method improves the efficiency of SNA and is used to determine social trends or phenomena.

Related Works

Social Network Analysis

Due to the popularity of Web 2.0 and the wide propagation of smart devices, there has been a diversity of research studies using SN [6, 7]. SNA represents the relationships between users in a graph. The most popular methods of SNA are degree centrality, closeness centrality, and betweenness centrality. Degree centrality is the index to figure out the importance in the whole SN, by measuring a node directly connected to other nodes [8]. Degree centrality computes the related degree of other users to a user v, when there is the user v of n users. It is represented by Eq. (1).

$$C_D(v) = \frac{deg(v)}{n-1} \tag{1}$$

Closeness centrality represents how closely related one user is to another user. If there are two users P_i and P_k of n users, this formula is represented by $C_C(P_k)$, as shown in Eq. (2).

$$C_C(P_k) = \left[\sum_{i=1}^{n} d(P_i, P_k)\right]^{-1} \tag{2}$$

In (2), the subscript C is an abbreviation of the word Closeness, $d(P_i, P_k)$ is the number of the shortest path from node i to node $k(i \neq k)$, and n is the number of users. Betweenness centrality is an index to measure how well a node performs as a mediator in the SN. It is computed by formula (3). In (3), the subscript B of C_B is Betweenness, and δ_{st} is the number of the shortest path between two nodes s and t. $\delta_{(v)}$ is the number of passing v in the path. If the betweenness centrality is large, it has an effect on the information flow in the SN.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\delta_{(v)}}{\delta_{st}} \tag{3}$$

The Shortest Path Search Method

Breadth-First Search (BFS) [8] and Depth-First Search (DFS) [9] are ways to transform a graph to state space, in order to explore the shortest path from the starting node to the goal node. BFS is a strategy to search nodes from nodes located at a close level, to nodes located at a distant level, in sequence. DFS is a recursive search method, which explores from the start node, to the node of distant level. DFS has less efficiency on a huge graph in that it needs to know how to search each node level.

The Best-First search is based on a heuristic method, to measure the shortest path between two nodes in graph. The Best-First search is an algorithm using a heuristic method, which is formalized by human experience. In [5], the authors proposed a method to estimate the shortest path in a large graph, using Monte Carlo Simulation. The proposed method in [5] measures the distance of two nodes, by comparing complete enumeration and Monte Carlo sampling in a large graph, which consists of 10,000 nodes. However, the method of [5] has a limitation on being applied to an SN, which graph changes frequently, because it requires the process of extracting a model. Another method in [10] suggested a self-organizing strategy, to detect the shortest path on complex networks. Likewise, there are the same shortcomings, as described in [5, 11].

Improved Social Network Analysis Method

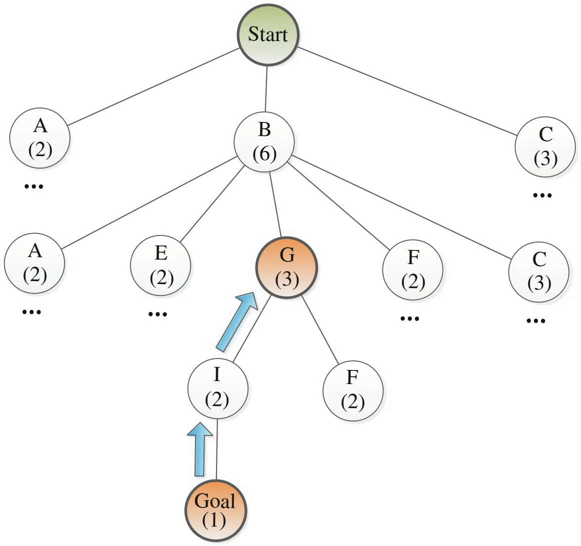
Closeness centrality and betweenness centrality, excepting degree centrality, in the typical SNA method are based on computation of the shortest path [12]. Thus, much time to compute closeness centrality and betweenness centrality is required in a large SN. Due to this problem, we came up with a method to add preprocessing steps with the Best-First search method, in order to compute the shortest path in an online SN. In an online SN, only a few users have a great number of connections; however, most other users have few connections. Hence, the users who have many connections perform a hub role in the SN [13]. Thus, finding users who have many connections in advance is desirable for the enhancement of search probability, since those are more influential on others in the SN. For this reason, the degree of node v_n is used as a heuristic evaluation function in a graph $G = (V, E)$, consisting of a set V of node v and set E of connections, when there are an existing n users, such as formula (4).

$$f(n) = \text{The number of degree}(v_n) \tag{4}$$

If a heuristic evaluation function is applied to the best-first search, such as formula (4), the shortest path is computed with comparison to a BFS. The heuristic evaluation function as shown in formula (1), however, does not rapidly increase path search performance, because its worst case occurs when nodes are skewed in state space. Thus, in this paper, the best-first search is utilized, after setting the root node of a skewed tree as a sub goal, if the search path contains a skewed tree.

Figures 1 and 2 show phrases of the proposed shortest path method. In Fig. 1, 'Start' and 'Goal' indicate a start node and a goal node, respectively. The values in parenthesis of each node are the evaluation value gained through formula (4) for each node. Before finding the path between 'Start' and 'Goal', a pertinent node is assigned in a search of the last part of a skewed tree, if the goal node is located in a

Fig. 1 Selecting a sub goal



leaf node of the skewed tree. Next, the best-first search with the number of degrees as evaluation function is used for finding the shortest path, as depicted in Fig. 2.

Table 1 is the proposed algorithm for searching the shortest path. The path can be detected quickly by preferential search of a hub user in large SN, when using the proposed shortest path search method. However, our method differs from complete enumeration, in terms of accuracy.

Fig. 2 Finding the shortest path

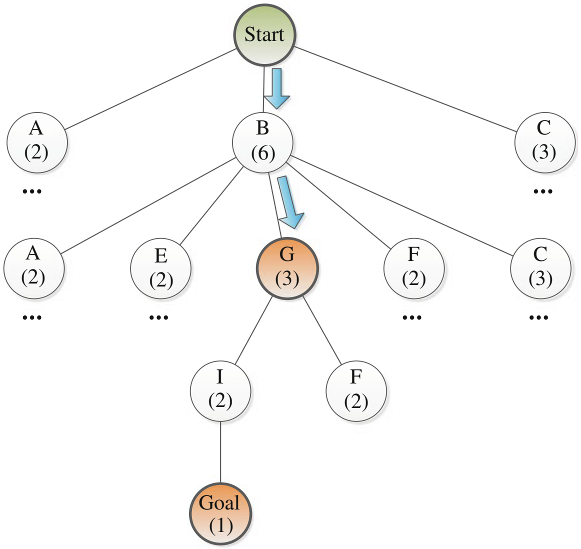


Table 1 An improved shortest path finding algorithm

```
Algorithm: Modified_Best_First_Search
Input: start_Node, goal_Node
Output: path

call module preprocessing(goal_Node)
open := [start_Node]
closed := [ ]
while open != [ ] do
 if there is no list in open then return fail
 move first node of open to closed, call it X
 if X == goal_Node then return path
 else if X == subgoal_Node then
     path = path + subPath
         return path
     else
         generate children of X
         for each children of X do
         case
            child is not on open or closed
               evaluate the child by heuristic function // in Equation (4)
               add the child to open
            child is already on open
               if the child was reached by a shorter path
               then give the state on open the shorter path
            child is already on closed
               if the child was reached by a shorter path
               then remove the state from closed and add
               the child to open
                put X on closed
                reorder states on open by evaluated value

module preprocessing(goal_Node) // Search and select a subgoal_Node if the
goal_Node is located in a terminal node of a skewed tree, as depicted in Fig. 1
        subPath := [ ]
        if the degree of parent == 2 then
            add the parent to subPath
            goal_Node = parent
            call module preprocessing(goal_Node)
                else
                  subgoal_Node = parent
                  return subgoal_Node
```

Evaluation

We conducted experiments to validate the SNA efficiency of the proposed shortest path search method. A computer with Intel Core2duo CPU, 2 GB RAM, and Java as the development toolkit, with a MySql database server, are used for the implementation.

First, we collected 164,910 SNS users for a dataset from Twitter,[1] as shown in Fig. 3. Most users have from one to five links; a few of them have more than five links. We converted the collected datasets into a relational database table, as

[1] http://www.twitter.com

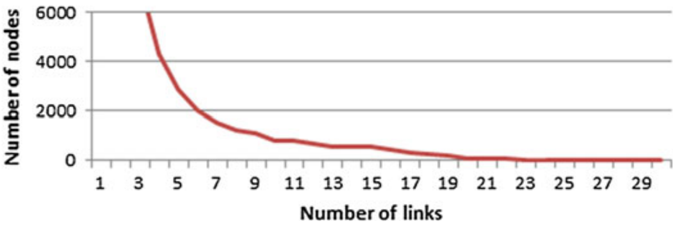


Fig. 3 Distribution of collected social network

demonstrated in Table 2. In Table 2, 'Id' is a unique identification number of each row, 'User' is a twitter user, and 'Following' is a user who follows the 'User'. For instance, if a user follows another ten users, then ten rows are inserted. Next, we designed a java program to accomplish the BFS algorithm and the proposed method, and then selected a pair of 100 arbitrary nodes, to compare the average search performance time.

We compared the proposed shortest path search method and BFS method on the SN, as described in Table 3. As the BFS method gradually searches from the first node to near-level node, it demonstrates the most accurate shortest path. In addition, as the suggested method searches nodes that have a lot of edges, it extracts the same or larger paths better, than those of the BFS method.

The BFS method always calculates the optimal shortest path, because it enumerates every possible path, to find the shortest one. On the other hand, the proposed method lists several paths selectively, using the heuristic evaluation function, so that it cannot always assure optimal values. Our proposed method can compute an average shortest path of about 80 %, compared to that of the BFS method. The method searches nodes less than those of the BFS method, so the average searching time is much faster, by about 7 times.

In addition, we calculated the betweenness centrality and closeness centrality, using performance comparison of the SN datasets, as depicted in Table 4. This shows the search results for both the BFS method and the proposed searching

Table 2 Social network database table

Id	User	Following
1	1,142,682	384,752
2	1,142,682	957,841
3	384,752	248,571
…	… omitted…	…

Table 3 Performance comparison of finding the shortest path

Algorithm	Mean path	Mean search	Mean time
BFS	2.45	83,069	1,781.69
Proposed	2.97	38,062	240.86

Table 4 Performance measurement of both the BFS and the proposed method

Algorithm	Criteria	BFS	Proposed	± Ratio
Betweenness centrality	Average time (s)	153,474	22,331	6.8 times
	Average number of searched nodes	255.33	181.36	−1.4 times
Closeness centrality	Average time (s)	109,336	59,418	1.8 times
	Average number of searched nodes	217.92	169.34	−1.3 times

method, to calculate the betweenness centrality and closeness centrality. In Table 4, the time to compute the betweenness centrality improved by about 6.8 times, but the number of searched nodes decreased by about 1.4 times. Besides, the time to compute closeness centrality improved by about 1.8 times, but the number of searched nodes decreased by about 1.3 times.

Conclusion

As web services are based on Web 2.0 and Social Web, online SNS users are gradually increased. At the same time, SNS posts share information between themselves, and impact on various fields. Then, many researchers in a variety of areas, such as Sociology, Economics, Politics, etc. have attempted to analyze SNSs. Since most real online SNSs consist of huge amounts of nodes, however, it is difficult for diverse SNA methods to be applied to real SNSs.

In this paper, we suggested a shortest path search method to improve the time performance for SNA, such as betweenness centrality, closeness centrality, and degree centrality. In other words, we presented a modified heuristic method, which is appropriate to online SNS. Though the method has less accuracy than BFS by about 80 %, it can compute closeness and betweenness centralities more than 7 times faster. Moreover, we can improve betweenness and closeness centrality analysis efficiency by 6.8 times and 1.8 times respectively.

Using our heuristic method, we can apply various SNA methods to large datasets. The method can also be used to determine social phenomena, user trends, and political traits in various fields, such as Politics, Sociology, Economics, etc.

References

1. Ellison NB, Steinfield C, Lampe C (2007) The benefits of facebook "friends:" social capital and college students' use of online social network sites. J Comput Mediated Commun 12(4):1143–1168
2. Kwak H, Lee C, Park HS, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, North Carolina, USA, ACM, pp 591–600
3. Sohn JS, Chung IJ (2013) Dynamic FOAF management method for social network in the social web environment. J Supercomput

4. Otte E, Rousseau R (2002) Social network analysis: a powerful strategy, also for the information sciences. J Inf Sci 28(6):441–453
5. Huh MH, Lee YG (2011) Applying Monte-Carlo method in social network analysis. Appl Stat Res 24(2):401–409
6. Cho ID, Kim NK (2011) Recommending core and connecting keywords of research area using social network and data mining techniques. J Intell Inf Syst 17(1):127–138
7. Kim HK, Choi IY, Ha KM, Kim JK (2010) Development of user based recommender system using social network for u-healthcare. J Intell Inf Syst 16(3):181–199
8. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):26–113
9. Hummon NP, Doreian P (1990) Computational methods for social network analysis. Soc Netw 12(4):273–288
10. Shen Y, Pei WJ, Wang K, Wang SP (2009) A self-organizing shortest path finding strategy on complex networks. Chin Phys B 18(9):3783
11. Nan D, Wu B, Pei X, Wang B, Xu L (2007) Community detection in large-scale social networks. In: Proceedings of joint 9th WEBKDD and 1st SNA-KDD
12. Ahmet ES, Kamer K, Erik S, Umit VC (2013) Incremental algorithms for network management and analysis based on closeness centrality. In: Proceedings of arXiv:1303.0422v1
13. Heer J, Boyd D (2005) Vizster: visualizing online social networks. Information visualization. In: Proceedings of IEEE symposium on INFOVIS

Part II
Special Session

Power-Saving Scheduling Algorithm for Wireless Sensor Networks

Ting-Chu Chi, Pin-Jui Chen, Wei-Yuan Chang, Kai-Chien Yang and Der-Jiunn Deng

Abstract Due to advances in communications technology in recent years, prompting more extensive application of wireless sensor networks. Such sensors are limited in battery energy supply and produce the energy hole problem, this even causes paralysis of part of the system. In the cluster architecture, burden of the cluster head is bound to become the energy consumption of the maximum point, so our method is focused on reducing energy consumption of the cluster head. To this end, we propose a new scheduling mechanism based on cluster architecture. In this mechanism, we use the "polling" method to make the cluster head have an absolutely effective data receiving. In addition, we also introduced the "sleeping" mechanism to ensure that the cluster head can achieve power saving under the premise of the most effective data receiving.

Keywords Power saving · Wireless sensor networks

T.-C. Chi · P.-J. Chen · W.-Y. Chang · K.-C. Yang · D.-J. Deng (✉)
Department of Computer Science and Information Engineering National Changhua,
University of Education, Taiwan, China
e-mail: djdeng@cc.ncue.edu.tw

T.-C. Chi
e-mail: a22533884@hotmail.com

P.-J. Chen
e-mail: jeffery2936@hotmail.com

W.-Y. Chang
e-mail: v123582@gmail.com

K.-C. Yang
e-mail: as5374942@hotmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_14, © Springer Science+Business Media Dordrecht 2014

Introduction

In wireless sensor networks, a large number of sensor nodes need to sense data sent to the data collection point, therefore, many protocol for data transmission were been developed. And the easiest way is the direct transmission protocol [1]. This protocol need to consider the distance from sensing node and the base station. Energy consumption will be generated when sensor nodes send data directly to the data collection point. The distance between sense node and data collection point will affect the energy for sensing nodes to transmit data, thus limits the scope of transmission. In order to increase the scope of transmission, a way to transfer information by node relaying in multiple-hop was generated, which is called the multi-hop transmission [2]. Although these two transmission purposes are the same to send data to the data collection point, but the latter can increase the effective sensing scope.

Multiple-hop architecture can be subdivided into cluster-based architecture, chain-based architecture and tree-based architecture to collect data. In multiple-hop transmission method, the nodes far from data collection point will send data to the nodes closer to the base station first, this way can avoid a large amount of energy-consuming because of the transmission from distant distance [3]. However, the nodes closer to the base station need to relay the data from the other nodes very frequently, therefore, the energy consumption will be particularly high. When the loss of these nodes happened, the transmission distance of other nodes will be increased, and the energy consumption will also be increased, these will cause the fail of data transmission. As a result, it will lead to the partial sensing node lose effectiveness, even more, the energy hole problem will be generated. As the following diagram shown in Fig. 1, the energy consumption

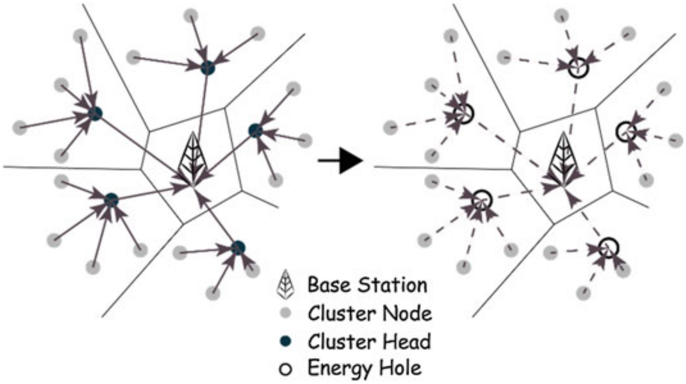


Fig. 1 Energy hole

of the overall system will be extremely uneven, and leading to the shortening of life-cycle [4].

The energy hole problem is difficult to be completely avoided, only to defer the occurrence of this problem. Thus, as much as possible to make the energy consumption of each node be balanced to avoid the energy hole problem casing by the loss of partial sensing nodes. That is, to extend the life cycle of the system by making the energy be balanced [5].

In this paper, we use the cluster's infrastructure architecture. In this architecture, cluster head usually become the point which has the heaviest energy load, and it is easily to cause becoming of the energy hole. So we mainly investigate the problem of the energy consumption of the cluster head. We take a particular scheduling method to make cluster head can achieve the optimal power saving schedule under the premise that satisfying the time tolerance.

Networkmodel

We use cluster-based network architecture, using the homogeneous sensor nodes, but divide them into two classes, a cluster head and several Cluster-Nodes, and doing operation by taking the architecture shown in Fig. 2.

Cluster is a circle, the cluster head as the center, and the distance R is the radius of the circle. The deployment of the nodes is a topology formed by taking random uniform distribution.

Propose Scheme

Initial Phase

The cluster head must tell other nodes in the network that it is the cluster head. Cluster head broadcasts an initial message (Ini-msg) to all nodes in the cluster at

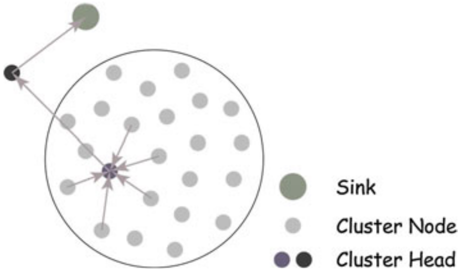


Fig. 2 Cluster-based network architecture

first. The Ini-msg contains an ID of the cluster head and some environmental parameters, such as sample rate, channel rate, etc. After each non-cluster-head node receives the Ini-msg, it transmits a Join-req (join-request) message to ask for joining the cluster by using CSMA/CA protocol. After the cluster head receives the Join-req message, it will base on the relevant environmental parameters to decide whether to accept this request. Once the non-cluster-head node is joined, the cluster head will create a corresponding token buffer.

Running Phase

When the systems are in this stage, it can be divided into four mechanisms:

- **Token Buffer Mechanism**

The cluster head establish token buffer, token buffer products token over sample rate, to let it know that it should receive the data from corresponding node.

- **Detection Mechanism**

Cluster head scans every Token buffer. Once finding the Token, it will add node ID into the Poll list and start scanning again from the first token buffer.

- **Polling Mechanism**

If Poll list is empty, cluster head will keep in sleep mode. If Poll list is not empty, cluster head will switch to active mode, and sent a polling packet to the corresponding cluster node and the node can know it should deliver the data to cluster-head. When cluster node received the polling packet it started to deliver the data, until cluster head reply a power saving poll (PS-Poll) packet and remove the node ID from the Polling list. After the completion of action system will restart Polling Mechanism.

- **Node Processing**

Cluster node keeps in sleep mode and sensing data. When node sensed data, the device will switch into active mode, and receiving the Polling packet from cluster head. Once cluster node receives PS-Polling packet, the node switches to sleep mode again.

Pseudo code

```
Function Polling_List_Creating(){
   WHILE (TRUE){
     head scans all nodes;
     IF (scan == token){
         add nodeID to Polling list;
         pointer move back to the first buffer;}
     ELSE{scan again from the first buffer to the last; }}
 scan again from the first buffer to the last;
}

Function Polling_Sending(){
   WHILE (TRUE){
    IF (Polling list != empty){
       head turns to active mode;
       send polling pkt to node i;}
    ELSE { head turns to sleeping mode; }}
}
```

Simulation

We refer to [6] and [7], our simulation environment is built in a circle. Other parameters are listed in the following (Table 1):

In Fig. 3, we consider that the energy consumption in our method will dynamically increase with the increasing of the node. In this way, we can reduce unnecessary waste by according to the amount of nodes. When there is no node needs to transmit, the head will also turn into sleep mode to conserve battery power. In addition, our approach will no longer receive data in the time of the upper bound of nodes in the system. The reason is that the information are time out, even if it receives is for naught, that is, to reduce energy consumption by avoiding unnecessary receiving. Because of our sleeping mechanism and the filter

Table 1 The variables of simulation

Parameters	
Sample period	30,000 unit time
Poll time/Ack time/data time	5/1/50 unit time
Tolerable time of node	23,000 unit time
Energy consumption of sending/receiving/sleeping	280/204/14 mA
Head power capacity	1.31072 * 10^{11} m Ah
Channel rate	2 Mbps

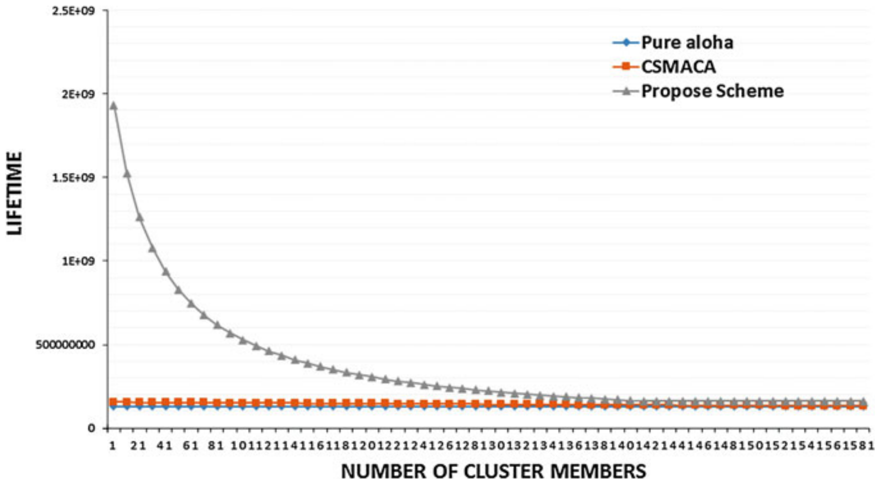


Fig. 3 Simulation of lifetime

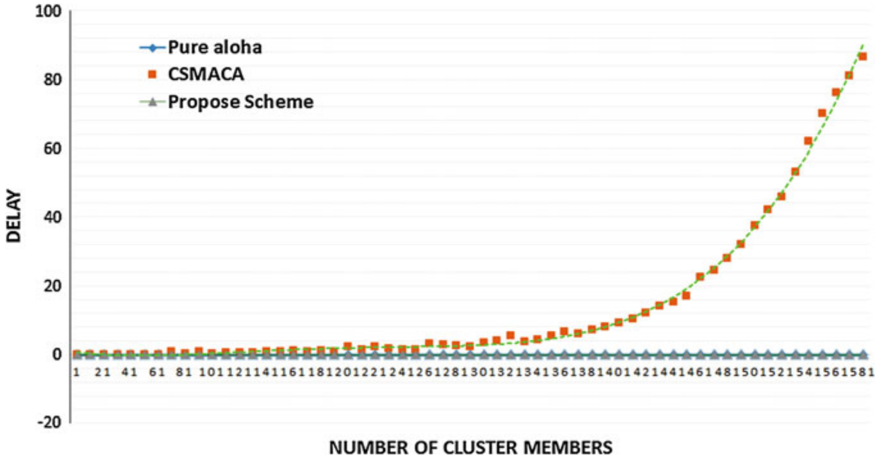


Fig. 4 Simulation of delay

mechanism in the initial step, the power saving is achieved, and the lifetime of the cluster head is also be prolonged.

Figure 4 shows the simulation of delay, the vertical axis represents the delay time of the cluster, and the horizontal axis represents the number of cluster. Because the data transmitting and receiving in cluster in our method will follow the schedule by using our "polling" method to make the cluster head have an absolutely effective data receiving, so that there is no delay in the cluster.

Conclusion

Burden of the cluster head will cause the cluster head more quickly lose energy, and this makes the cluster head naturally become the energy hole location because of its maximum energy consumption in the cluster. Thus, to reduce the energy consumption of the cluster head to be an important issue. And the most important part is to prolong the cluster head's life cycle in which prioritization is key to optimizing the overall performance of the cluster. The "polling" method to make the cluster head has an absolutely effective data receiving. In other hand, the "sleeping" mechanism to ensure that the cluster can provide the most effective data receiving under the premise of saving more power of the cluster.

References

1. Muruganathan SD, Ma DCF, Bhasin RI, Fapojuwo AO (2005) A centralized energy-efficient routing protocol for wireless sensor networks. IEEE Commun Mag 43:S8–S13
2. Noori M, Ardakani M (2008) Characterizing the traffic distribution in linear wireless sensor networks. IEEE Commun Lett 12(8):554–556
3. Guo P, Jiang T, Zhang K, Chen HH (2009) Clustering algorithm in initialization of multi-hop wireless sensor networks. IEEE Trans Wireless Commun 8(12):5713–5717
4. Dietrich I, Dressler F (2009) On the lifetime of wireless sensor networks. ACM Trans Sens Netw 5(1):Article 5
5. Li J, Mohapatra P (2005) An analytical model for the energy hole problem in many-to-one sensor networks. In: Proceedings of IEEE 62nd vehicular technology conference, vol 4, pp 2721–2725
6. Chen B, Jamieson K, Balakrishnan H, Morris R (2002) Span: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks. ACM Wireless Netw 8(5):481–494
7. Bouabdallah F, Bouabdallah N, Boutaba R (2009) On balancing energy consumption in wireless sensor networks. IEEE Trans Veh Technol 58(6):2909–2924

Localization Algorithm for Wireless Sensor Networks

Yin-Chun Chen, Der-Jiunn Deng and Yeong-Sheng Chen

Abstract In recent years, many localization algorithms are proposed for wireless sensor networks because that is crucial to identifying the accurate positions of sensor nodes. This study proposes an analytic localization algorithm by utilizing radical centers. Assume that a target node can measure its distances to four or more anchor nodes. By picking four distance measurements to four anchor nodes, a radical center is computed and treated as the target node location. To further improve and fuse these estimations, effective filtering mechanisms are then proposed to filter out the improper estimations. Afterwards, the remaining radical centers are averaged, and the solution is the final estimation of the target node location. The location errors of the proposed method and the conventional Minimum Mean Square Error method (MMSE) are analytically compared. Extensive computer simulations were carried out and the results verify the advantage of the proposed location algorithm over the MMSE approach.

Keywords: Wireless sensor networks · Localization · Radical centers · Minimum mean square error

Y.-C. Chen · D.-J. Deng (✉)
Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua, Taiwan
e-mail: djdeng@cc.ncue.edu.tw

Y.-C. Chen
e-mail: fpwking32@hotmail.com

Y.-S. Chen
Department of Computer Science, National Taipei University of Education, Changhua, Taiwan
e-mail: yschen@tea.ntue.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_15, © Springer Science+Business Media Dordrecht 2014

Introduction

Wireless sensor networks (WSNs) have great potential in a lot of control and monitor applications such as data collection, environment observation, and battlefield surveillance, and so on. Since most sensor nodes are randomly deployed without the knowledge of their positions, localization of sensor nodes is an essential issue for the operation, management, and applications of WSNs.

A lot of researchers have proposed many different solutions for the localization problem in WSNs [1–8]. Minimum Mean Square Error (MMSE) estimation method [1, 2] uses least squares solution to estimate the position of the target node. It has been widely studied for 2D localization in wireless sensor networks. However present network environment and technologies demand 3D localization [3–5]. In this paper, a novel range-based localization algorithm in wireless sensor networks is developed. Assume that the target node can measure its distances to four or more anchor nodes. We propose a simple and efficient mechanism for deriving better location accuracy based on the existing technologies for distance measurements using the conventional trilateration approach. The contribution of this study is that based on the existing technologies for distance measurements and without any extra hardware cost, the proposed mechanism provides an efficient algorithm for localization with better accuracy.

Location Estimation Using Trilateration

Trilateration is a common localization algorithm to identify the position of target node by using similar geometric concept of triangulation [9, 10]. Let A(x_A, y_A, z_A), B(x_B, y_B, z_B), C(x_C, y_C, z_C) and D(x_D, y_D, z_D) denote four anchor nodes. The actual distances from the target blind node T(x, y, z) to A, B, C and D are denoted as d_A, d_B, d_C and d_D; whereas, the estimated distances from T to A, B, C and D are denoted as e_A, e_B, e_C and e_D.

Consider node A. Equation $(x - x_A)^2 + (y - y_A)^2 + (z - z_A)^2 = d_A^2$ represents the sphere with center (x_A, y_A, z_A) and radius d_A. Ideally, the target node T(x, y, z) must be a point on the spherical surface. However, in practical implementation, due to the measurement errors, only the estimated distances (e_A, e_B, e_C and e_D) can be derived. Similarly, after also considering nodes B, C and D, we have the following system of equations.

$$\begin{cases} (x - x_A)^2 + (y - y_A)^2 + (z - z_A)^2 = e_A^2 \\ (x - x_B)^2 + (y - y_B)^2 + (z - z_B)^2 = e_B^2 \\ (x - x_C)^2 + (y - y_C)^2 + (z - z_C)^2 = e_C^2 \\ (x - x_D)^2 + (y - y_D)^2 + (z - z_D)^2 = e_D^2 \end{cases} \tag{1}$$

In the realistic case, there will be no pair of coordinates (x, y, z) satisfying (1). To tackle this problem, based on the theory of the radical centers of four spheres, an efficient analytic solution to the localization problem that can be formulated as (1) is proposed as described as follows.

Localization Utilizing Radical Centers

Related Definitions and Theorems

Definition 1 (*Power of a point*) The power of a point X with respect to a sphere with center O and radius r is defined as $(XO)^2 - r^2$ [12].

Definition 2 (*Radical Plane*) The radical plane of two spheres is the locus of points that have the same power with respect to both spheres [12].

Theorem 1 *For two spheres* A *and* B *whose centers and radii are* $(x_\mathrm{A}, y_\mathrm{A}, z_A)$, $(x_\mathrm{B}, y_\mathrm{B}, z_B)$, d_A *and* d_B, *respectively, the equation of the radical plane of spheres* A *and* B *is*:

$$\begin{aligned} & x_A^2 - x_B^2 + y_A^2 - y_B^2 + z_A^2 - z_B^2 - x(2x_A - 2x_B) - y(2y_A - 2y_B) - z(2z_A - 2z_B) \\ & \quad = d_A^2 - d_B^2 \end{aligned}$$

Theorem 2 *The radical planes of four spheres (no two of them are concentric) are concurrent or parallel or coincident* [13, 14]. *The point of concurrence is called the 'radical center' of the four spheres. (For brevity's sake, the proof is omitted.)*

Computation and Selection of the Radical Center

In practical implementation, the coordinates of the radical center can be easily derived by using Cramer's Rule [11], the solution [i.e., the coordinates of the radical center (x', y', z')], is

$$\begin{cases} x' = \frac{1}{2} \times \dfrac{\begin{vmatrix} \left(e_A^2 - e_B^2 - x_A^2 + x_B^2 - y_A^2 + y_B^2 - z_A^2 + z_B^2\right) & (y_B - y_A) & (z_B - z_A) \\ \left(e_B^2 - e_C^2 - x_B^2 + x_C^2 - y_B^2 + y_C^2 - z_B^2 + z_C^2\right) & (y_C - y_B) & (z_C - z_B) \\ \left(e_C^2 - e_D^2 - x_C^2 + x_D^2 - y_C^2 + y_D^2 - z_C^2 + z_D^2\right) & (y_D - y_C) & (z_D - z_C) \end{vmatrix}}{K} \\ y' = \frac{1}{2} \times \dfrac{\begin{vmatrix} (x_B - x_A) & \left(e_A^2 - e_B^2 - x_A^2 + x_B^2 - y_A^2 + y_B^2 - z_A^2 + z_B^2\right) & (z_B - z_A) \\ (x_C - x_B) & \left(e_B^2 - e_C^2 - x_B^2 + x_C^2 - y_B^2 + y_C^2 - z_B^2 + z_C^2\right) & (z_C - z_B) \\ (x_D - x_C) & \left(e_C^2 - e_D^2 - x_C^2 + x_D^2 - y_C^2 + y_D^2 - z_C^2 + z_D^2\right) & (z_D - z_C) \end{vmatrix}}{K} \\ z' = \frac{1}{2} \times \dfrac{\begin{vmatrix} (x_B - x_A) & (y_B - y_A) & \left(e_A^2 - e_B^2 - x_A^2 + x_B^2 - y_A^2 + y_B^2 - z_A^2 + z_B^2\right) \\ (x_C - x_B) & (y_C - y_B) & \left(e_B^2 - e_C^2 - x_B^2 + x_C^2 - y_B^2 + y_C^2 - z_B^2 + z_C^2\right) \\ (x_D - x_C) & (y_D - y_C) & \left(e_C^2 - e_D^2 - x_C^2 + x_D^2 - y_C^2 + y_D^2 - z_C^2 + z_D^2\right) \end{vmatrix}}{K} \end{cases} \tag{2}$$

Definition 3 (*Valid set of four anchor nodes*) For four anchor nodes, if their radical center is not at a point of infinity or no intersection, the set of these four anchor nodes is called *valid.*

Theorem 3 *Four anchor nodes* A(x_A, y_A, z_A), B(x_B, y_B, z_B), C(x_C, y_C, z_C) *and* D(x_D, y_D, z_D) *form a valid set iff* $K = \begin{vmatrix} (x_B - x_A) & (y_B - y_A) & (z_B - z_A) \\ (x_C - x_B) & (y_C - y_B) & (z_C - z_B) \\ (x_D - x_C) & (y_D - y_C) & (z_D - z_C) \end{vmatrix} \neq 0$

(For brevity's sake, the proof is omitted.) Theorem 3 provides simple and useful rules to pick proper anchor nodes for computing the radical center.

Further Investigation of the Errors of the Radical Center

To analyze the error between the radical center and actual location of the target node, we can check the values of $x - x'$, $y - y'$ and $z - z'$. Let e and d respectively denote the estimated and actual distance from the target node to an anchor node. Thus, we have $e = d + \varepsilon$, where ε is the measurement error. In this study, it is assumed that $\varepsilon = \rho d$, where ρ is called the *measurement error coefficient*. That is, the inherent location error of the radical center is formulated as (3), (4) and (5).

$$\begin{aligned} x - x' = & \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_A^2 - e_B^2\right)[(y_C - y_B)(z_D - z_C) - (y_D - y_C)(z_C - z_B)]}{2K} \\ & + \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_B^2 - e_C^2\right)[(y_D - y_C)(z_B - z_A) - (y_B - y_A)(z_D - z_C)]}{2K} \\ & + \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_C^2 - e_D^2\right)[(y_B - y_A)(z_C - z_B) - (y_C - y_B)(z_B - z_A)]}{2K} \end{aligned} \tag{3}$$

$$\begin{aligned} y - y' = & \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_A^2 - e_B^2\right)[(z_C - z_B)(x_D - x_C) - (z_D - z_C)(x_C - x_B)]}{2K} \\ & + \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_B^2 - e_C^2\right)[(z_D - z_C)(x_B - x_A) - (z_B - z_A)(x_D - x_C)]}{2K} \\ & + \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_C^2 - e_D^2\right)[(z_B - z_A)(x_C - x_B) - (z_C - z_B)(x_B - x_A)]}{2K} \end{aligned} \tag{4}$$

$$
\begin{aligned}
z - z' = {} & \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_A^2 - e_B^2\right)\left[(y_C - y_B)(x_D - x_C) - (y_D - y_C)(x_C - x_B)\right]}{2K} \\
& + \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_B^2 - e_C^2\right)\left[(y_D - y_C)(x_B - x_A) - (y_B - y_A)(x_D - x_C)\right]}{2K} \\
& + \frac{\left(\frac{1}{(1+\rho)^2} - 1\right)\left(e_C^2 - e_D^2\right)\left[(y_B - y_A)(x_C - x_B) - (y_C - y_B)(x_B - x_A)\right]}{2K}
\end{aligned}
\tag{5}
$$

Filtering Mechanism

In (3), (4) and (5), it is likely that the smaller the value of the denominator K is, the larger the error $(x - x',\ y - y'$ and $z - z')$ will be. Thus, K can be taken as a metric for filtering out the improper selections of anchor nodes. We utilize K by sorting all the estimated points according to its K value, and then some certain percentage of the estimated points with small K values is filtered out. This mechanism is called K-filtering.

Comparisons with MMSE and Simulation Results

With the MMSE method, we also check the values of $(x - x',\ y - y'$ and $z - z')$ as follows. Let $P = (x_1 - x_n)^2 + \cdots + (x_{n-1} - x_n)^2$, $Q = (x_1 - x_n)(y_1 - y_n) + \cdots + (x_{n-1} - x_n)(y_{n-1} - y_n)$, $R = (x_1 - x_n)(z_1 - z_n) + \cdots + (x_{n-1} - x_n)(z_{n-1} - z_n)$, $S = (y_1 - y_n)^2 + \cdots + (y_{n-1} - y_n)^2$, $T = (y_1 - y_n)(z_1 - z_n) + \cdots + (y_{n-1} - y_n)(z_{n-1} - z_n)$, $U = (z_1 - z_n)^2 + \cdots + (z_{n-1} - z_n)^2$, we have

$$
\begin{bmatrix} x - x' \\ y - y' \\ z - z' \end{bmatrix} = \frac{-1 * \left(\frac{1}{(1+\rho)^2} - 1\right)}{2\ *\ (PSU + 2\ *\ QTR - R^2S - Q^2U - T^2P)}
* \begin{bmatrix}
[(SU - T^2)(x_1 - x_n) + (RT - QU)(y_1 - y_n) + (QT - RS)(z_1 - z_n)](e_1^2 - e_n^2) + \cdots + [(SU - T^2)(x_{n-1} - x_n) + (RT - QU)(y_{n-1} - y_n) + (QT - RS)(z_{n-1} - z_n)](e_{n-1}^2 - e_n^2) \\
[(RT - QU)(x_1 - x_n) + (PU - R^2)(y_1 - y_n) + (RQ - PT)(z_1 - z_n)](e_1^2 - e_n^2) + \cdots + [(RT - QU)(x_{n-1} - x_n) + (PU - R^2)(y_{n-1} - y_n) + (RQ - PT)(z_{n-1} - z_n)](e_{n-1}^2 - e_n^2) \\
[(QT - RS)(x_1 - x_n) + (RQ - PT)(y_1 - y_n) + (PS - Q^2)(z_1 - z_n)](e_1^2 - e_n^2) + \cdots + [(QT - RS)(x_{n-1} - x_n) + (RQ - PT)(y_{n-1} - y_n) + (PS - Q^2)(z_{n-1} - z_n)](e_{n-1}^2 - e_n^2)
\end{bmatrix}
\tag{6}
$$

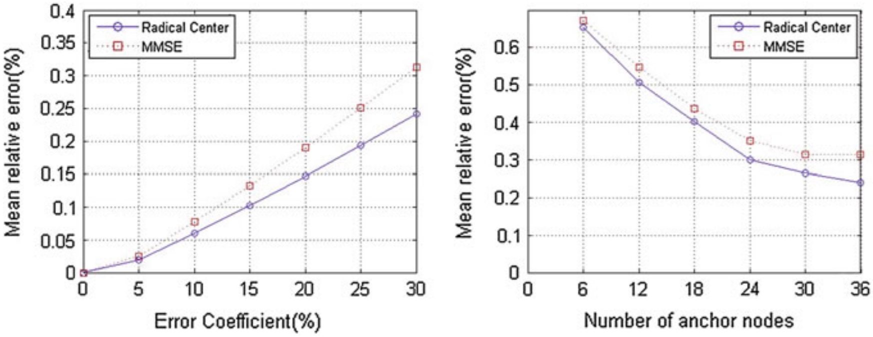


Fig. 1 Filtering percentage = 10 %

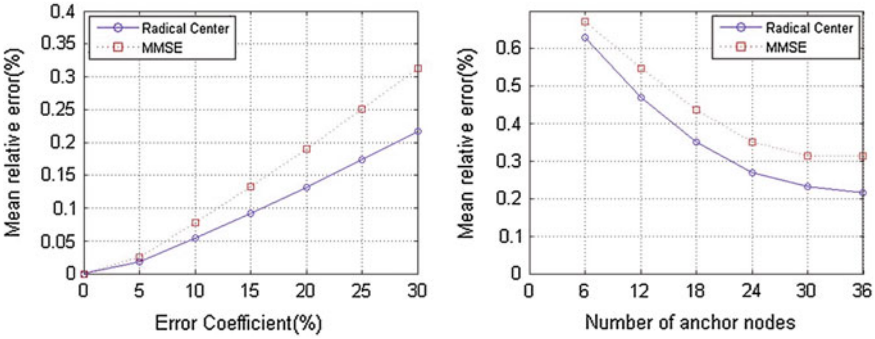


Fig. 2 Filtering percentage = 30 %

From (6), we can see that with MMSE, the location error is the summation of n-1 items. However, with the proposed algorithm, the estimated error, is composed of only three items. Thus, it is easy to reason that the resulting location error from the MMSE approach will be larger than the proposed algorithm. In the simulation environment, there is a target node and 36 anchor nodes and they are randomly placed. The distance measurement error coefficients are set to be 5, 10, 15, 20, 25 or 30 %. The filtering percentages in the filtering process are set to be 10, 20, or 30 %. We have conducted extensive simulations to look into the effectiveness of the proposed approach. For brevity's sake, parts of the simulation results are described here (Figs. 1 and 2).

As shown in Figures, the proposed approach always outperforms the MMSE method in accuracy and efficiency and also has better location accuracy than the MMSE in different number of anchor nodes.

Conclusions

Localization is one of the key issues in WSNs. We propose a novel range-based localization algorithm by utilizing radical centers. With our proposed algorithms, the target node computes the radical centers for location estimation with any four anchor nodes. Then, it can effectively filter out the improper estimations (radical centers) with the proposed filtering mechanisms so as to enhance the location accuracy. The advantages of the proposed algorithm over the conventional MMSE approach have been analytically analyzed and compared; and the conduced simulations demonstrated that the proposed algorithm effectively derive better location accuracy than the MMSE approach.

References

1. Wan J, Yu N, Feng R, Wu Y, Su C (2009) Localization refinement for wireless sensor networks. Comput Commun 32:1515–1524
2. Qi Y, Kobayashi H, Suda H (2006) Analysis of wireless geolocation in a non-line-of-sight environment. IEEE Trans Wireless Commun 5:672–681
3. Kuruoglu S, Erol M, Oktug S (2009) Three dimensional localization in wireless sensor networks using the adapted multi-lateration technique considering range measurement errors. In: Proceedings of IEEE GLOBECOM, pp 1–5
4. Zhang Y, Liu S, Jia Z (2012) Localization using joint distance and angle information for 3D wireless sensor networks. IEEE Commun Lett 16:809–811
5. Davis JG, Sloan R, Peyton AJ (2011) A three-dimensional positioning algorithm for networked wireless sensors. IEEE Trans Instrum Meas 60:1423–1432
6. Barsocchi P, Lenzi S, Chessa S, Giunta G (2009) A novel approach to indoor RSSI localization by automatic calibration of the wireless propagation model. In: Proceedings of IEEE 69th vehicular technology conference, pp 1–5
7. Gracioli A, Fröhlich A, Pires RP, Wanner L (2011) Evaluation of an RSSI-based location algorithm for wireless sensor networks. IEEE Latin Am Trans 9:830–835
8. Niculescu D, Nath B (2003) Ad hoc positioning system (APS) using AOA. In: Proceedings of 22nd annual joint conference of the IEEE computer and communications societies (INFOCOM 2003), vol 22, pp 1734–1743
9. Evrendilek C, Akcan H (2011) On the complexity of trilateration with noisy range measurements. IEEE Commun Lett 15:1097–1099
10. Yang Z, Liu Y, Li XY (2010) Beyond trilateration: on the localizability of wireless ad hoc networks. IEEE/ACM Trans Netw 18:1806–1814
11. Anton H, Rorres C (2010) Elementary linear algebra: applications version, 10th edn. Wiley, New York
12. Coxeter HSM, Greitzer SL (1967) Geometry revisited. The Mathematical Association of America, Washington, DC, pp 27–34
13. Dorrie H (1965) 100 great problems of elementary mathematics. Dover Publications, New York, p 153
14. Weisstein EW (2003) CRC concise encyclopedia of mathematics, 2nd edn. CRC Press, USA, p 502

Exploring Community Structures by Comparing Group Characteristics

Guanling Lee, Chia-Jung Chang and Sheng-Lung Peng

Abstract In recent years, more and more researchers devoted to identifying community structure in social networks. The characteristics of the social network are analyzed by clustering the social network users according to user's relationships. However, the users of current popular social networks such as LiveJournal and Flickr, can join to or create the communities according to their interests. Instead of grouping the users according to the cluster strategies which are wildly used in previous works, the purpose of the paper is to explore the structures and characteristics of the social networks according to the community the users actually joined. Moreover, we experiment on four real datasets, LiveJournal, Flickr, Orkut and Youtube, to analyze the characteristics hidden behind the social networks.

Keywords Social network · Community structure · Cluster

Introduction

Accompanying the growth of social networks, such as Facebook, more and more users rely on the social networks to communicate with their friends and share their daily life. In order to understand the users' behavior, identifying the characteristics behind the connections of the social network users become an important research topic. In [1, 2], the problem of how to measure the effect of information dissemination in a social network is discussed. By analyzing the number of users that will be affected by a certain user, the roles of *leader* and *follower* in a social network are defined in [3]. Moreover, in [4–6], the problem of how to partition the

G. Lee (✉) · C.-J. Chang · S.-L. Peng
Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien 974, Taiwan, Republic of China
e-mail: guanling@mail.ndhu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_16, © Springer Science+Business Media Dordrecht 2014

users into the clusters is discussed. In previous works, the users are categorized into the same cluster if the degrees of intra connection is high and inter connection is low. However, in current popular social network, the users have the right to join to or create his/her own *communities*. Therefore, the purpose of the paper is to explore the structures and characteristics of the communities that the users actually joined, and use the characteristics to represent the relationship among the social network users.

The datasets we used to analyze the community structure are collected from four popular social networks, LiveJournal, Flickr, Orkut and Youtube. Moreover, we propose several measurements to model the characteristics and structures of the communities. The paper is organized as follows. The problem and three measurements are presented in section Problem Definition. We experiment on four real datasets and discuss the results in section Experimental Results. And finally, Conclusion concludes the work.

Problem Definition

As discussed in [7], a social network is defined as an *interaction Graph* $G = (V, E)$, where V denotes the vertex set of G and represents the social network users, E denotes the edge set and e_{ij} is contained in E if nodes i and j are friends in the network. In previous works, the users (nodes in G) are partitioned into clusters according to the connection degree among them. That is, the nodes are grouped into a cluster if the intra similarity is much larger than inter similarity. And similarity is usually measured by a function of the number of connections among the nodes. However, as mentioned above, in current popular social network, the users have the right to join to or create his/her own *communities*. Moreover, a user can join many communities according to his/her interests. That is, differ to the cluster concept proposed in previous works, in the real community model, a node can belong to many communities and a community is not necessary to have a tight connectivity. Therefore, the measurements for modeling a cluster are not suitable to measure the characteristics of the community. In the following, we propose three ideas for measuring the characteristics of a community.

Center of the community: If a user has many friends in the community, then the information posed by him would be noticed by many users in the same community. Therefore, we define the center of a community is the node whose number of connections in the community is much larger than that of other nodes in the same community. Therefore, the *center degree* of $node_i$ in community A, denoted as c_i^A, can be measured by the following equation.

$$c_i^A = \frac{\mathrm{deg}ree_i^A}{|N_A| - 1} \tag{1}$$

In the equation, $degree_i^A$ denotes the number of connections (friends) of $node_i$ in community A and $|N_A|$ is the number of members in community A. Therefore, $|N_A| - 1$ is the maximum connections of a node in community A. When c_i^A is larger than a predefined threshold α, $node_i$ is said to be a center of community A. Moreover, the center set of community A, denoted as C_A, is the collection of centers in community A.

Intra connection degree: We propose the idea of *intra connection degree* to measure the inner structure of a community. And the complete connections concept is adapted to measure it. The intra connection degree of community A which is denoted by $Intra_A$, is measured by the following equation.

$$Intra_A = \frac{|E_A|}{(|N_A| \times (|N_A| - 1))/2} \tag{2}$$

In the equation, $|E_A|$ denotes the number of edges contain in A. The denominator indicates the maximum number of edges that community A can have. Therefore, the larger the value of $Intra_A$, the tighter the members in community A is.

Inter connection degree: We use the idea of centers of community to measure the inter connection degree. The basic concept is that if a center of community A is also the center of other community, then community A has a strong connection to the other community. And we denote the node which is the center of at least two communities as OC. The inter connection degree of community A, denoted as $Inter_A$, is measured by

$$Inter_A = \frac{|\{n|\, n \in C_A\, and\, n\, is\, OC\}|}{|C_A|} \tag{3}$$

The numerator of the equation is the number of centers which is also an OC in community A. A large $Inter_A$ indicates the proportion of OC in C_A is high. And therefore, community A has a strong connection to other communities.

In next section, a set of experiment is performed on real datasets. And by comparing the proposed measurements to the characteristics of the dataset, a thorough discussion is made.

Experimental Results

Four real datasets of online social network platforms, LiveJournal, Orkut, Flickr and Youtube, are collected from [8] to perform the experiment. LiveJournal was built by Brad and Fitzpatrick in 1999, and is a vibrant global social media platform where users share common passions and interests. Orkut is a social network service provided by Google. Users can build their own virtual social links in the internet by using the platform. Flickr was developed by Ludicorp company. In the platform, users can upload and share their pictures. Moreover, users can create *tags* for the pictures to ease the browsing process. YouTube is a video-sharing website, on

Fig. 1 The average intra connection degree of small size communities

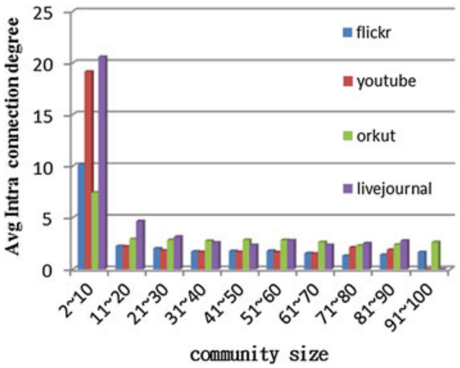
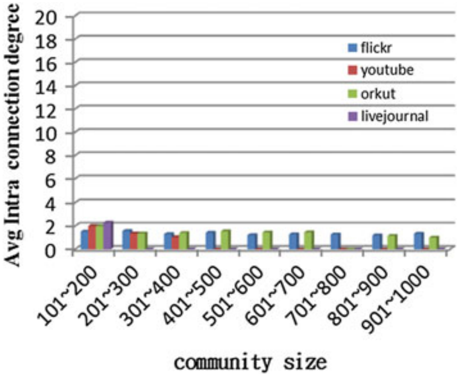


Fig. 2 The average intra connection degree of large size communities



which users can upload, view and share videos. Most of the content on YouTube has been uploaded by individuals, although media corporations including CBS, and other organizations offer some of their material via the site. In the platform, unregistered users can watch videos, while registered users can upload an unlimited number of videos.

The first experiment shows the relationship between the number of community members and the intra connection degree of the community. The results for small size and large size communities are shown in Figs. 1 and 2, respectively.

As shown in the results, the average intra connection degree of the communities whose sizes are within 2–10 is much larger than that of other group of communities. It indicates that the relationship between the members in a small community is very tight. Moreover, when the community size exceeds 10, the intra connection degree becomes quite small, which means the communities have a loose connection structure. This is because the users join the community according to their interests, which means the members belong to the same community have the similar interests and they are not necessary to know each other.

The second experiment shows the relationship between the number of community members and the inter connection degree of the community. The results

Fig. 3 The average inter connection degree of small-size communities

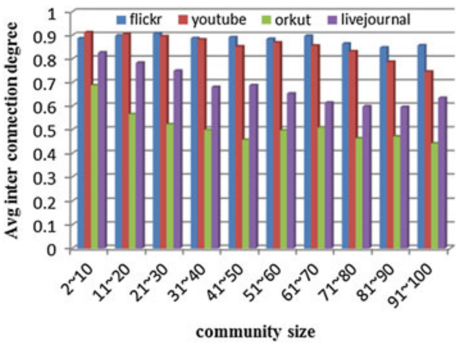
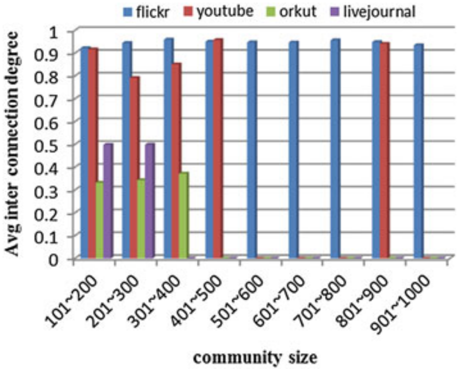


Fig. 4 The average inter connection degree of small-size communities



for small-size and large-size communities are shown in Figs. 3 and 4, respectively. As indicated in the results, the inter connection degrees of Flickr and Youtube are quite large, which means the centers of a community are also likely to be the centers of other community. The simulation result shows that the connection between real communities is high, and is a very different result comparing to the idea of cluster analysis proposed in previous works.

Conclusions

In this paper, by exploring the structures and characteristics of the social networks according to the community which the users actually joined, the characteristics of social networks are discussed. We propose several measurement methods to model the characteristics of the social network based on the community structure. To measure the inter connection degree of a community, we introduce the concept of centers and define what is a strong connection between the communities. Moreover, we experiment on four real datasets, LiveJournal, Flickr, Orkut and Youtube, to analyze the characteristics hidden behind the social networks. According to the experiment results, we find that the community has a loose connection structure when its size exceeds 10. Moreover, the inter connection between communities is high especially in Flickr and Youtube. It is a very different result by comparing to the concept of cluster analysis proposed in previous works.

Acknowledgments This work was partially supported by the National Science Council of Taiwan, under contracts NSC 101-2221-E-259-002 and NSC 101-2221-E-259-004.

References

1. Chen W, Wang Y, Yang S (2009) Efficient Influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, pp 199–208
2. Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of SIGKDD 2003, pp 3–9
3. Goyal A, Bonchi F, Lakshmanan L (2008) Discovering leaders from community actions. In: Proceedings of ACM conference on Information and knowledge management, pp 3–7
4. Biryukov M (2008) Co-author network analysis in DBLP: classifying personal names. Springer, Berlin Heidelberg, pp 403–407
5. Lancichinetti A, Fortunato S, Kertesz J (2009) Detecting the overlapping and hierarchical community structure of complex networks. Physics 11(3):4–14 (arxiv.org)
6. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proceedings of ACM KDD 2006, pp 44–53
7. Licamele L, Getoor L (2006) Social capital in friendship-event networks. In: Proceedings of the sixth IEEE international conference on data mining, ICDM 2006, pp 1–12
8. http://socialnetworks.mpi-sws.org/data-imc2007.html

Emergency Broadcast in VANET by Considering Human Satisfaction

Yu-Shou Chang, Shou-Chih Lo and Sheng-Lung Peng

Abstract The emergency broadcast is an important service in Vehicular Ad Hoc Network (VANET) for the safety of vehicle drivers. The design of a broadcast scheme in a city environment becomes challenging. In this paper, we extend our previous work of using the concept of water wave propagation by further considering human satisfaction. As long as the drivers have sufficient time to react to an emergency event, a lazy rebroadcast approach is applied to significantly reduce rebroadcast times and network traffic.

Keywords VANET · Emergency broadcast · Human satisfaction

Introduction

Vehicular Ad Hoc Network (VANET) [1] is a type of mobile wireless network which supports for multi-hop wireless communications between vehicles. The most important application of VANET is to disseminate emergency messages to drivers in case of dangerous events [2]. A warning message needs to be delivered with low delay and high reliability to those vehicles that are located within a warning area. Also, this warning message needs to be delivered to those vehicles that newly enter the warning area within a warning time. This kind of service relies on emergency broadcast.

The broadcast design in VANETs suffers from new challenges beside the well-known broadcast storm problem [3], and these challenges are the connection hole problem, the building shadow problem, and the intersection problem [4]. These problems become more serious in a city environment with many street roads.

Y.-S. Chang · S.-C. Lo (✉) · S.-L. Peng
Department of Computer Science and Information Engineering, National Dong Hwa University Hualien, Hualien 974, Taiwan, Republic of China
e-mail: sclo@mail.ndhu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_17, © Springer Science+Business Media Dordrecht 2014

The core technique of any broadcast scheme is to select next rebroadcast nodes. That is, when a node listens to a broadcast packet, this node should follow a certain criterion to verify whether to rebroadcast this packet or not. Since there might have several nodes listening to the same broadcast, a contention or selection mechanism is applied to these nodes. Based on the different philosophies of selection mechanisms, a comprehensive survey of emergency broadcast schemes was given in [5].

Some existing broadcast schemes such as WPP [6], RBM [7], and UVCAST [8] cannot fully and efficiently solve the new challenges. In this paper, we propose an emergency broadcast scheme that is suitable for city environments. This scheme is extended from our previous one called Water-Wave Broadcast (WWB) [5] by further considering human satisfaction at the warning service. Unlike other broadcast schemes that always perform rebroadcasting immediately, our proposed scheme would perform rebroadcasting lazily as long as the drivers are satisfied with the warning service.

The remainder of this paper is organized as follows. The proposed scheme is illustrated in section Emergency Broadcast, followed by the performance evaluation in section Performance Evaluation. A brief conclusion is given in section Conclusions.

Emergency Broadcast

WWB is our previously proposed scheme that follows the concept of water wave propagation. This work further improves WWB by considering human satisfaction. Basically, we divide a warning area with circle shape into three ranges (WA_1, WA_2, and WA_3). Denote $A_{x,y}$ to be the area with the distance to the center of the warning area between x and y. For example, if the radius of the warning area (R) is 200 m, we have $\mathrm{WA}_1 = A_{0,100}$, $\mathrm{WA}_2 = A_{100,150}$, and $\mathrm{WA}_3 = A_{150,200}$. The allowable waiting time to listen to the emergency warning is t_i for WA_i. In our setting, $t_1 = 1$ s, $t_2 = 7$ s, and $t_3 = 11$ s. A person (vehicle, or node) is satisfied with the warning service if one of the following three cases is true:

Case 1: A node that has already located in WA_i can receive the warning within t_i after the beginning of the emergency broadcast.

Case 2: A node that is newly entering into WA_i can receive the warning within t_i after the entering.

Case 3: A node that is newly entering into WA_i encounters the ending of the emergency broadcast (i.e., warning time is over) within t_i after the entering, no matter whether the warning is received or not.

A warning wave is simulated as an emergency event is detected by a vehicle, and the propagation medium is vehicles on the road. The head wave will spread the event to the whole warning area. Vehicles hold the head wave as they move and forward it to other encounter vehicles. The warning area keeps rippling until the end of the warning time. Any vehicle entering into a ripple area will be notified of this event through other encounter vehicles. Accordingly, we identify the following types of tasks that are performed by a vehicle (or node).

- Normal-node task: Regular jobs to each node.
- Source-node task: Dedicated jobs to a node triggering an emergency event.
- Head-node task: Dedicated jobs to a node promoting a head wave.
- Receiver-node task: Dedicated jobs to a node receiving a broadcast packet.

Normal-Node Task

Each node periodically announces its status to all its one-hop neighbors by broadcasting a hello packet. The cycle time is called a *hello interval*. Hello packets have two formats: basic and extended. These packets in both formats carry the current location of a node. An extended hello packet additionally carries an event list that summaries what emergency events have been received and are still valid for a node. This extended hello packet is broadcast only when any new neighbor is found.

Moreover, each node maintains two tables: neighbor table and event table. The neighbor table records the information of its one-hop neighbors by listening to hello packets from them. The successive location data received from a neighbor are used to estimate the motion vector of this neighboring node. The event table records valid emergency events that have been received or generated. An event list is generated by listing the identification numbers of all entries in the event table.

Source-Node Task

If an emergency event is detected by a node, this node additionally performs this type of task. An emergency broadcast packet which specifies a warning area and a warning time is generated and broadcast then. An emergency event is *valid* for a node if the warning time is not expired and this node is currently located within the warning area. Next, this node ends the source-node task and starts performing the head-node task.

Head-Node Task

The mission of this task is to carry an emergency broadcast packet and forward it by broadcasting to the other neighbors. After the broadcast, a *reliability check* is performed to confirm that any neighbors continue rebroadcasting the packet. Otherwise, this node rebroadcasts the packet once again. Then, this node decides whether to give up this head-node task or not according to a *head-node-selection* criterion.

A node S satisfies the head-node-selection criterion, if there is no neighbor of S that is ahead of or behind S and driving the same direction as S. This node S usually locates on the border of current event propagation and cannot temporally spread the event forward or backward. Therefore, this node S carries the emergency broadcast packet and starts rebroadcasting the packet when encountering a new neighbor.

Receiver-Node Task

If any hello or emergency broadcast packets are received, a node starts performing this type of task. The mission of this task is to parse a received packet and performs jobs accordingly. The receiver node updates its neighbor table when receiving a hello packet. If an extended hello packet is received, the receiver node checks whether the node sending this packet misses any emergency events by comparing its own event list with the received one. The closest node to this sending node will handle any event missing by locally rebroadcasting the missing events. By considering human satisfaction, this local rebroadcast need not be performed immediately, or in other words, need not be performed each time. This is called a *lazy rebroadcast* approach. The lazy rebroadcast is based on a *local rebroadcast probability* as computed in (1).

$$\text{Local rebroadcast probability} = \begin{cases} 1, \quad \text{if } T_R gt; \alpha \cdot T_W \\ (1 - D/R) \times 0.5 + (T_R/(\alpha \cdot T_w)) \times 0.5, \quad \text{otherwise} \end{cases} \tag{1}$$

T_R is the remaining warning time and T_W is the warning time. When T_R is less than and equal to a certain portion of T_W (controlled by parameter α, $0 \leq \alpha \leq 1$), the local rebroadcast probability becomes small, since the warning time is almost over. Moreover, we consider the distance between the node of missing an event to the center of the warning area (denoted as D). If D is large, the local rebroadcast probability becomes small too, since this warning event is not urgent to this node.

If a non-locally broadcast packet is received, a node i waits for a certain time period as computed in (2) and then decides whether to rebroadcast the packet or not based on a *global rebroadcast probability.*

$$\text{Waiting time} = WT_{\max} \times \textit{random}\,(0, NB_i)/NB_i \tag{2}$$

$WT_{\max}$ is set to be twice the average one-hop communication delay. NB_i is the number of neighbors of node i (including node i itself). During the waiting period, node i may listen to ongoing rebroadcasts of the same packet from its neighbors. After the waiting time is over, node i records these rebroadcast nodes including the original broadcast node in set RS_i (Rebroadcast Set). This node i then need not rebroadcast the packet if all its neighbors are under the communication ranges of (in other words, are covered by) these nodes in RS_i. Here, we make an assumption that a node can receive any packets with high probability from another node within the communication range. This coverage situation is used to compute a global rebroadcast probability in (3). Denote $|RS_i|$ to be the number of nodes in the set. CV_i is the number of neighbors of node i that are also covered by nodes in RS_i.

$$\text{Global rebroadcast probability} = \begin{cases} 1, & \text{if } |RS_i| = 1 \\ 0, & \text{if } NB_i = 1 \\ (1 - CV_i/(NB_i - 1))^{|RS_i|}, & \text{otherwise} \end{cases} \quad (3)$$

For every packet rebroadcast, a reliability check is performed always to increase delivery reliability. Finally, we check whether the receiver node is suitable to be a head node by checking the head-node-selection criterion.

Performance Evaluation

To evaluate the performance, we carried out simulations using NS-2. We consider a real street environment which is imported from the TIGER [9]. A city street map of size 2,000 × 2,000 m is used as in the paper [5]. Under the street model, vehicles are generated and their moving patterns are controlled by the tool VanetMobiSim [10]. For each simulation run, one vehicle is randomly selected as an event source node. The warning area is a circle centered at the current location of the source node. The default parameter settings in our simulation are listed in Table 1.

We compare our proposed scheme (WWB) with WPP, RBM, and UVCAST. The cost metrics are *satisfy ratio* (percentage of the number of nodes satisfying the warning service in the warning area to the number of nodes entering into the warning area), *rebroadcast times* (number of times that an emergency packet is

Table 1 Parameter settings

Parameter	Value
Transmission radius	100 m
MAC protocol	IEEE 802.11p
Propagation model	Two-ray ground
Number of nodes	50–250
Vehicle speed	20–60 km/hr
Warning radius (R)	200 m
Warning time	180 s
Hello interval	1 s
WT_{max}	4 ms
Simulation time	240 s

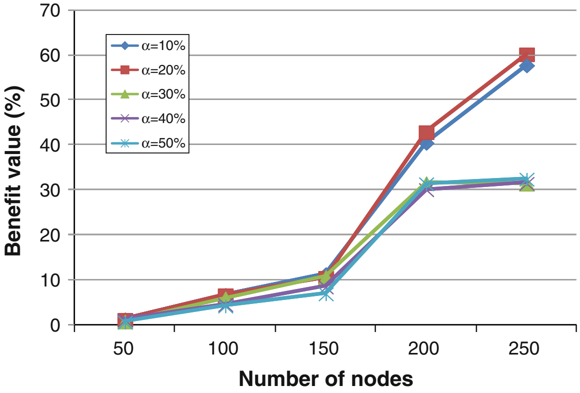


Fig. 1 The setting of parameter α

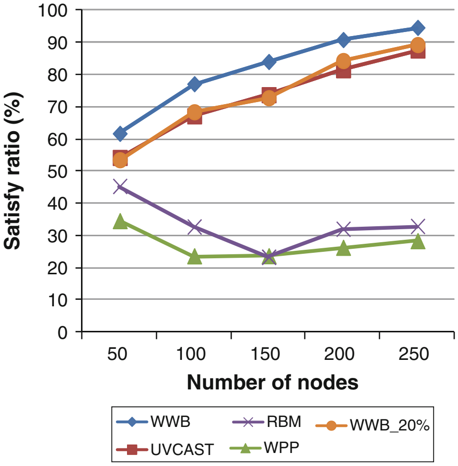


Fig. 2 Comparison of satisfy ratio

rebroadcast), and *delay time* (average elapsed time from the moment when a node enters into a warning area to the moment when the node receives the emergency packet).

At first, we evaluate the setting of α in (1) by observing the benefit value (the percentage of the amount of reduced rebroadcast times to the amount of decreased satisfy ratios). The benefit value is the best as $\alpha = 20\ \%$ (Fig. 1), which implies to use the lazy rebroadcast when the remaining warning time is less than 36 s. If the lazy rebroadcast is applied, we sacrifice a certain amount of satisfy ratios, but the proposed scheme can still compete with other schemes (Fig. 2). Also, we greatly reduce the rebroadcast times (Fig. 3), and the delay time is still acceptable (Fig. 4).

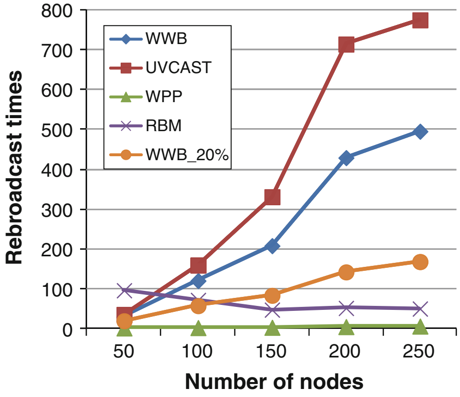


Fig. 3 Comparison of rebroadcast time

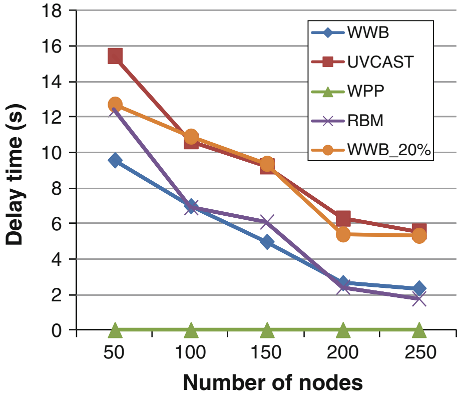


Fig. 4 Comparison of delay time

Conclusions

The dissemination of safety-related messages is an important application in a vehicular network environment. Our proposed scheme follows the water wave propagation to disseminate emergency warning messages along the street. To reduce the rebroadcast times in the whole network, we propose a lazy rebroadcast scheme by considering human satisfaction at the warning service.

References

1. Yousefi S, Mousavi MS, Fathy M (2006) Vehicular ad hoc networks (VANETs): challenges and perspectives. In: Proceedings of international conference on ITS telecommunications 2006, pp 761–766
2. Toor Y, Muhlethaler P, Laouiti A (2008) Vehicle ad hoc networks: applications and related technical issues. IEEE Commun Surv Tutorials 10:74–88

3. Ni SY, Tseng YC, Chen YS, Sheu JP (1999) The broadcast storm problem in a mobile ad hoc network. In: Proceedings of ACM MOBICOM 1999, pp 151–162
4. Yang CY, Lo SC (2010) Street broadcast with smart relay for emergency messages in VANET. In: Proceedings of IEEE AINA workshop 2010, pp 323–328
5. Lo SC, Gao JS, Tseng CC (2012) A water-wave broadcast scheme for emergency messages in VANET. Wireless Pers Commun. doi:10.1007/s11277-012-0812-2
6. Wisitpongphan N, Tonguz OK, Parikh JS, Mudalige P, Bai F, Sadekar V (2007) Broadcast storm mitigation techniques in vehicular ad hoc networks. IEEE Wirel Commun 14:84–94
7. Khakbaz S, Fathy M (2008) A reliable method for disseminating safety information in vehicular ad hoc networks considering fragmentation problem. In: Proceedings of international conference on wireless and mobile communications 2008, pp 25–30
8. Viriyasitavat W, Tonguz OK, Bai F (2011) UV-CAST: an urban vehicular broadcast protocol. IEEE Commun Mag 49:116–124
9. TIGER: topologically integrated geographic encoding and referencing http://www.census.gov/geo/www/tiger
10. VanetMobiSim http://vanet.eurecom.fr/

The Comparative Study for Cloud-Game-Based Learning from Primary and Secondary School Education Between Taiwan and America

Hsing-Wen Wang and Claudia Pong

Abstract A new way of learning has been introduced and it is here to stay: e-learning. E-learning stands for electronic learning and includes m-learning (mobile learning), u-learning (ubiquitous learning), computer-based learning, web-based learning and cloud-game-based Learning. The aim of this paper is to study cloud-game-based Learning in Taiwan and the United States of America through primary and secondary education and explore the fundamental differences between these countries. The most important ambition is to reach collaborative learning, or providing an environment where both teachers and students can debate and achieve synergy so that it is necessary to stimulate interest, improve children's performance and increase efficiency through games, problem solving and teamwork. This new learning paradigm seems to be vastly popular because it completely integrates people into learning, no matter what socioeconomic status or urban/rural area; the only two things that are significant are the willingness to learn and the teacher's effort. The paper is divided into the follows sections, including the section covers the purpose and relevance of this study supported by a review of recent e-learning and cloud-game-based Learning research; the section contains the analysis of cloud-game-based Learning in primary and secondary schools in Taiwan; the section examines cloud-game-based Learning in primary and secondary schools in the USA; the section compares cloud-game-based Learning in primary and secondary schools in Taiwan and USA. Finally, the last section covers the conclusion and suggestions.

H.-W. Wang (✉)
Department of Business Administration, Changhua University of Education, Changhua, Taiwan, Republic of China
e-mail: shinwen@cc.ncue.edu.tw

C. Pong
Department of Licenciatura en Economía, Universidad Nacional del Sur, Bahía Blanca, Argentina

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_18, © Springer Science+Business Media Dordrecht 2014

Keywords E-learning · Ubiquitous learning · Cloud-game-based learning · Comparative study · Primary and secondary schools

Introduction

A new learning paradigm has emerged, it is called e-learning and it is student centered. The main objectives are increase motivation, effectiveness and fun in learning activities. The purpose of this paper is to study Cloud-game-based Learning in Taiwan and the United States of America through primary and secondary education and explore the fundamental differences between these countries. The research methodology will be a review of a set of e-learning and Cloud-game-based Learning in the US and Taiwan. The first step is to define a game, its qualities and elements, relation to learning theory and advantages of Cloud-game-based Learning.

In the second step, it is going to be explained four games designed in Taiwan for primary and secondary schools: (1) 3D role play for learning anti-Japanese war during Qing dynasty and geography of Southern Taiwan; (2) Communicative language teaching for English learning; (3) Chinese language learning and (4) Gjun system. In third place, five games are famous in the USA for teaching primary and secondary students: (1) Making history about World War II; (2) Massive multiplayer online game (MMOG) for Mathematics, Language Arts, Science and Social Studies; (3) Dimention MTM for Mathematics; (4) Immune Attack for Science and (5) Survival Master for STEM (Science, Technology, Engineering and Mathematics).

In section Comparison Between Cloud-Game-Based Learning in Primary and Secondary Schools in Taiwan and USA, a summary is presented through a table where it is compared the target, design, hardware and software, objectives and improvements obtained by using Cloud-game-based Learning in Taiwan and USA. Finally, in section Conclusions and Suggestions, the conclusion and suggestions for Taiwan.

Literatures Reviews of Joyful Learning

In 2003, a movement was started for using video games in teaching and training. This initiative, known as serious games, has changed the way that educators viewed instruction to meet the needs of the Net generation. The perceived change in learning needs of the ‘Games Generation’ (Prensky 2001) or ‘Net Generation’ (Oblinger 2004) coupled with the ongoing growth in use and acceptability of a range of communications technology that has precipitated a growing interest in the potential of games and computer games for learning.

There are some elements that define an activity as a game: (1) Competition: the score-keeping element and/or winning conditions which motivate the players and provide an assessment of their performance. (2) Engagement: or intrinsic motivation means that once the learner starts, he or she does not want to stop before the game is over and the four sources are challenge, curiosity, control and fantasy (Beck and Wade 2004; Prensky 2006). (3) Immediate Rewards: Players receive victory, points or descriptive feedback, as soon as goals are accomplished.

Games fulfill a number of educational purposes. Some games are explicitly designed with educational purposes, while others may have incidental or secondary educational value. All types of games might be used in an educational environment. Educational games are games that are designed to teach people about certain subjects, expand concepts, reinforce development, understand an historical event or culture or assist them in learning a skill as they play (Aldrich 2004; Foreman et al. 2004; Prensky 2001; Quinn 2005).

However, a game is educational when it makes learning integral to scoring and winning. It is not enough to simply incorporate course material into a game and if it is possible to score and win without learning, students are likely to do so. There are different kinds of games [1]: (1) Video Games: These are played over the Internet, on personal computers or on specific game consoles hooked up to televisions. (2) Role-Playing Games: These are generally cooperative and highly engaging with a subtle way of handling scoring. (3) Board and Card Games: These tend to emphasize strategy elements rather than being completely random games of chance. Some of board and miniature games take hours or even days to play. (4) Sports: Students do not need to be physically fit to enjoy running around chasing things. (5) Scavenger Hunts, Raffles, etc.: When these events are organized as fundraisers for students and they tend to be quite popular with students.

Not only does the integration of learning with gaming make science more fun; it also motivates students to learn through doing, immerses them in the material so they learn more effectively and encourages them to learn from their mistakes. Games are such a great escape from the real world because bad consequences are rarely serious or lasting, they are only a game and if students lose, they can start the game over and try again.

These findings frame the three key aspects to Cloud-game-based Learning: motivation, skill development and immersive learning environments. The very nature of games provides three main factors for motivation: fantasy, challenge and curiosity (Malone 1981). Fantasy relates to the use of imagination and the child's inherent inclination towards play (Opie and Opie 1969). There is freedom to fail, experiment, fashion identities, freedom of effort and interpretation that create a learning space where new ideas and problem-solutions can emerge (Klopfer et al. 2009).

Beyond increased motivation, teachers using games in classroom have also noted improvement in several key skills areas (Joyce et al. 2009): personal skills (such as initiative, persistence, planning and data-handling), spatial and motor skills (such as coordination and speed of reflexes), social (such as teamwork, communication, negotiating skills and group decision-making) and intellectual (such as problem-solving, strategic thinking and application of numbers).

About learning environments, games allow players to enter environments that would be impossible to access in any other way; for instance, going back in history, understanding the complexity of running a major city, managing entire civilizations or nurturing families. They require engagement with complex decisions like exploring the effects of different choices and a multiplicity of variables offering ongoing and responsive feedback on choices. They also stimulate conversation and discussion; players share ideas, hints and tips in what increasingly tend to be lively and supportive learning communities (ELSPA 2006).

According to James Paul Gee (2003), digital games create ‘semiotic domains’ which are any set of practices that recruits one or more modalities (for example, oral or written language, images, equations, symbols, sounds, gestures, graphs, artifacts, etc.) to communicate distinctive types of meanings. The semiotic domain for a game is the world or culture it creates and is shared by those participating in the game together where they share knowledge, skills, experiences and resources. Active and successful participation in a semiotic domains demonstrated by ‘active learning’, where group members gain there sources and skills to solve problems within and perhaps beyond the domain as well as ‘critical learning’, which includes thinking about the game at a ‘meta’ level so that they cannot only operate within the game but within the social structure that surrounds the game as well (Williamson 2003).

However, teachers are also consistently found to be critical components ineffective Cloud-game-based Learning. Where the game is just the tool, the teacher is essential to effective implementation of the game through direction of the learning approach, discussion, debrief and support in construction of the social learning culture that surrounds the game-play. Numerous researchers have stated that learning with educational video games is not likely to be effective without additional instructional support and effective strategies for implementation (Leemkuil et al. 2003; O’Neil et al. 2005; Wolfe 1997).

Research Methodology and Issues

The paper aims to identify key issues and themes arising from the literature reviewed, the case studies produced and the consultation undertaken. The review comprises a meta-review that is a review of literature reviews, and literature has been grouped in relevant categories according to selected themes or issues. Literature was sourced from keyword searches of electronic databases, key journals in the field and a general search of the internet. Selected criteria include significant meta-reviews, relevance to Cloud-game-based Learning and empirical studies of the use of games. The criteria were used to identify relevant literature for inclusion

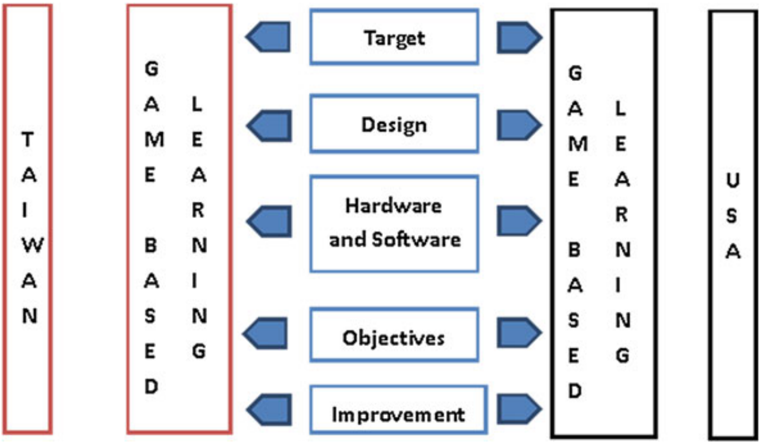


Fig. 1 The research structure of our framework

in the paper. Recommendations from experts in the field were also used to identify key articles and texts relating to examples from the practice.

In the introduction and literature review, there were a definition of game, its qualities, types and elements, relation to learning theory and advantages of Cloud-game-based Learning. In section Literatures Reviews of Joyful Learning, it is developed four games designed in Taiwan for primary and secondary schools: (1) 3D role play for learning anti-Japanese war during Qing dynasty and geography of Southern Taiwan; (2) Communicative language teaching for English learning; (3) Chinese language learning and (4) Gjun system.

In section Research Methodology and Issues, it is explained five games that are famous in the USA for teaching primary and secondary students: (1) Making history about World War II; (2) Massive multiplayer online game (MMOG) for Mathematics, Language Arts, Science and Social Studies; (3) Dimention MTM for Mathematics; (4) Immune Attack for Science and (5) Survival Master for STEM (Science, Technology, Engineering and Mathematics).

In section Comparison Between Cloud-Game-Based Learning in Primary and Secondary Schools in Taiwan and USA, a summary is presented through a table where it is compared the target, design, hardware and software, objectives and improvements obtained by using Cloud-game-based Learning in Taiwan and USA. Finally, in section Conclusions and Suggestions, the conclusion and suggestions for Taiwan. The following figure shows briefly section Comparison Between Cloud-Game-Based Learning in Primary and Secondary Schools in Taiwan and USA mentioned above: Fig. 1

Comparison Between Cloud-Game-Based Learning in Primary and Secondary Schools in Taiwan and USA

When looking for papers for the USA about Cloud-game-based Learning in primary and secondary school, most of them were orientated to college and university applications. While in Taiwan, most of the research papers are based on Cloud-game-based Learning for primary and secondary education. One reason for this is that Taiwan wants their student to get accustomed to information technology and computing devices so that it is increased the competitiveness of the students and their clerical skills. However, as part of its Connected Educator Month, the U.S. Department of Education notes that Cloud-game-based Learning is gaining considerable attention as more and more young people are learning from games outside of school, and more and more teachers are leveraging the power of games to engage students in school. Well-designed games can motivate students to actively engage in meaningful and challenging tasks, and through this process to learn content and sharpen critical-thinking and problem-solving skills.

Most of the games found for the USA were focus on one specific lesson instead of the complete subject in primary and secondary education. In Taiwan, the games that were analyzed include exercises and answers for the students as well as test and record graphics. However, these functions need more equipment and requirements i.e. while teaching only one lesson needs only one software (that can be saved with other programs), in order to teach a complete subject in primary and secondary schools, it is necessary a hardware or electronic device for the software or video game.

On one hand, Cloud-game-based Learning in the USA mainly pays attention to the achievement of higher grades in the students' subjects and the improvement of skills like problem solving, team working and strategic planning. On the other hand, in Taiwan is important to increase motivation in students and develop a deep understanding, an enjoyable experience and cultural immersion inside and outside the classrooms Table 1.

Table 1 Comparison between cloud-game-based learning in primary and secondary schools in Taiwan and the USA

	USA	Taiwan
Target	Universities	Primary and secondary schools
Design	Only one lesson	Subject and exams
Hardware and software	Software	Software and hardware
Objectives	Achievement	Motivation
Improvement	Problem solving	Deep understanding
	Team working	Enjoyable experience
	Strategic planning	Cultural immersion

Conclusions and Suggestions

It is clear from the data that Cloud-game-based Learning presents an opportunity to engage students in activities, which can enhance their learning. Like any successful pedagogy, outcomes need to be well planned and classrooms carefully organized to enable all students to engage in learning. What is notable about using games for learning is the potential they have for allowing many children to bring their existing interests, skills and knowledge into the classroom and then use games as a hook or stimulus to build the activities for learning around them. In many ways these findings reflect those of earlier media education programs, which sought to capitalize on children's own interest in television and film and build activities around them.

Although, it is good to have Cloud-game-based Learning in primary and secondary schools in Taiwan, it would be convenient to extend Cloud-game-based Learning to college and kindergarten too as well as develop games in different languages so that different countries can take advantage of them because students perceived a range of educational benefits as a result of participating in the Cloud-game-based Learning approaches, including increased collaboration, creativity and communication.

For future research, it would be valuable to investigate how to cultivate more interpersonal relationships, how to improve privacy and security in using online video games and integrate Cloud-game-based Learning with u-learning and m-learning.

Acknowledgments This work was supported in part by the National Science Council (NSC), Taiwan, ROC, under Grant NSC Taiwan Tech Trek 2012, NSC 101-2511-S-018-016.

References

1. Blunt RD (2005) Knowledge area module V: a framework for the pedagogical evaluation of video cloud-game-based learning environments, Applied Management and Decision Sciences. Walden University, Minneapolis
2. Carleton College http://serc.carleton.edu/introgeo/games/index.html
3. Groff J et al. (2010) The impact of console games in the classroom: evidence from schools in Scotland. Futurelab, Innovation in Education. www.futurelab.com.uk
4. Watson WR et al. (2010) A case study of the in-class use of a video game for teaching high school history. Elsevier, computers and education, Purdue University, Department of curriculum and instruction, United States, contents lists available at Science Direct. www.elsevier.com/locate/compedu
5. Swearingen DK (2011) Effect of digital game based learning on ninth grade students' mathematics achievement. University of Oklahoma, ProQuest LLC, Oklahoma
6. Hacker M, Kiggens J (2011) Gaming to learn: a promising approach using educational games to stimulate STEM learning. In: Barak M, Hacker M (eds) Sense publishers, fostering human development through engineering and technology education, pp 257–279

7. Leonard AA (2008) Video games in education: why they should be used and how they are being used. Theor Pract, N C State Univ 47(3):229–239
8. Whitton N (2007) Motivation and computer game based learning. Education and Social Research Institute, Manchester Metropolitan University, Singapore
9. de Freitas S (2006). Learning in immersive worlds. A review of cloud-game-based learning. JISC e-learning programme
10. Shih JL et al. (2010) Designing a role-play game for learning Taiwan history and geography. IEEE computer society, 2010 IEEE international conference on digital game and intelligent toy enhanced learning
11. Wang YH (2010) Using communicative language games in teaching and learning english in Taiwanese primary schools. J Eng Technol Educ, Kainan Univ 7(1):126–142
12. National Changhua University of Education–NCUE (2012) Chinese language e-learning site content, effectiveness and relationship management for user satisfaction
13. Hwang WY et al (2011) Effects of reviewing annotations and homework solutions on math learning achievement. Br J Educ Technol 42(6):1016–1028
14. Huang YM et al (2012) A ubiquitous English vocabulary learning system: evidence of active/passive attitudes vs. usefulness/ease-of-use. Comput Educ 58(1):273–282 Elsevier, Science Direct
15. Huang YM et al (2011) The design and implementation of a meaningful learning-based evaluation method for ubiquitous learning. Comput Educ 57(4):2291–2302 Elsevier, Science Direct
16. Huang YM et al (2011) Development of a diagnostic system using a testing-based approach for strengthening student prior knowledge. Comput Educ 57:1557–1570 Elsevier, Science Direct
17. Huang YM et al (2010) An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization. Comput Educ 55:1483–1493 Elsevier, Science Direct
18. Huang YM et al (2011) Applying adaptive swarm intelligence technology with structuration in web-based collaborative learning. Comput Educ 52:789–799 Elsevier, Science Direct
19. Huang YM et al (2009) An adaptive testing system for supporting versatile educational assessment. Comput Educ 52:53–67 Elsevier, Science Direct
20. Huang YM et al (2008) Toward interactive mobile synchronous learning environment with context-awareness service. Comput Educ 51:1205–1226 Elsevier, Science Direct
21. Huang YM et al (2010) Effectiveness of a mobile plant learning system in a science curriculum in Taiwanese elementary education. Comput Educ 54:47–58 Elsevier, Science Direct
22. Huang YM et al (2008) A blog-based dynamic learning map. Comput Educ 51:262–278 Elsevier, Science Direct

The New Imperative for Creating and Profiting from Network Technology Under Open Innovation: Evidences from Taiwan and USA

Hsing-Wen Wang, Raymond Liu, Natalie Chang and Jason Chang

Abstract In the rapidly changing Web 2.0 era innovation has become a key focus in organizations around the world. Innovation can be described as the creation of new products that take advantage of changing markets and improved technology, but innovation also means the ability to adapt to new technologies and create new networks. In this paper we focus on organizations in different fields that attempt to adapt to advances in network technology. The fields we examine are online web stores. With regards to social media, we will examine how the design and allowance of user feedback affects the success of E-stores. In addition we also compare the studies done on these fields from Taiwan and America in order to understand how regional cultural biases affect research style and direction. This paper will study the effects of being innovative in the three fields described above as well as the different views of being innovative in America and Taiwan.

Keywords Web 2.0 · Open innovation · E-stores · Network technology

H.-W. Wang (✉)
Department of Business Administration, Changhua University of Education, Changhua, Taiwan, Republic of China
e-mail: shinwen@cc.ncue.edu.tw

R. Liu
Department of Economics and International Relations, University of California Davis, Davis, CA, USA

N. Chang
Department of Economics, Simon Fraser University, Burnaby, Canada

J. Chang
Department of Business, University of Iowa, Iowa, USA

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_19, © Springer Science+Business Media Dordrecht 2014

Introduction

The amount of people using the internet to communicate, search for information, and create new web pages has grown tremendously in the past 20 years. Not only has the user base for the internet grown, its capabilities and purposes have expanded greatly as well. With such an explosive growth in users over the past twenty, no organization in the world can ignore the importance of integrating themselves to the online network. Older functions of the internet included looking up information and communication via IRC channels. Now in the Web 2.0 the internet provides cloud storage capabilities, e-commerce, and gigantic social media community sites such as Facebook. Companies must be adaptable to the growth of the internet user base and to the creation of new platforms of internet usage. Organizations achieve this by being innovative and taking the initiative to understand each new internet application before it becomes widespread.

The rapid improvement in network technologies has made it possible for over two billion users to access the internet. In addition the internet now processes over 21 exabytes of information per month. Organizations are eager to integrate themselves with the digital world but there is still a lack of research regarding the most efficient ways to use web 2.0 network technologies. Companies are looking for the best ways to design their online store websites; they want to understand which features promote user connectivity and make communication easier. Successful E-stores need to quick and simple, as well as having the platforms necessary to form user communities that will help promote brand products (Fig. 1).

Organizations have also begun focusing their attention on popular social media sites such as Facebook to serve as a new advertising and marketing platform. With over 873 million users (http://www.checkfacebook.com/), Facebook is by far the largest social media site. Companies and other independent organizations see great potential in the Facebook fan page application, a page which allows Facebook users to show they "like" a brand name or cause. Companies can upload pictures and videos, provide updates, and create special Facebook apps in order to build rapport with customers and improve their brand image. Fan pages are free and represent an efficient minimal cost advertising technique. Fan pages save companies money and raises brand name visibility, increasing profits.

The ability to store data online via cloud storage is a boon to institutions such as hospitals. With thousands of patient files, hospitals can benefit tremendously from online data storage, as long as they are able to keep that information secure. One such application by hospitals is the use of balance score cards as a feedback tool to inspect the efficiency level of internal operation. These score cards are costly and time consuming evaluations that could become more efficient if the process of collecting and sorting data became online-based. If these score cards could be shared online, low scoring hospitals could study higher rated hospitals and adapt their management techniques and fund allocations.

Adverse drug event (ADE) is an injury resulting from the use of a drug. Hospitals need to be careful of the drug being use in the institution. Although the medications

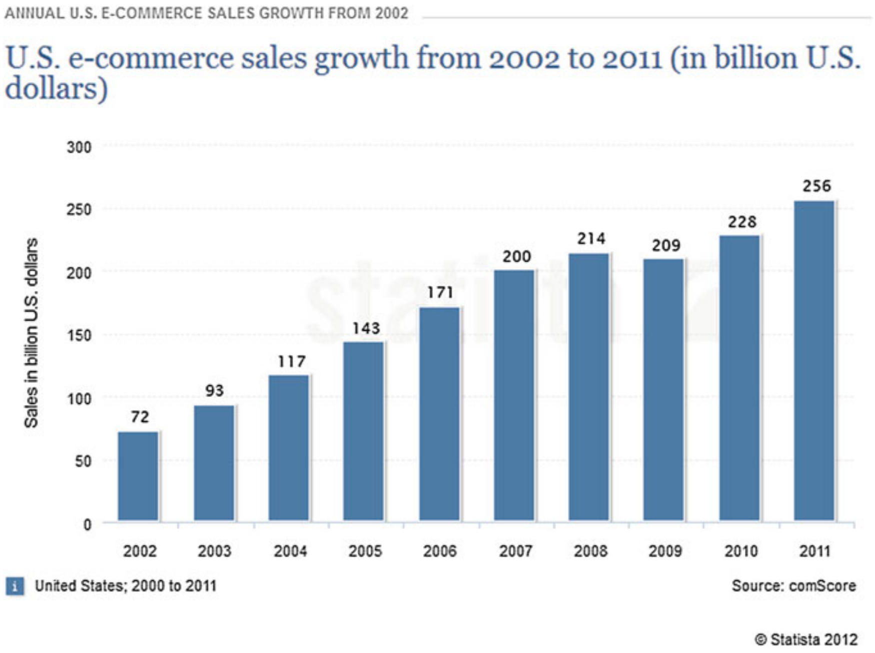


Fig. 1 The trend toward e-commerce in America

being used are tested, constant monitoring to insure safety is still required. Hence, hospitals are also dependent on online database collect and monitor patients with drug usage. Prior to 2007, adverse drug event is under recognized and underreported within the U.S. Department of Veterans Affair. Two methods for extracting and collating ADEs into national databases where developed in years 2007. They are the Adverse Reaction Tracking package and Veteran Affairs Adverse Drug Event Reporting System. It is necessary to conduct multiple studies on how to achieve the most efficient usage of all this Web 2.0 technology. We intend to examine research done in Taiwan and the United States in order to gain two different perspectives on how innovation can be best achieved. Researchers from each country will have their own cultural backgrounds that affect how their study will be designed and what conclusions they will draw from their datasets. By examining the work of two different countries, we hope that our conclusions will show how to profit from the global future of network technologies.

Literatures Reviews

E-commerce refers to a broad range of online business activities for products and services. Using private network, Electronic Data Interchange (EDI), to transact business between companies was the early form of e-commerce. As defined by

Kestenbaum and Straight [1], e-mail, electronic fund transfer, electronic data interchange, and any related technological integration could be identified as elements of a business system. However, different definitions to technological network has brought up by Kalakota and Whinston [2]. Consolidating the principles of e-commerce can be described as four types: business to business, business to consumer, consumer to consumer, and consumer to business, which are all usually associated with conducting any commercial transaction through the Internet.

With a rapid change in network technology, firms are vigorously seeking innovative strategies to advance performance and it has aroused numerous researchers' interests to study this phenomenon. Many researchers focus their studying on how corporations' existing business model influences by innovation and how they adapt these changes and integrate with old business model to form new marketing strategies. According to lots of research papers, utilizing Internet technology to expand business has more pros than cons. And also most of reports have pointed out the common reasons that corporations have chosen to launch online service are based on the factors of transaction costs, consumer behaviors, and services improvement.

The marketing strategies that the majority of companies utilize for e-business have listed out in the studies as well; such strategies are as auction, point collection, and lottery. Auction is the process of sale in which goods are sold to the highest bidder. Reward point is a program for customers to redeem points they accumulated from purchasing. And lottery involves the drawing of lots for a prize. The reward programs are powerful mechanisms for raising sales or brand loyalty [3] and the two mechanisms that make positive sales impact are "points pressure" and "rewarded behavior" [4]. Effort to earn a reward by increasing purchase is a short term impact, described as points pressure, and the long term impact whereby customers rise their purchase rate after obtaining the reward, can be defined as rewarded behavior.

Opening an e-business market can acquire profitability by offering businesses the opportunity to reduce their costs dramatically. Companies are realized that their online systems are more valuable than the companies themselves. And this could be happened is because of the innovative technology assisting corporations to maximize the benefit of transactional cost, which it always refers to the six transactional sigma identified by Downes and Mui [5]. The six types of transaction costs are search costs, information costs, bargaining costs, decision costs, policing costs, enforcement costs and IT costs. Firms can less depend on broker dealer to sell products and have fully control on the quality of products and services. Additionally, it trims down the cost on time and money to find responsible suppliers and developing new customers. Via the Web, companies can directly interact with their customers to strengthen consumer relationship and also attract more new customers who are relying on online shopping by sharing digital content. Virtual stores can provide fastest and recent updates of new products and user experiences than physical stores as well. Widespread Internet marketing has certainly opened many various unique possibilities for firms and helps to expand the corporate image. Innovative technologies can improve different aspects for a corporation and has presented an optimistic outlook of the e-business market.

The Research Framework

The topics of innovative web design, targeted Facebook fan page marketing and maximizing hospital efficiency are all centered about the usage of network technology and the importance of organizations being innovative. However, each topic deigns its own research methodology as a study designed to research hospital efficiency lacks the tools necessary to understand why consumers prefer Starbuck's fan page over Subway's fan page. In this section we examine our three different research methodologies. Hypothesis: H1: User interaction affects purchase intention; H2: Brand image affects purchase intention; H3: Event creation affects user participation; H4: Consumer attitude affects purchase intention; H5: Heuristic Information Processing is a driving force behind purchase intention; H6: Network technology usage increases profits.

Comparative methodology will be using in this research to compare and analyze the companies in manufacturing sector, 3 M and Procter and Gamble (P&G), between US and Taiwan. Both of 3 M and Procter and Gamble are multinational manufacturers and sell from home and leisure products to health care products. As attributions of markets in diverse regions, these two companies in US and Taiwan would operate on different tactics of online marketing to compete domestically. Since the online order center of 3 M in US is only open for the channel partners and other business customers, the virtual store for daily consumers have no longer existed, Procter and Gamble as a supportive sample in the comparison.

3 M in Taiwan has opened an online shopping site to serve its customers directly and conveniently which customers can purchase various 3 M products from displays and graphics, health care, to home and leisure. Since virtual store has helped 3 M sell products without broker, 3 M have fully control to promote its products in low prices irregularly to raise sells. There are different tactics can be seen on online web site of 3 M Taiwan, (1) Auction: 3 M sets a $1 in New Taiwan dollar reserve price by offering consumers up for bid. (2) Reward point: Customers can obtain points when they purchase products from 3 M online and which customers can use these collected points to exchange 3 M products; (3) Lottery: Once customers have become members online, they will receive a lottery number while logging into the Web. Customers will win a prize when his/her numbers are drawn (Fig. 2).

The listed marketing strategies are as allures to attract more shoppers and also can help 3 M to understand consumer behavior. From studying consumer behavior, 3 M can implement dynamic product adjustment to correspond with market trend and can acquire immediate responses from customers when new products are released.

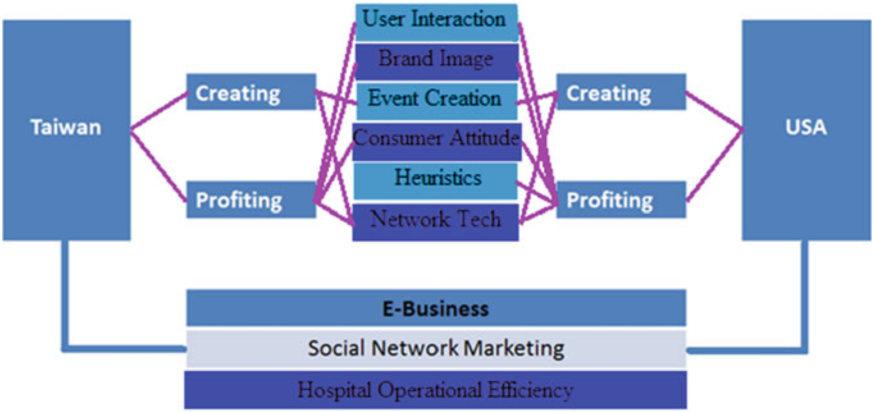


Fig. 2 The framework of this research

Additionally, 3 M has created a forum and fan page on Facebook to have live interaction with consumers. This has given a platform for consumers to share their experiences using 3 M products and for 3 M to share and update its recent news and new products vividly and instantly by uploading audios or pictures. Either social media or virtual store, innovations exponentially grow consumer affinity for the brand and mutually raise sales for 3 M and its retail partners.

As shown from the comparisons, 3 M and Procter and Gamble are vigorously seeking advance network technologies to strengthen their business. Since consumers move based on the development of the market and their shopping habits, e-business continues to grow faster than traditional retail. Percentage of e-business revenue is expected to more than double in next few years even now e-business total revenue is only 6–12 % of a traditional retailer's total revenue. Figure 1 indicates the future outlook of online shopping market is optimistic, therefore; 3 M and Procter and Gamble should gradually shift their concentration from traditional retail to e-business. 3 M and Procter and Gamble must adjust their marketing strategies as attributions in diverse regions and operate on different ideals of network technology to compete domestically.

Results and Discussions

Posts on 3 M Taiwan Facebook and individual brands of Procter and Gamble US show different atmospheres to online users. The tactic that Procter and Gamble used in social media is to partner with celebrities and it would lead to a creation of commercial polished environment which shopping decisions of some customers will influence by the celebrities. 23 % of the respondents strongly agree celebrity endorsement is a method of persuasion. Sales and brand visibility will tend to react with media exposures and reputation of celebrities positively because a high-status endorser can facilitate reassure the true quality of a product to consumers [6]. 50.7 % of the respondents agree that using celebrity endorsement will make the brand stand out in the clutter. Economic value of celebrity endorsement is profitable. As the statement from P&G India, the effect of celebrity endorsement helped one of the major brands of P&G to rank as the market leader with over 45 % market share. Different than what Procter and Gamble US approach, 3 M Taiwan provides a simple platform for consumers to share experiences and reviews. Certainly, the simplicity builds an intimate friendship between 3 M Taiwan and consumers.

Conversely, 3 M Taiwan shows creativity on its online shopping store than Procter and Gamble US. 3 M Taiwan offers a loyalty program including reward point, lottery and auction, to optimize its sales and also establish an unique way to interact with shoppers. Reward program is a promotional tool to incentivize consumers on basis of cumulative purchases from a firm [7]. From a case study, a consumer will spend between 20 and 25 % more per visit in a well-constructed loyalty program. As Lottery, shoppers require to log into 3 M virtual store in order to receive a daily lottery number. Lottery acts as a temptation to consumers for visiting 3 M virtual store everyday and it successfully increases chances of customers to purchase products even customers are not planning to. E-Store, name of Procter and Gamble's virtual store which is partnership with PFS Web, does not offer any reward programs and it makes E-Store less competitive to the market (Fig. 3).

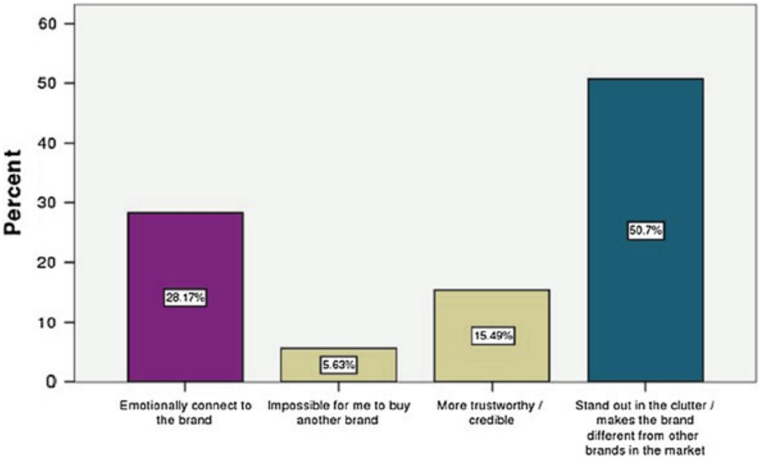


Fig. 3 The result of this research

Conclusions and Suggestions

More similar characteristics of companies will start an online shopping business to emulate and for advancing their services and the existed virtual stores should be out with the old to make way for the new in order to strive for success. The suggestions will be: (1) Strengthen relationship with online consumers. Companies must divide up the market serves to prioritize the primary target markets they will focus on, which the segments will facilitate companies to profoundly comprehend market characteristics and types of consumers. From the analysis, companies can release different kinds of discounts for distinct customers and update to keep websites continuously fresh. (2) Partnership with other firms in different industries. Companies can cooperate with other types of reward programs to reinforce services. For example, if Procter and Gamble signed an agreement with Air Miles (Air Miles is a Canada based reward program offering flight mileage on a multiplicity of products and services, which was launched in the United States in 1992): (a) Members can also earn Air Mile points when they purchase products from any brands of Procter and Gamble; (b) Members can transfer their points or flight miles between both reward programs; (c) Members can redeem Air Miles points for Procter and Gamble products.

Advantages of partnership are attracting more prospective customers and customers can enjoy a variety services from two or more companies. Either 3 M or Procter and Gamble can consider forming partnership with other firms in different industries to expand their businesses. Additionally, 3 M or Procter and Gamble can complement with its partners to better allocate their business. Besides, firms can design customer satisfaction surveys and questionnaire for existing and prospective customers to obtain information about: (1) Appealing of web pages atmosphere; (2) How good selection of products was present; (3) How satisfied consumers are with their purchases via virtual store; (4) Age range; (5) What kind of goods consumers usually buy online; (6) Online purchase intention.

Surveys and questionnaire quantify firms' strengths in growth opportunities, area for improvement, and competitive threats. Furthermore, firms are able to view their performances objectively and adjust business model and target market on the right track. As attributions of markets in diverse regions, 3 M and Procter and Gamble in US and Taiwan would operate different ideals of network technology to compete domestically and internationally. If 3 M or Procter and Gamble and other loyalty programs collaborate on rewarding, it will enhance brand visibility in multiple countries and consumers will acquire diversifies services. Besides, they can mutually help one another in business. Network technology brings professional and creativity to the business market and convenience to either firms or consumers. Innovation is inevitable.

Acknowledgments This work was supported in part by the National Science Council (NSC), Taiwan, ROC, under Grant NSC Taiwan Tech Trek 2012, NSC 101-2511-S-018-016.

References

1. Kestenbaum MI, Straight RL (1996) Paperless grants via the internet. Public Adm Rev 56(1):114–120 Washington, D.C
2. Kalakota R, Whinston AB (1996) Frontiers of Electronic Commerce. Addison Wesley Longman Publishing Co., Inc., Redwood City, pp 1–3
3. Kopalle PK, Neslin SA (2003) The economic viability of frequency reward programs in a strategic competitive environment. Rev Mark Sci 1(1):1–39
4. Taylor GA, Neslin SA (2004) The current and future sales impact of a retail frequency reward program. http://dbs.ncue.edu.tw:2057/science/article/pii/S0022435905000692
5. Downes L, Mui C (2012) Unleashing the killer app: digital strategies for market dominance. Amazon.com
6. Chung KYC, Derdenger TP, Srinivasan K (2012) Economic value of celebrity endorsements. http://www.andrew.cmu.edu/user/derdenge/CelebrityEndorsements.pdf
7. Kim BD, Shi M, Srinivasan K (2001) Reward programs and tacit collusion. http://dbs.ncue.edu.tw:2105/stable/pdfplus/3181632.pdf?acceptTC=true
8. Je-How Y (2009) A study of applying internet marketing strategy to achieve clicks-and-mortar business
9. Liu SH, Huang JX (2011) FB million fans operating technique. Common Wealth Magazine p 464
10. Venkatesh V, Morris MG, Davis GB, Davis FD (2003) User acceptance of information technology: toward a unified view. MIS Q 27(3):425–478
11. VerticalNet (1999). http://www.verticalnet.com
12. Webopedia (what is database management system) (2012) http://www.webopedia.com/TERM/D/database_management_system_DBMS.html
13. Tseng WH (2010) A study of the effects of brand image and purchase intention on e-learning facebook fan page with moderator of cloud service. National Changhua University of Education, Changhua

Part III
Exploration of Scientific Evidence on Affective Learning

What is Affective Learning?

Wen-Yen Wang, Ling-Chin Ko, Yueh-Min Huang, Yao-Ren Liu and Shen-Mao Lin

Abstract Affective computing can be used to evaluate human psychological reactions after using affective equipment. And, learners produce intrinsic and extrinsic affective reactions in the learning process called as affective learning. However, some work evaluates affective learning through non-computing method as compared to computing one. Based on the survey, this work demonstrates two examples that used different methods to evaluate affective learning. In addition, the characteristics of the methods have been discussed and explored to understand affective learning further.

Keywords Affective computing · Affection

Introduction

What is affective learning? Affective learning means that learners generate intrinsic and extrinsic affective reactions in the learning process [1]. The reactions include personal emotion, feeling, fancy, attitude, and so on. Learners may also produce personal emotion performance and extrinsic reactions facing specific courses, teaching materials and subjects in the learning. These affective reactions

W.-Y. Wang (✉) · Y.-R. Liu · S.-M. Lin
Department of Information Engineering, Kun Shan University, No. 949, Da-Wan Road,
Yung-Kang, Tainan city 71003, Taiwan, Republic of China
e-mail: wwang@mail.ksu.edu.tw

L.-C. Ko
Information and Communication, Kun Shan University, No. 949, Da-Wan Road,
Yung-Kang, Tainan city 71003, Taiwan, Republic of China

Y.-M. Huang
Department of Engineering Science, National Cheng Kung University, Taiwan, No. 1,
University Road, Tainan city 701, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_20, © Springer Science+Business Media Dordrecht 2014

have cognitive changes with personal favor degree, referring to the learners' cognitive thoughts and behavioral performance [1]. The changes in the emotion performance can be regarded as a class of emotion. The extrinsic affective reactions of the learners in studying various things can generate different emotion combinations, which are regarded as a learning reaction of affective learning.

As the equipment for affective computing is complicated and expensive, instead of using affective computing for discussion, some studies observed the learners' intrinsic and extrinsic emotion in learning, and used adaptive strategy for discussion, so as to evaluate the learner's affective reactions.

Related Studies

Christian (2010) indicated the affective computing technology could be used to simulate the human cognitive inference capability, and implemented the test environment for simulation [2]. It was a 3D emotional space displaying the continuous development combination of human inner thoughts and somatic reactions. The facial expression of emotional change could be calculated by affective computing, so as to know how the affective reactions generated continuous combination process of a person's emotional response to body change through the effect of cognition. Kiavash et al. (2012) proposed an improved learning framework, using web camera and microphone for learning [3]. The camera and microphone were used to collect the learners' facial expressions, operating conditions and learning reactions. The learners' learning data were analyzed by affective computing, so as to help the learners observing their learning behaviors, and improve their learning effectiveness, flexibility and expandability. Nik and Tanya (2012) indicated that the facial expression of learners could be observed by computational analysis of affective learning and computer arithmetic, so as to identify the present emotion type of the learners [4]. The results could be helpful in evaluating the different emotional reactions in various environments. The learning changes and states of the learners could be analyzed effectively. Guhe et al. [5] designed an emotion mouse. When the users used the mouse, the mouse could sense the users' physiological data, including heart rate, hand temperature, conductivity of skin, and so on. The present physiological state of the users was recorded to analyze the learners' physiological state in various learning environments. Arindam and Amlan (2012) found that the affective computing technology could improve learning [6]. The learners' facial expressions were analyzed and classified by collecting the biological signal of emotion detection. The learners' response was known from the classified facial expression, and the suitable learning style was found as detailed classification for adjusting learning.

The affective learning state can calculate and measure the affective state of the learner in learning. Some scholars have observed the intrinsic and extrinsic emotion of the learner in learning, and used adaptive strategy for discussion for evaluating the learners' affective reactions. This is called affective learning, but is seldom discussed. Anderson and Krathwohl [7] proposed the theoretical five-hierarchy architecture of affective learning, including receiving, responding, valuing, organization and internalization from bottom to top. At the low hierarchy of the primary structure of emotion theory, the behavior of emotion hierarchy is more specific and apparent; at the high hierarchy, the hierarchy of emotion is more abstract and complex.

Although affective learning implements different strategies by using or not using affective computing, they aim to discuss the emotions of learners in learning, and to use these emotions to improve the learning state of learners.

Example of Affective Computing Strategy

Kwok et al. [8] suggested helping students to understand and utilize Six Thinking Hats in SAMAL to generate creative solution, instead of asking them to wear six colors of hats, and using the six interactions and emotions to evoke their mental states for problem solving. The SAMAL (emotional atmosphere learning) provides an unique integrated environment, using cognitive and environmental emotions to improve learning. Bono's Six Thinking Hats approach in learning process proposed a research model to check the learners' affective experience, learning participation and creativity in positive SAMAL environment. In the research model, compared with physical setting, SAMAL setting uses learners' vision, hearing, sense of smell and interactive feeling to evoke appropriate affective and psychological conditions, thus stimulating learners to participate in the process of Six Thinking Hats, and generating creative solutions in SAMAL. Therefore, the ambient stimulation or message is delivered to learner. The learner produces the personal perception after receiving the delivery, then express emotion that the perception affects his or her thinking. The entire procedure is named as ambient affective computing as shown in Fig. 1.

Example of Strategy not Using Affective Computing

Besides measuring biological features of learners, some scholars have discussed another aspect of affective learning, which is the basic concept of emotion. First, the learner receives learning triggering action, so that the internal learner responds and evaluates the value and experience obtained from learning. The value and experience of learning are reorganized, and the combined value and personal value

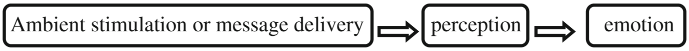


Fig. 1 Ambient affective leaning

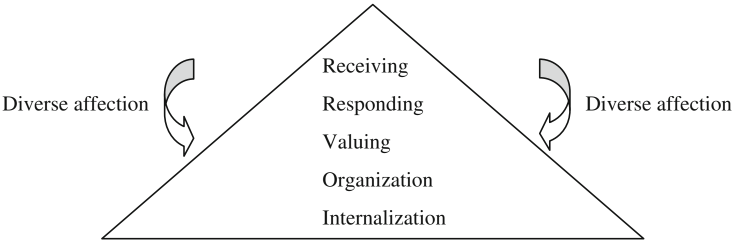


Fig. 2 Affective learning with diverse affection

are combined. The architecture of affective learning theory is formed systematically to attain the deep level objective of educational psychology aspect. Krathwohl et al. [9] proposed the research direction of using five learning phases to observe the learners' emotional change and external response in learning. The extrinsic learning of learners is converted into personal psychological features, including intrinsic interest, attitude and value. Finally, American educationalists Anderson and Krathwohl (2001) proposed the overall theory of affective learning [7]. The theory of affective learning, from bottom to top, is divided into five hierarchies, including receiving, responding, valuing, organization, and internalization as Fig. 2. At the low hierarchy of the primary structure of the emotion theory, the behavior of emotion hierarchy is more specific and apparent; at the high hierarchy, the hierarchy of emotion is more abstract and complex. The architecture of affective learning is formed systematically to attain the deep level objective of educational psychology aspect. This is another part discussed in this study. The five hierarchies are introduced as follows. Receiving means the learner is willing or active to receive or participate in learning activities with some kind of stimulation. Responding indicates the learner is willing to participate in learning activities, and participates in learning as interested in learning. Thereafter, the affection is passed to valuing. On this stage, in terms of the learner's cognition or impression of persons and objects in an environment, the objects are judged and measured by personal inner assessment standard. It can be regarded as personal intrinsic value measurement criteria. Then, the organization stage expresses the learner conceptualizes the learned or referenced values, and then classifies and integrates them into a new value by systematization. Finally, the learner characterizes various exotic values, integrated with personal value by personal judgment into personal shared value belief system as the basis of behaving and dealing with matters in the future. This system can integrate belief, concept and value structures into consistent intrinsic system.

Conclusion and Future Research

Using auxiliary system to measure the features of learners in learning can reach very high accuracy, however, such measurement and computing are very costly and complex, and cannot be used extensively. Thus, the affective learning computing has not yet been extensively used in learning so far [1].

In terms of whether affective learning has used affective computing, there has not yet been questionnaire for affective learning in the affective assessment of affective learning in studies. Future studies can design a questionnaire for affective learning factors.

References

1. Picard R, Papert S, Bender W, Blumberg B, Breazeal C, Cavallo D et al (2004) Affective learning: a manifesto. BT Technol J 22(4):253–269
2. Becker-Asano C, Wachsmuth I (2010) Affective computing with primary and secondary emotions in a virtual human. Auton Agent Multi-Agent Syst 20(1):32–49
3. Bahreini K, Nadolski R, Westera W (2012) FILTWAM-a framework for online affective computing in serious games. Procedia Comput Sci 15:45–52
4. Thompson N, McGill TJ (2012) Affective tutoring systems: enhancing e-learning with the emotional awareness of a human tutor. Int J Inf Commun Technol Educ 8(4):75–89
5. Guhe M, Gray WD, Schoelles MJ, Liao W, Zhu Z, Ji Q (2005) Non-intrusive measurement of workload in real-time. In: Proceedings of the human factors and ergonomics society annual meeting, 2005. SAGE Publications, pp 1157–1161
6. MacLean EL, Matthews LJ, Hare BA, Nunn CL, Anderson RC, Aureli F et al (2012) How does cognition evolve? phylogenetic comparative psychology. Anim Cogn 15(2):223–238
7. Anderson LW, Krathwohl DR, Airiasian W, Cruikshank K, Mayer R, Pintrich P (2001) A taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational outcomes, Complete edition. Longman, New York
8. Kwok R, Cheng SH, Ho-Shing Ip H, Kong J (2011) Design of affectively evocative smart ambient media for learning. Comput Educ 56(1):101–111
9. Krathwohl DR, Bloom BS, Masia BB (1964) II: handbook II: affective domain. David McKay, New York

The Influences of Emotional Reactions on Learning Gains During a Computerized Self-Assessment Test

Yueh-Min Huang, Chin-Fei Huang, Ming-Chi Liu and Chang-Tzuoh Wu

Abstract This study aims to examine learning gains and emotional reactions by receiving applause during computerized self-assessment testing for elementary school students. The participants were asked to solve mathematics problems in a computer-assisted self-assessment system with or without pre-recorded applause as emotional feedback while EEG measurements were taken. The results of this study provide support for the belief that it is useful to improve students' learning achievement by using emotional reactions such as applause during computer-assisted self-assessment testing, especially for male students. It is suggested that teachers may create such a positive emotional self-assessment learning environment as encourage students to learn by themselves more efficiently.

Keywords Applause · EEG · Emotion · Computer-assisted self-assessment

Introduction

Assessment is one of the useful instruments in the education domain since it can be frequently carried out by teachers to evaluate students' learning gains. However, from the students' perspective, assessment normally turns into an invisible source of stress and anxiety if they cannot achieve the expected grades [1]. A self-assessment test system is typically considered as an effective instructional strategy for training students to evaluate their own learning progress and helping them

Y.-M. Huang (✉) · M.-C. Liu
Department of Engineering Science, National Cheng Kung University,
Tainan City, Taiwan, Republic of China
e-mail: chiaju1105@gmail.com

C.-F. Huang · C.-T. Wu
Graduate Institute of Science Education, National Kaohsiung Normal University,
Kaohsiung, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_21, © Springer Science+Business Media Dordrecht 2014

prepare to face anxiety and other emotional states during tests [2]. Self-assessment provides a practice chance for students to rehearse the course content and discover unfamiliar content, so they will be prepared for in-class tests [3]. Incidentally, an increase in students' test achievement and a decrease in their anxiety will occur through the training of self-assessment (Snooks 2004).

Although self-assessment is beneficial for students to prepare for tests and to maintain learning motivation, paper-based assessment always creates the stressful experience of a typical exam [4]. The evidence from past studies indicates that computer-based assessments (CBA) could be more user friendly for students [2]. Besides, the advantages of CBA include security, reducing the time and cost of assessment, recording students' test results automatically, and providing instant feedback [5]. To sum up, it is both important and suitable for learners to integrate self-assessment and computer-based assessment [2], (Nicol and Macfarlane-Dick 2006). In this study, we therefore combined self-assessment and computer-based assessment into a computerized self-assessment test.

Affect is the basis of human experience. There are two types of affect, positive and negative [6], where positive affect (e.g. acceptance, joy, confidence, etc.) has a positive impact on learning, memory, and thinking. On the contrary, negative affect (e.g. anxiety, anger, fear, sadness, etc.) has a negative impact on motivation, and leads to inattention. Moridis and Economides [2] indicated that negative affect impedes learning. More negative affect is caused when the learner replies to a question with the wrong answer. Therefore, affect has a tremendous impact on learners' learning. However, in paper-based assessment, it is difficult to understand learners' affect and to give them immediate feedback [7].

Issues of immediate affective feedback for computer-assisted self-assessment have been increasingly discussed in recent studies [2]. These studies all suggest that immediate affective feedback for computer-assisted self-assessment could help to promote students' self-confidence and performance. In this study, we adopted neuroscience technology to explain the reasons for the gender differences and why the rewarding feelings influence students' test performance. The results of this study could provide psychological evidence to interpret the underlying reasons for the findings and could provide useful suggestions for the design of further systems.

Methodolgy

This study adopted the Hot Potatoes System to design the computer-self-assessment tests. The experimental task and the controlled task designs are shown in Fig. 1. In the first step, the single choice question type was chosen to compose the tests for both the experimental and controlled tasks. Second, fifteen single choice type mathematics questions were individually imported into the computer-assisted test for both the experimental and controlled tasks. In other words, there are 30 single choice test questions in this study allotted to the two parallel tasks. Third, it was designed that the participants in the experimental task would receive applause when they got the right answers, whereas the participants in the controlled task would not. Fourth, all of the participants needed to complete the computer-assisted self-assessment test. The correlation of these two tests for the two tasks is 0.96. This high correlation indicates that the two tests are highly consistent with each other.

The details of the third step are that a pre-recorded sound of applause was imported into the Hot Potatoes System for the experimental task. If the students get the correct answer, the applause sound is played, accompanied by the word "correct" shown on the computer screen. If the students get the wrong answer, no sound is played and the word "wrong" is shown on the screen. In the controlled task, there is no applause played if the answer is correct; however, the word "correct" is shown on the screen. If the students get the wrong answer, the word "wrong" is shown. The computer-assisted self-assessment tests for the experimental and controlled tasks were exported to create the achievement tests for this study.

This study was conducted at an urban university in Taiwan. A total of 30 students (n = 30, 15 males, 15 females; mean age ± S.D. = 19.2 ± 2.0 years) participated in this study, divided into two groups, one male group and one female group. All participants were asked to complete both the experimental and controlled task tests using neuroscience technology throughout the whole process to collect the psychological data. This study conformed to The Code of Ethics of the World Medical Association (Declaration of Helsinki) and was approved by the ethics committee of National Kaohsiung Normal University.

In this study, students' anxiety levels were measured using the State-Trait Anxiety Inventory (STAI) [8] which was the same instrument used in the research of [2]. The self-report inventory included 20 items to assess the participants' state of anxiety. Responses to the items ranged from 1 to 4, as follows: (1) not at all; (2) somewhat; (3) moderately so, and (4) very much, according to the students' feelings. Scores range from a minimum of 20 (highest anxiety) to a maximum of 80 (lowest anxiety). The validity of the contents of the Chinese version was assessed by three professional psychologists, and the Cronbach's alpha was 0.91 for the state subscale.

Two mathematics tests on the addition of two-digit numbers were administered in this study, consisting of 15 single choice type questions with a perfect score of 30. The correlation of these two tests is 0.96, and the Cronbach's alpha values were 0.84 and 0.87 for the two tests respectively.

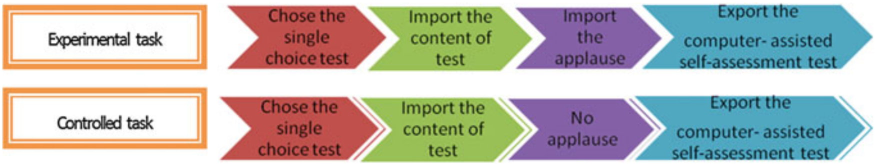

Experimental task
Chose the single choice test
Import the content of test
Import the applause
Export the computer- assisted self-assessment test
Controlled task
Chose the single choice test
Import the content of test
No applause
Export the computer- assisted self-assessment test

Fig. 1 The hot potatoes system

In this study, the EEG supplied neuroscience data indicative of the effect of gender differences on emotional reactions during a computer-assisted self-assessment test, and to the best of our knowledge, this is the first study to provide concrete evidence regarding this issue. The neuroscience technology adopted in this study is EEG (Electroencephalography) which is a procedure to measure the electrical activity of the brain through the skull and scalp [9]. When participants recognize specific emotional reflections, the corresponding electrical activities in the brain are induced [10].

All students were advised that they needed to take a computer-assisted final mathematics exam in this study. They were also told that they could participate in a computer-assisted self-assessment test twice as practice before the exam. All students in this study took the computer-assisted self-assessment test twice, the controlled test first followed by the experimental test. The two tests were separated by an interval of one week. The duration of each test was approximately 20 min.

The state of anxiety questionnaire was distributed both before and after the controlled and the experimental tests. In the process of completing the tests, all participants had to wear the electrode cap in order to collect the EEG data. At the beginning of the EEG experiment, all students were asked to sit down and relax. The EEG of their rest state was collected to be the individual brain wave baseline.

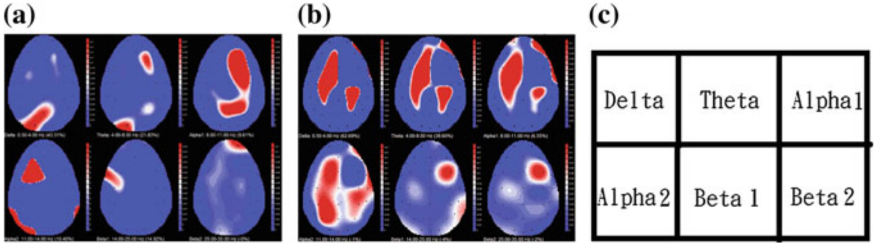
We recorded all participants' EEG signals when they were completing the experiments. Frequency analysis was performed in the delta (1–4 Hz), theta (4–7 Hz), alpha (8–12 Hz), beta (13–30 Hz) and gamma (>30 Hz) frequency bands. The evidence of brain activities from these frequencies and power values could help to verify the hypothesis that the part of the brain that generates feelings of reward is more active in males than in females during the computer-assisted self-assessment test. The extracted data were analyzed using a *t* test and ANCOVA analysis (SPSS version 17.0).

Results and Discussion

Computer-assisted self-assessment tests with and without applause were administered in this study and the statistical data from the state of anxiety questionnaire were collected before and after each test. The first to be administered was the controlled test without applause as emotional feedback. The results of this test are discussed first. The results show that the male group scored 12.4 ± 1.5 points compared with 11.3 ± 1.8 points for the female group in the controlled task mathematics test. There is no significant difference between the two groups' performance on the test. However, for the male group, the state of anxiety after the controlled test is significantly higher than before the test.

Then, t-test analysis was used to assess the differences in the scores of the state of anxiety of the male and female groups before the controlled computer-assisted self-assessment test. The results show that the females had a significantly higher state of anxiety before the test ($t = -2.05$, $p < 0.05$) than the males. Hence, analysis of covariance (ANCOVA) was used to assess the differences in the scores of the state of anxiety of the two groups after the test. The result shows that there are no significant differences (F = 2.96, $p > 0.05$, $\eta 2 = 0.099$) in the scores of state of anxiety between gender after the controlled computer-assisted self-assessment test without applause. This finding is consistent with the results of the research of [2] who suggested that the main effect of gender on state of anxiety is not significant.

Our findings by anxiety questionnaire mostly in line with those found in literatures. However, we argue that anxiety should be a continuous state rather than an outcome of a period. Apparently, the data from questionnaire in the aftermath of the test presumably regarded as the emotion state needs to be confirmed. In short, a continuous observation is a substantial step to conclude the questionnaire. For the purpose, the EEG data of the males and females who performed the experimental computer-assisted self-assessment test with applause were analyzed. The topographical map of the brain is shown in Fig. 2. For the male group, the result reveals that the power values of the alpha 1 and alpha 2 frequencies are more active on the two sides of the frontal lobe than they are for the female group. Blackhart et al. [11] mentioned that the power values of the alpha 1 and alpha 2 frequencies from the two sides of the frontal lobe are often induced by the appearance of positive emotions [11]. In other words, the higher power values of the alpha 1 and alpha 2 frequencies mean more positive emotion activities. Therefore, the findings indicate that, in the experimental computer-assisted self-assessment test with applause, the part of the brain that generates feelings of reward is more active in males than in females during computer-assisted self-assessment testing (Table 1).



Delta	Theta	Alpha1
Alpha2	Beta 1	Beta 2

Fig. 2 The hot potatoes system. **a** Females. **b** Males. **c** Frequency table

Table 1 The scores of state of anxiety before and after the computer-assisted self-assessment test for males and females not receiving applause (controlled task)

Task	Gender group	Variables	Mean ± S.D.	*t*	Cohen's *d*
Controlled task	Male group	State of anxiety (pretest)	63.87 ± 3.70	−2.41*	0.030
		State of anxiety (posttest)	61.93 ± 3.13		
	Female group	State of anxiety (pretest)	59.73 ± 6.90	−0.65(n.s.)	0.049
		State of anxiety (posttest)	60.07 ± 6.94		

Huang and Liu [12] mentioned that the frontal lobe of the human brain dominates humans' high-level thinking, mental rotation and math calculations. Ho et al. [10] also indicated that the delta frequency in the frontal lobe of the human brain is related to humans' high-level cognitive processing. For this reason, this study further analyzed the delta frequency power values in the frontal lobe (F4 electrode). The results show that both the male and female groups had higher delta frequency power values when completing the controlled computer-assisted self-assessment test without applause feedback than for the test with applause feedback (see Fig. 2).

References

1. Caraway K, Tucker CM, Reinke WM, Hall C (2003) Self-efficacy, goal orientation, and fear of failure as predictors of school engagement in high school students. Psychol Schools 40(4):417–427
2. Moridis CN, Economides AA (2012) Applause as an achievement-based reward during a computerised self-assessment test. Br J Educ Technol 43(3):489–504
3. Kulik JA, Kulik CLC, Bangert RL (1984) Effects of practice on aptitude and achievement-test scores. Am Educ Res J 21(2):435–447
4. Ricketts C, Wilks SJ (2002) Improving student performance through computer-based assessment: insights from recent research. Assess Eval High Educ 27(5):475–479
5. Terzis V, Economides AA (2011) The acceptance and use of computer based assessment. Comput Educ 56(4):1032–1044
6. Moridis CN, Economides AA (2008) Toward computer-aided affective learning systems: a literature review. J Educ Comput Res 39(4):313–337
7. Sung YT, Chang KE, Chiou SK, Hou HT (2005) The design and application of a web-based self- and peer-assessment system. Comput Educ 45(2):187–202
8. Spielberger CD (2005) State-trait anxiety inventory for adults. Mind Garden, Redwood City
9. Coles MGH, Rugg MD (1995) Event-related brain potentials: an introduction. In: Rugg MD, Coles MGH (eds) Electrophysiology of mind: event-related brain potentials and cognition. Oxford University Press, New York, pp 1–26
10. Ho MC, Chou CY, Huang CF, Lin YT, Shih CS, Han SY, Shen MH, Chen TC, Liang CL, Lu MC, Liu C-J (2012) Age-related changes of task-specific brain activity in normal aging. Neurosci Lett 507:78–83
11. Blackhart GC, Kline JP, Donohue KF, LaRowe SD, Joiner TE (2002) Affective responses to EEG preparation and their link to resting anterior EEG symmetry. Pers Individ Differ 32:167–174
12. Huang CF, Liu CJ (2012) An event-related potentials study of mental rotation in identifying chemical structural formulas. Eur J Educ Res 1(1):37–54

A Conceptual Framework for Using the Affective Computing Techniques to Evaluate the Outcome of Digital Game-Based Learning

Chih-Hung Wu, Yi-Lin Tzeng and Ray Yueh Min Huang

Abstract That's an interesting issue for how the outcome that educators use angry bird to teach projectile motion physics theorem. For verifying the possibility of playing angry bird to learn the projectile motion physics problem, this study design an experiment that include two different learning methods. One is the tradition learning method, and the other one is to learn the projectile motion using Angry Bird. When student learning, their eye movement data, brain wave and heart beat will be measured for analyzing their attention, emotion and the strategy of solving problem. After learning, they take a posttest to prove the digital game-based learning method can help student learning.

Keywords Affective computing · Eye movement · Brain wave · Heart rhythm coherence · Digital game-based learning

Introduction

The motivation of games could be combined with curricular contents into "Digital Game-Based Learning (DGBL)" [1]. Games that encompass educational objectives and subject matter are believed to hold the potential to render learning of academic subjects more learner-centered, easier, more enjoyable, and more

C.-H. Wu (✉) · Y.-L. Tzeng
National Taichung University of Education, Taichung, Taiwan, Republic of China
e-mail: chwu@ntcu.edu.tw

Y.-L. Tzeng
e-mail: bit099101@gm.ntcu.edu.tw

R. Y. M. Huang
National Cheng Kung University, Tainan, Taiwan, Republic of China
e-mail: huang@mail.ncku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_22, © Springer Science+Business Media Dordrecht 2014

interesting. Although games are believed to be motivational and educationally effective, the empirical evidence to support this assumption is still limited and contradictory [2]. The mobile game “Angry Bird” is a very famous game. Because “Angry Bird” has relations with projectile motion physics theorem. Some educators combine this game and physics course to enhance the student’s motive. But the educators use angry bird to teach that is just for fun, or really can promote student’s grade in real course. That’s an interesting issue.

For verifying the possibility of playing angry bird to learn the projectile motion physics problem, this study design an experiment that include two different learning methods. One is the tradition learning method, and the other one is to learn the projectile motion using Angry Bird. When student learning, their eye movement data, brain wave and heart beat will be measured for analyzing their attention, emotion and the strategy of solving problem. After learning, they take a posttest to prove the digital game-based learning method can help student learning.

Literature Review

Digital Game-Based Learning

Several studies found the digital game can change the players’ emotion. Ravaja et al. [3] looked at facial EMG activity and skin conductance responses as well as assessments of mood during game-play (joy, pleasant relaxation, fear, anger, and depressed feeling) in response to short duration emotional game events. They found the game events did in fact lead to emotion state [3].

Affective Computing in Learning

Since Affective Computing was proposed, there has been a burst of research that focuses on creating technologies that can monitor and appropriately respond to the affective states of the user [4]. Because this new Artificial Intelligence area, computers able to recognize human emotions in different ways. Why human emotion is an important research area? The latest scientific findings indicate that emotions play an essential role in decision-making, perception, learning and more [5].

The Physiological Input Signals this Study Selected

The physiological input signals of eye movement, EEG and ECG were selected to input our learning affective recognition system. According the past studies, several techniques need to be combined to estimate the state of attention and emotion. Eye movements provides information about location of attention and the nature, sequence and timing of cognitive operations [6]. With the emergence of Electroencephalography (EEG) technology, learner's brain characteristics could be accessed directly and the outcome may well hand-in-hand supported the conventional test recognize a learner's Learning Style [7]. And the arousal state of the brain [8], alertness, cognition, and memory [9, 10] also can be measure. Heart rate variability from ECG, has gained widespread acceptance as a sensitive indicator of mental workload [6]. And positive emotions may change the HF components of HRV [11].

Method

Research Hypotheses

To examine the relationships among different learning methods, learner attention, emotion, strategy and learning outcome, this study utilized the following two different learning methods: digital game-based learning, study the projectile motion physics problem by Angry Bird; tradition learning, study by text description, coordinates and formula. To fairly compare how digital game-based learning affect learning attention, emotions, strategy and learning outcome, the two learning methods in this study have the same learning content and learning objectives; that is, the same learning materials are presented in different methods. Figure 1 shows the relationship framework of the discussed research variables in this study (Tables 1, 2).

Research Variables

The input and output variables in this study are shown in Table 3. Learner attention is recognized by the Neurosky system, it was used to detect neuron electric triggering activity, and it has the earphone appearance. According the NeuroSky proprietary Attention and Meditation eSense algorithms, NeuroSky can report the attention score each second. The range of attention score is 1–100 (1 = very low attention level and 100 = very high attention level). Learner emotion is recognized by the emWave system, which uses human pulse physiological signals to identify Coherence score every 5 s. Coherence score have 0, 1 and 2 (0 = negative emotion, 1 = peaceful and 2 = positive emotion). Strategy includes visual attention and sequential analysis. To successfully solve this problem, participants need to distinguish the key factors. Fixation duration on options and factors provide the data to show that learners' thinking will pay much

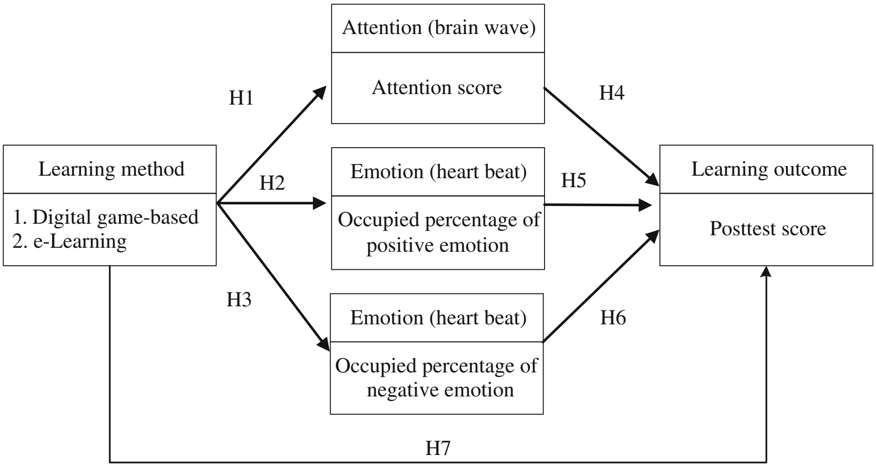


Fig. 1 Relationship framework of the research variables discussed in this study

Table 1 Multi physiological feature system review

		Physiological input signals					
Research object	Reference	Eye	EEG	ECG	Facial	Speech	SCR
Emotion recognition	[12]			×			×
Neonatal seizures	[13]		×	×			
Emotion recognition	[14]				×	×	
Emotion recognition	[6]	×		×	×		
Visual search task	[15]	×	×				
Emotion recognition	[16]		×		×		
Emotional distractors	[17]				×		
Emotion recognition	[18]					×	
Emotion recognition	[19]			×			
Brain computer interface	[20]						
Reading process	[21]	×	×				
Emotion recognition	[22]	×			×		
Learning state	[23]			×			
Driver fatigue	[24]			×			
Driver fatigue	[25]		×	×			
Emotion recognition	[8]		×				
Epilepsy state	[26]		×	×			
Learning state	*My research*	×	×	×			

According to the table, we found the physiological signals of eye movement, EEG and ECG have become the research trends. But it not exist a system combined these signals to recognize the affective of human

attention to refer which key factors. In addition, sequential analysis can effectively infer the overall behavioral path patterns during learners' thinking. We can observe the problem solving logic of learners. The assessment of learning outcome is based on pretest and posttest results.

Table 2 Hypotheses and reference in this study

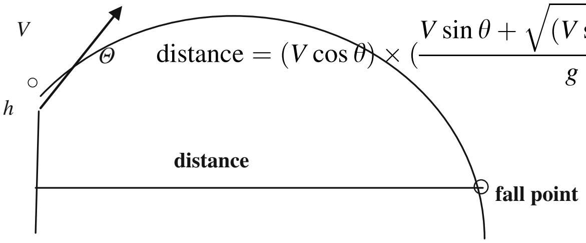
Hypotheses	Reference
H1. Learners use digital game-based learning method that have better attention score, and have significant difference	[27]
H2. Learners use digital game-based learning method that have more occupied percentage of positive emotion, and have significant difference	[3, 28]
H3. Learners use digital game-based learning method that have less occupied percentage of negative emotion, and have significant difference	[3, 28]
H4. There is a significant positive correlation between attention and learning outcome	[29, 30]
H5. There is a significant positive correlation between positive emotion and learning outcome	[23, 31]
H6. There is a significant negative correlation between negative emotion and learning outcome	[23, 31]
H7. Learners use digital game-based learning method that have better learning outcome, and have significant difference	[2]

Table 3 Input and output variables in this study

Input device	Input variables	Output
Neurosky	Attention score will be calculated each second. (The range of attention score is 1 to 100; 1 = very low attention level and 100 = very high attention level.)	Attention
emWave	Coherence score will be calculated every 5 s. (Coherence score have 0, 1 and 2; 0 = negative emotion, 1 = peaceful and 2 = positive emotion)	Emotion
Eye tracker	1. Visual attention: fixation duration on options and factors 2. Scan paths for sequential analysis	Strategy

Participants and Design

Thirty university students will participate in this study. All of them took the fundamental projectile motion course in the past. Therefore, they already possessed some prior knowledge for solving the physics problem. All participants had good visions and passed the eye-tracking calibrations. For verifying the possibility of playing angry bird to learn the projectile motion physics problem, this study refers a multiple-choice science problem solving study [32]. Four images in which four factors (velocity in slingshot, degree in slingshot and pigs from a variety of structures) were included were designed to be inspected by each participant during the problem solving task. These four factors correspond with the velocity, degree, texture and fall point in projectile formula. In projectile motion course, the projectile formula is shown in following:



$$\text{distance} = (V\cos\theta) \times \left(\frac{V\sin\theta + \sqrt{(V\sin\theta)^2 + 2gh}}{g}\right) \quad (1)$$

where V is velocity, θ is degree, h is the distance between ball and ground, distance is the length of the ball from slingshot flying to fall point.

Procedure

Participants will test individually. On arrival and after attaching physiological sensors, the participants will be asked to read an introduction. Participants will be told that four images in which four factors (velocity in slingshot, degree in slingshot, birds and pigs from a variety of structures) will be displayed for 5 m on a screen in front of them. On the top of the screen, the problem will be initiated by a statement describing, "According the first shot, please select an image inferring a best combination you will choose and justify your selection." Before the experiment begins, all of the participants pass the calibrations by eye tracker and Ware have signal input. Next, the experiment begin, the emotion and eye movement data will be recorded when participant solving the problem. In addition, participants

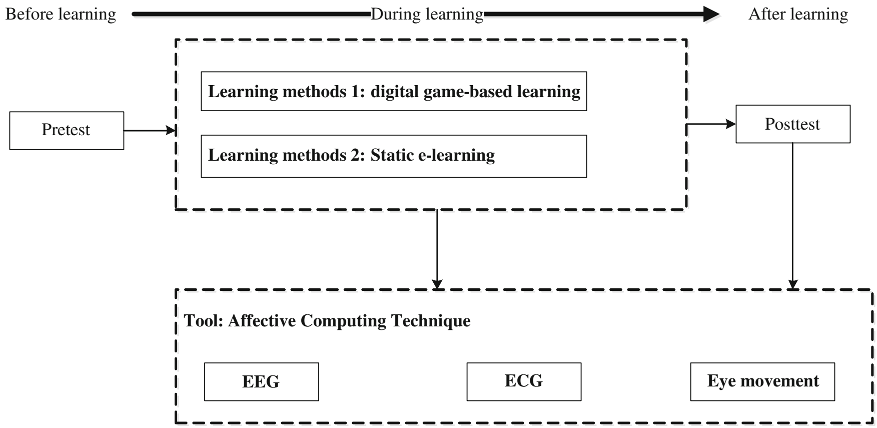

Before learning
During learning
After learning
Learning methods 1: digital game-based learning
Learning methods 2: Static e-learning
Pretest
Posttest
Tool: Affective Computing Technique
EEG
ECG
Eye movement

Fig. 2 The flow of experiment

will be asked to think aloud while solving the problem. By doing so, we can accumulate participants' justifications which are used to check against their selection. A think aloud training will be conducted before the experiment start. The entire experiment will lasting approximately 10 min. Participants' eye movement data, emotion and attention state and question responses will be recorded. The flow of experiment is shown in Fig. 2.

References

1. Prensky M (2003) Digital game-based learning. ACM Comput Entertainment 1:1–4
2. Marina P (2009) Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. Comput Educ 52:1–12
3. Ravaja N, Turpeinen M, Saari T, Puttonen S, Keltikangas-Jarvinen L (2008) The psychophysiology of James Bond: Phasic emotional responses to violent video game events. Emotion 8:114–120
4. Picard R (1997) Affective computing. MIT Press, Cambridge
5. Ben Ammar M, Neji M, Alimi AM, Gouardères G (2010) The affective tutoring system. Expert Syst Appl 37:3013–3023
6. Lin T, Imamiya A, Mao X (2008) Using multiple data sources to get closer insights into user cost and task performance. Interact Comput 20:364–374
7. Rashid NA, Taib MN, Lias S, Sulaiman N, Murat ZH, Kadir RSSA (2011) Learners' learning style classification related to IQ and stress based on EEG. Procedia Soc Behav Sci 29:1061–1070
8. Zhang Q, Lee M (2012) Emotion development system by interacting with human EEG and natural scene understanding. Cogn Syst Res 14:37–49
9. Berka C, Levendowski DJ, Cvetinovic MM, Petrovic MM, Davis G, Lumicao MN, Zivkovic VT, Popovic MV, Olmstead R (2004) Real-Time Analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. Int J Hum Comput Interact 17:151–170
10. Berka C, Levendowski DJ, Lumicao MN, Yau A, Davis G, Zivkovic VT, Olmstead RE, Tremoulet PD, Craven PL (2007) EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat Space Environ Med 78:B231–244
11. von Borell E, Langbein J, Després G, Hansen S, Leterrier C, Marchant-Forde J, Marchant-Forde R, Minero M, Mohr E, Prunier A, Valance D, Veissier I (2007) Heart rate variability as a measure of autonomic regulation of cardiac activity for assessing stress and welfare in farm animals—A review. Physiol Behav 92:293–316
12. Kim KH, Ban SW, Kim SR (2004) Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput 42:419–427
13. Greene BR, Boylan GB, Reilly RB, de Chazal P, Connolly S (2007) Combination of EEG and ECG for improved automatic neonatal seizure detection. Clin Neurophysiol 118:1348–1359
14. Ruffman T, Henry JD, Livingstone V, Phillips LH (2008) A meta-analytic review of emotion recognition and aging: implications for neuropsychological models of aging. Neurosci Biobehav Rev 32:863–881
15. Latanov AV, Konovalova NS, Yermachenko AA (2008) EEG and EYE tracking for visual search task investigation in humans. Int J Psychophysiol 69:140
16. Zhang Q, Lee M (2010) A hierarchical positive and negative emotion understanding system based on integrated analysis of visual and brain signals. Neurocomputing 73:3264–3272
17. Srinivasan N, Gupta R (2010) Emotion-attention interactions in recognition memory for distractor faces. Emotion 10:207–215

18. Yang B, Lugger M (2010) Emotion recognition from speech signals using New Harmony features. Signal Process 90:1415–1423
19. Murugappan M, Ramachandran N, Sazali Y (2010) Classification of human emotion from EEG using discrete wavelet transform. J Biomed Sci Eng 3:390–396
20. Lee EC, Woo JC, Kim JH, Whang M, Park KR (2010) A brain–computer interface method combined with eye tracking for 3D interaction. J Neurosci Methods 190:289–298
21. Dimigen O, Sommer W, Hohlfeld A, Jacobs AM, Kliegl R (2011) Coregistration of eye movements and EEG in natural reading: analyses and review. J Exp Psychol Gen 140:552–572
22. Schmid PC, Schmid Mast M, Bombari D, Mast FW, Lobmaier JS (2011) How mood states affect information processing during facial emotion recognition: an eye tracking study. Swiss J Psychol 70:223–231
23. Chen C-M, Wang H-P (2011) Using emotion recognition technology to assess the effects of different multimedia materials on learning emotion and performance. Libr Inf Sci Res 33:244–255
24. Patel M, Lal SKL, Kavanagh D, Rossiter P (2011) Applying neural network analysis on heart rate variability data to assess driver fatigue. Expert Syst Appl 38:7235–7242
25. Zhao C, Zhao M, Liu J, Zheng C (2012) Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator. Accid Anal Prev 45:83–90
26. Valderrama M, Alvarado C, Nikolopoulos S, Martinerie J, Adam C, Navarro V, Le Van Quyen M (2012) Identifying an increased risk of epileptic seizures using a multi-feature EEG–ECG classification. Biomed Signal Process Control
27. Wen-Hao H (2011) Evaluating learners' motivational and cognitive processing in an online game-based learning environment. Comput Hum Behav 27:694–704
28. Baldaro B, Tuozzi G, Codispoti M, Montebarocci O, Barbagli F, Trombini E, Rossi N (2004) Aggressive and non-violent videogames: short-term psychological and cardiovascular effects on habitual players. Stress Health 20:203–208
29. Annette MS (2004) Attention performance in young adults with learning disabilities. Learn Individ Differ 14:125–133
30. Robinson K, Winner D (1998) Rehabilitation of attentional deficits following brain injury. J Cogn Rehabil 16:8–15
31. Goleman D (1995) Emotional intelligence. Bantam Books, New York
32. Tsai M-J, Hou H-T, Lai M-L, Liu W-Y, Yang F-Y (2012) Visual attention for solving multiple-choice science problem: an eye-tracking analysis. Comput Educ 58:375–385

Adopt Technology Acceptance Model to Analyze Factors Influencing Students' Intention on Using a Disaster Prevention Education System

Yong-Ming Huang, Chien-Hung Liu, Yueh-Min Huang and Yung-Hsin Yeh

Abstract This paper explores the potential of geographic information system (GIS) in disaster prevention education. Open source GIS is applied to build a disaster prevention education system used to assist students in strengthening their knowledge of typhoon prevention and enhancing awareness of typhoon disaster. An experiment which the technology acceptance model was applied as the theoretical fundamental was designed to investigate students' intention on using the system. A total of 34 university students participated in using the proposed system. Results show that (1) perceived ease of use has a positive and significant influence on attitude toward use and perceived usefulness; (2) perceived usefulness has a positive and significant influence on attitude toward use and behavioral intentions; (3) attitude toward usage does not have a significant influence on the students' intention to use the system.

Keywords GIS · Disaster prevention education · Technology acceptance model

Y.-M. Huang
Department of Applied Informatics and Multimedia, Chia Nan University of Pharmacy and Science, Tainan, Taiwan, Republic of China
e-mail: ym.huang.tw@gmail.com

C.-H. Liu (✉)
Department of Network Multimedia Design, Hsing Kuo University of Management, Tainan, Taiwan, Republic of China
e-mail: chliu@mail.hku.edu.tw

Y.-M. Huang · Y.-H. Yeh
Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan, Republic of China
e-mail: huang@mail.ncku.edu.tw

Y.-H. Yeh
e-mail: nimo.tw@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_23, © Springer Science+Business Media Dordrecht 2014

Introduction

Disasters always caused the losses of life, property damage as well as social and economic disruption. Disaster is a serious disruption of functions of a community/society, which includes human, material, economic or environmental losses [1]. Disasters involve natural disasters, technological disasters, and man-made disasters [2]. Earthquake, flood, landslide, windstorm, drought and wildfire are a type of natural disaster [2]. Industrial and transport accident as well as bomb explosion are a type of technological disaster [2]. Man-made disaster is an event brought extensive damage and social disruption through complex technological, organizational and social process such as terrorist activity [2, 3]. Once disasters occur, it always led to huge loss no matter what caused it. Thus, it is a vital issue to develop a sound approach to mitigate the effect of disasters.

Disaster education is one of useful ways to mitigate the effects of disasters [4–7]. Early on, Vitek and Berta proposed that education is the most reliable means of gaining information about disasters and learning how to react during emergencies [7]. Later, Becker reported a technological disaster education project which was used to foster the ability of students to collaboratively deal with a chemical or nuclear disaster [4]. Ronan and Johnston examined the role of disaster education for increasing youths' resilience to disasters, and their findings revealed that disaster education was helpful to increase youths' resilience to disasters [5]. Recently, Tanaka explored whether disaster education can enhance people's readiness for disaster. His results showed people with disaster education are more prepared than people without disaster education [6]. Overall, disaster education can assist people in realizing the seriousness of disasters and promote people's capacity for handling disasters further.

Among the studies of disaster education, natural disaster education has been regarded as the most important issues in some countries such as Taiwan, because such disasters such as typhoons and torrential rains often caused huge property damage. In these countries, typhoons easily brought severe wind, floods, landslides, and debris flows from Kalmaegi (July 2007), Sinlaku (September 2007), and Jangmi (September 2008). A famous case is that Typhoon Morakot hit Taiwan on 8 August 2009 which caused the second highest damages of school facility in history [8]. Consequently, it is a vital issue to assist the people in these countries in strengthening their knowledge of disaster prevention and promoting disaster awareness for reducing the loss of life and property damage.

In this study, we used an open source geographic information system to develop a disaster prevention education system, and help students strengthen their knowledge of natural disaster prevention and promote natural disaster awareness. To explore the perspectives of students on the system, an experiment based on the technology acceptance model (TAM) was constructed [9, 10]. Specifically, we implemented the system and deployed it at a university. A questionnaire was designed to explore students' perspectives on the system. Finally, a series of analyses were conducted to examine the model and draw a conclusion about the analyses.

Research Design

Research Tool

In this work, we aimed to develop a disaster prevention education system and intended to support student engagement in a typhoon prevention education curriculum. To this end, MapGuide Open Source was used to develop the system. MapGuide Open Source is a web-based platform that enables researchers to develop and deploy web mapping applications. More importantly, it provides users with interactive system design that includes support for feature selection, property inspection, map tips, and operations. Figure 1a shows the user interface which A area shows the name of the system, B area shows the menu of system that supports students in choosing the learning topic, C area shows the description of system, and D area shows the usage of the system. Furthermore, in order to support student engagement in learning more realistically, the historical data such as photo and video is included to ensure that students can be fully immersed in the learning and achieve further meaningful learning. Figure 1b shows an example. System will play the then disaster video when students view the historical disaster event of a certain region.

Research Model and Hypotheses

TAM is regarded as one of significant roles in the successful development of e-learning system [11, 12]. It is one of famous means to evaluate users' perspective on acceptance of technology [9, 10], which was developed by Davis and his colleagues. Davis et al. proposed four main perceived constructs to develop TAM, that is, perceived ease of use (PEU), perceived usefulness (PU), attitude toward use (AT), and behavioral intentions (BI). PEU refers to a person believes that using a technology would be free of effort [9]. PU refers to a person believes that using a technology would enhance his/her job performance [9]. AT refers to a person's

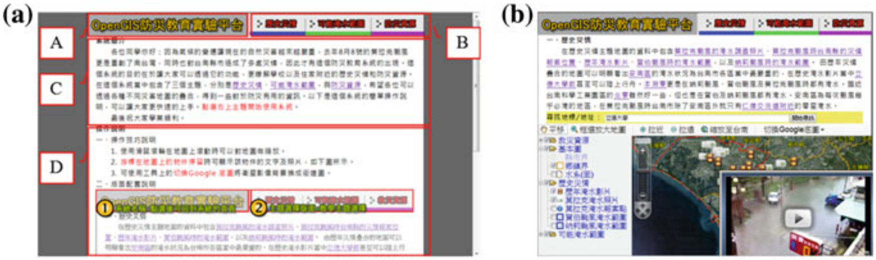


Fig. 1 The disaster prevention education system interface. **a** User interface. **b** Student views the historical disaster event of a certain region

general feeling of a favorableness or unfavorableness toward some stimulus object [13]. BI refers to a person's subjective probability that he/she will perform a specified behavior [14]. Through TAM, researchers can understand whether the system meets users' requirements and demonstrate the systems' value further. Consequently, TAM is adopted to investigate students' perspectives on the disaster prevention education system. Figure 2 shows the research model, which originates from TAM theory. The model consists of five hypotheses, which are described as follows:

From the studies of TAM [9, 10], PEU was hypothesized to influence PU and AT, and subsequently PU was hypothesized to influence AT and BI, and AT was hypothesized to influence BI. Consequently, the third to the seventh hypothesis are shown as follows:

H1. PEU is positively related to PU.
H2. PEU is positively related to AT.
H3. PU is positively related to AT.
H4. PU is positively related to BI.
H5. AT is positively related to BI.

Participants, Questionnaire, and Procedure

The participants were students from a university in Tainan City, Taiwan. A total of 34 students enrolled in the experiment. The framework for questionnaire design based on a review of prior studies [9–12, 14] as well as feedback from two experts. The questionnaire included four constructs, that is, PEU, PU, AT, and BI. At the start of the experimental procedure, all the participants executed a learning activity through the disaster prevention education system. In the activity, the participants used the system to strengthen their knowledge of typhoon prevention. When the activity was completed, the participants were asked to fill out the questionnaire that examined the proposed research model.

Results

In this study, the partial least squares (PLS) approach was used to analyze the questionnaire data, due to the small sample size. In this paper, SmartPLS 2.0 was used to assess the measurement and structural models [15]. The measurement

Fig. 2 Research model

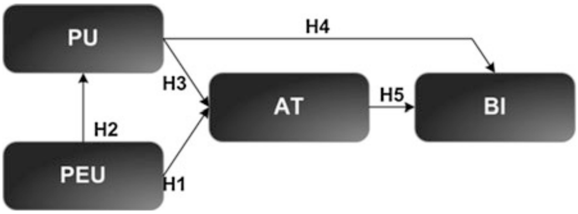
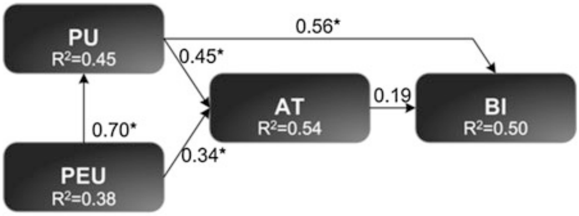


Table 1 The convergent validity, reliability of measure, discriminant validity for the measurement model

	Convergent validity	Reliability of measure		Discriminant validity			
	AVE	Composite reliability	Cronbach's alpha	Latent variable correlations			
				PEU	PU	AT	BI
PEU	0.83	0.93	0.90	0.91			
PU	0.91	0.96	0.95	0.70	0.95		
AT	0.88	0.95	0.93	0.66	0.69	0.93	
BI	0.77	0.91	0.85	0.67	0.70	0.58	0.87

Fig. 3 The results of the structural model



Note: Marked coefficients (*) are significant at $p<0.05$ ($T>1.96$).

model was assessed by convergent validity, reliability of measure, and discriminant validity. Table 1 shows the results of the measurement model is acceptable, since all the values meet the standard levels.

The structural model was used to verify the hypotheses by using path coefficients and R^2 value. The R^2 was used to assess the ability of the model to explain the variance in the dependent variables. The path coefficients were used to assess the statistical significance of the hypotheses. These results indicated that one hypothesis refuted the predictions, that is, H5; while the others confirmed the predictions. One reason for rejecting H5 is that the participants in this study did not perceive their use of technology to be mandatory, so that AT is not a significant predictor of intention to use technology. This result is consistent with previous research suggesting that AT is a significant predictor of intention to use technology mainly under mandatory conditions of technology use [16] (Fig. 3).

Conclusions

This study used an open source geographic information system to develop a disaster prevention education system to help students strengthen their knowledge of typhoon disaster prevention. To explore students' perspectives of the system, TAM was applied to build the research model, and PLS was used to assess the model. The results revealed that the system was successfully accepted by the students in the sample, but attitude toward use does not have a significant influence on the students' intention to use the system.

Limitations of this study include the type of the measurements, and the relatively small sample size. In this study, all of the measurements of this study are limited to students' self-reported perceptions. In future work, we will introduce additional measurements to explore the effects of the proposed system on disaster prevention education. Furthermore, increasing the sample size to obtain stronger evidence for the proposed system will be expected because the small sample size might limit the power of this study.

Acknowledgments The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 102-2511-S-041-001, and NSC 101-2511-S-432-001.

References

1. ISDR (2011) International strategy for disaster reduction. Retrieved 3 Apr 2011 from http://www.unisdr.org/eng/terminology/terminology-2009-eng.html
2. Mansourian A, Rajabifard A, Valadan Zoej MJ, Williamson I (2006) Using SDI and web-based system to facilitate disaster management. Comput Geosci 32(3):303–315
3. Shaluf IM, Ahmadun F, Said AM, Mustapha S, Sharif R (2002) Technological man-made disaster precondition phase model for major accidents. Disaster Prev Manage 11(5):380–388
4. Becker SM (2000) Environmental disaster education at the university level: an integrative approach. Saf Sci 35(1–3):95–104
5. Ronan KR, Johnston DM (2001) Correlates of hazard education programs for youth. Risk Anal 21(6):1055–1064
6. Tanaka K (2005) The impact of disaster education on public preparation and mitigation for earthquakes: a cross-country comparison between Fukui, Japan and the San Francisco Bay Area, California, USA. Appl Geogr 25(3):201–225
7. Vitek JD, Berta SM (1982) Improving perception of and response to natural hazards: the need for local education. J Geogr 81(6):225–228
8. Chen CY, Lee WC (2012) Damages to school infrastructure and development to disaster prevention education strategy after Typhoon Morakot in Taiwan. Disaster Prev Manage 21(5):541–555
9. Davis FD (1989) Perceived usefulness, perceived ease of use and user acceptance of information technology. MIS Quart 13(3):319–340
10. Davis FD, Bagozzi RP, Warshaw PR (1989) User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 35(8):982–1003
11. Liu IF, Chen MC, Sun YS, Wible D, Kuo CH (2010) Extending the TAM model to explore the factors that affect intention to use an online learning community. Comput Educ 54(2):600–610
12. Sanchez-Franco MJ (2010) WebCT—the quasimoderating effect of perceived affective quality on an extending technology acceptance model. Comput Educ 54(1):37–46
13. Fishbein M, Azjen I (1975) Belief, attitude, intention and behavior: an introduction to theory and research. Addison-Wesley, Reading
14. Chatzoglou PD, Sarigiannidis L, Vraimaki E, Diamantidis A (2009) Investigating Greek employees' intention to use web-based training. Comput Educ 53(3):877–889
15. Ringle CM, Wende S, Will A (2005) SmartPLS 2.0 (beta). Retrieved 22 Oct 2010 from http://www.smartpls.de
16. Teo T, Noyes J (2011) An assessment of the influence of perceived enjoyment and attitude on the intention to use technology among pre-service teachers: a structural equation modeling approach. Comput Educ 57(2):1645–1653

Designing an Interactive RFID Game System for Improving Students' Motivation in Mathematical Learning

Ho-Yuan Chen, Ding-Chau Wang, Chao-Chun Chen and Chien-Hung Liu

Abstract Game-based learning becomes a critical issue in the e-learning field. Many instructors want to make their students can do the studying with fun and their interest activities. For this reason, this paper designs an interactive RFID learning system for improving learners' motivation and performance by adopting game-based learning. There are several advantages for using this RFID learning system while learners behave well, such as learners will not feel they are engaging in a traditional learning environment when playing the RFID learning system. The purpose of this research was to develop an interactive RFID learning device and competitive learning environment for enhancing students' learning motivation and number sense in mathematics subject. The research is to investigate whether and how this RFID learning system can be developed to help users learn mathematics with enjoyment. This study considers ideas in game design, motivation issues, and mathematics learning to develop a strategy to engage users with the interactive RFID system. In this research, the learners can do the synchronic learning with other classmates outside of the classroom. Moreover, instructors can

H.-Y. Chen
Graduate School of Education, Chung Yuan Christian University, Jhongli,
Taiwan, Republic of China
e-mail: huc140@cycu.edu.tw

D.-C. Wang (✉)
Department of Information Management, Southern Tainan University,
Tainan, Taiwan, Republic of China
e-mail: dcwang@mail.stust.edu.tw

C.-C. Chen
Institute of Manufacturing Information and Systems, National Cheng-Kung
University, Tainan, Taiwan, Republic of China
e-mail: chaochun@mail.ncku.edu.tw

C.-H. Liu
Department of Network Multimedia Design, HsingKuo University of Management,
Tainan, Taiwan, Republic of China
e-mail: chliu@mail.hku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_24, © Springer Science+Business Media Dordrecht 2014

evaluate the individual's learning levels in mathematics in this research by analyzing the database of the game-based RFID learning system.

Keywords Interactive RFID applications · Game-based learning · Mathematics education · Number sense

Introduction

Clearly, instructional technology changes have influenced many educational activities, especially in the field of game-based learning and e-learning. In fact, instructors face a complex task in designing, developing, and evaluating e-learning courses, which include many different factors [7, 10, 11]. For this reason, program planners must consider several factors as they provide their learners with effective learning activities by using technology. Many researchers have pointed out that computer games as one kind of math learning tools with considerable potential for students to learn mathematics in game-based context. In reference to these concepts, we are developing the interactive RFID learning system which can be designed to assist students to learn mathematics.

We have to understand the composition of the interactive RFID learning system in order to design a game-based learning module first. Then we will discuss some concepts related to the computer-based learning. Moreover, we will describe several issues we faced to develop the interactive RFID learning system. Finally, the contributions of the RFID learning system in the game-based learning will be introduced. The research designs an interactive RFID learning system for improving learners' motivation and performance by adopting the car racing game-based learning. The RFID learning system meets individual's learning needs and provides various learning situations in the game-based learning activities.

There are several advantages for using a RFID learning system while learners behave well.

First, the interactive RFID learning system is a game-based learning device. Users will not feel they are engaging in a traditional learning environment when playing the RFID learning module.

Second, the competition system in the interactive RFID learning system is based on their prior knowledge in Mathematics. In this research, we designed a competition system as an example to deliver different level task-based questions according to the answer learners choose is right or wrong.

Third, many scholars have pointed out that distance education involves teachers and students separated by geographic and time factors [5]. The learners can do the synchronic learning with other classmates outside of the classroom, because we developed the interactive RFID learning system can be accessed through the internet.

The rest paper is organized as follows. Section System Architecture discusses the system architecture. We proposed the interactive RFID system designs for improving users' learning motivation, number sense, and performance in mathematics in section Designs of Game-Based Learning RFID Module. Then, section Demonstration shows the prototype implemented based on the interactive RFID system. Finally, we conclude our study in section Conclusion.

System Architecture

Figure 1 shows the reference architecture of the interactive RFID system. In the system, the database of interactive RFID learning system includes three factors: the ranking data of scores that users get, the inferential ability of the mathematics system, and the logic ability of the mathematics system. The logic ability of the mathematics system provided various logic questions for learners to figure out the right answers. The inferential ability of the mathematics system assists learners to learn the different concepts related to mathematics. Besides, the system can record users' learning processes for instructors to analyze the different learning behaviors and levels. In refer to this concept, instructors can meet individual's learning needs and situations based on this database.

Figure 2 shows the steering wheel which attached the RFID Tag. Learners can use the steering wheel device to choose the plus, minus, times, or divide when they are playing the game.

Designs of Game-Based Learning RFID Module

The game-based learning RFID system mainly include the RFID communication component (e.g., RFID reader and tag) and the database component. In this

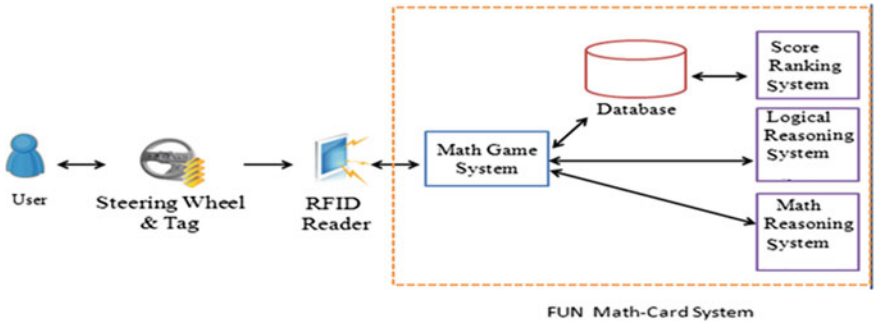


Fig. 1 System architecture diagram

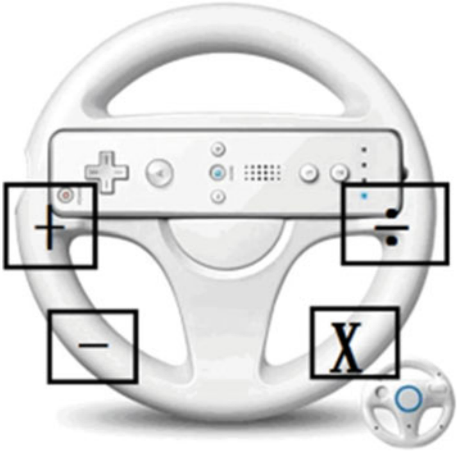


Fig. 2 The steering wheel which attached the RFID tag

section, we design the game-based learning RFID module in order to improve the users' learning motivation and number sense in mathematics field.

The Interactive RFID Learning Module

Figure 3 shows the interactive RFID learning system we designed. The RFID Reader can receive the information from the RFID Tag which is attached with steering wheel. Then, the RFID Reader delivers the learning data to the computer screen.

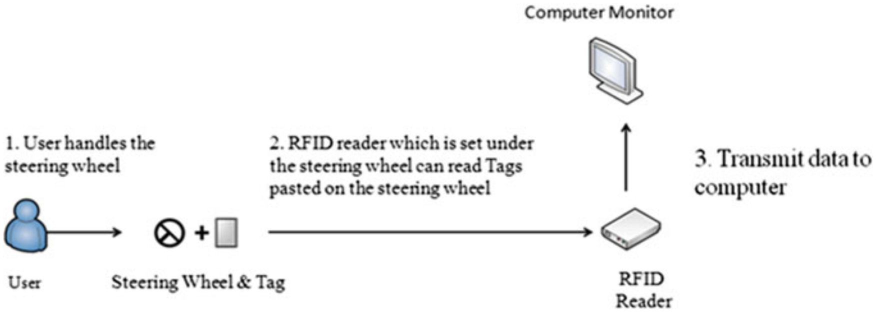

Computer Monitor
1. User handles the steering wheel
2. RFID reader which is set under the steering wheel can read Tags pasted on the steering wheel
3. Transmit data to computer
User
Steering Wheel & Tag
RFID Reader

Fig. 3 System operation diagram

The Playing Steps of the Interactive RFID System

There are several important play steps in this interactive RFID system:

Step 1 The screen of the computer will show the questions when users turn on the start button. Also, the car will be started to run in the computer interface when the users start to play the game.

Step 2 The right or wrong answers:

2.1 If the users choose the right answer, the interface will show the score of the question.

2.2 If the users choose the wrong answer, they could not get any score.

Step 3 The system will continue showing the new questions when users finish one question. The learners can have 5 min in each section.

Step 4 The users will receive the ranking of the scores from the computer screen when they finish each section.

Computer-Based Learning Environment

Computer-based learning (CBL) lets learners can learn new knowledge or skills from not only in traditional learning environment but also by computers. In the computer-based learning system, learners can practice themselves in anytime and anywhere. As the technology innovating, many instructors start to adopt computer-based learning in order to improve students' learning motivation [9].

Besides, many researches indicated that number sense of the learners had increased greatly by using computer-based learning tools [1, 4]. Moreover, the internet become more and more popular among out life, learners can do the competitive and collaborative learning activities via internet. The National Council of Teachers of Mathematics (NCTM 2000) indicated that the use of computer software can assist learners to study the mathematics.

Many researchers pointed out that game-base learning can attract learners' attentions and motivation [2, 9]. Also, the game-based learning can promote the math test performance more than the traditional learning approaches. In this research, the design strategy adopts a competitive game as the learning context in order to keep students' attention and motivate students to engage the learning activities.

In reference to this concept, we design the interactive RFID learning module to combine the game activities for learning mathematics. However, the design of effective game-based learning environment and to make learners to engage game-based learning environments is not simple. Moreover, motivation plays a critical role in the learning of mathematics [3]. Many researchers pointed out that the competition environment is an effective way to motivate learners to engage the learning environments [8]. For this reason, this research design a system for users

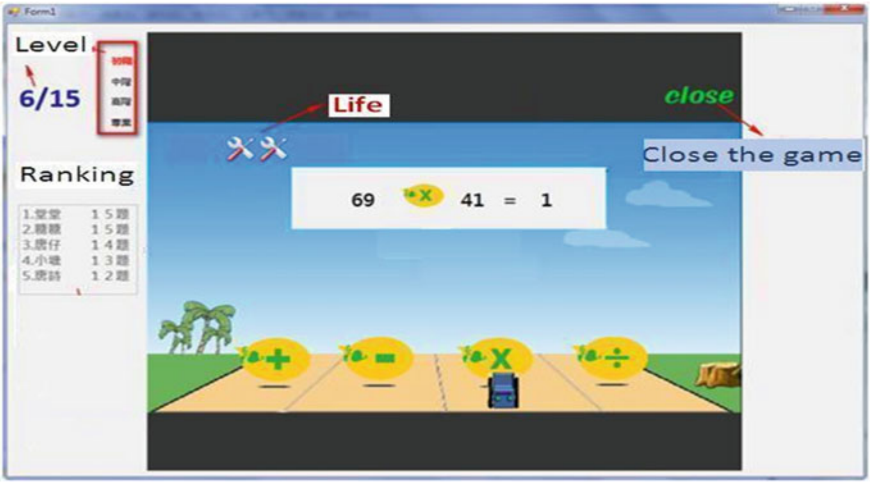


Fig. 4 Practice interface

which can help them receive the information related to the ranking of scores in this race game and who are also login at the same time. The main purpose of this research is to develop an online interactive RFID learning system on computer-based environment for learners competing themselves with other classmates. Learners will keep practicing the mathematical skills when they are trying to receive the higher ranking in this game-based learning environment. Finally, the users' records related to their learning processes of the interactive RFID learning system is recorded in the database of the system. In case good game-based learning quality is adopted in our research experiment. Moreover, we left the issue related to develop the system on mobile devices in future works.

Demonstration

We have developed the interactive RFID learning system for instructors to make students do the learning with fun. Figure 4 shows the main interface of the game-based learning system on computer. The interface will provide several information included questions, different sections, and the ranking. In educational field, games usually are learner-center and have several important characteristics such as rules design, competition, challenging activities, choices, and the ranking of the score points. One of the primary advantages of the interactive RFID learning system is that they can play the game in competition environments. Game-based learning environment has the potential to improve students' number sense in mathematics education. In this research, the interactive RFID learning system can automatically deliver appropriate feedback related to the answer is wrong or right in the interface of the race game. Each user has five chances to finish one question. This interactive RFID learning device can be used with a wide range of users along a continuum of different ability levels of mathematics. Learners can move to the advance level sections when they chance the right answer. If they could not complete the question during the five chances, the interactive RFID learning system would automatically deliver the other same level question for users to practice again. Besides, the users can check their ranking information in this result interface. Moreover, the interactive learning system will increase the speed of the racing car after the user complete one section in order to improve learners' number sense in the game-based learning context.

Conclusions

This research implemented a game-based learning environment by using an interactive RFID learning system to improve learners' learning motivation, number sense, learning performance, and also collected three different kinds of data: the logic ability of the mathematics, the inferential ability of the mathematics, and the ranking data of scores.

For the instructors, instructors usually adopt the exam to evaluate the student's learning level in the ending of the course in the traditional learning environment. However, teachers can analyze the database of the game-based learning system in order to realize the individual's learning levels in mathematics in this research and also realize how the game system can be develop to assist students to practice their number sense. Besides, the instructors can utilize the interactive RFID learning system to develop the self-learning game to increase learners' motivation.

In refer to the learners, the interactive RFID learning system is fun and will not make users feel they are studying or testing. Users can feel they are engaging a racing game and have the opportunity to compete with other classmates. The game-based learning becomes an important issue in e-learning field in order to make learners can do learning with fun and their interest subjects. In this research, learner could be encouraged to contribute more times to do the game-based learning when they get higher ranking of scores more than other classmates. At the same time, learners can improve their number sense and approach to the solution of a given mathematical questions.

Future Work

There are several future works needed to consider to improve this research. First, the mobile device technology has become very popular nowadays. In the e-learning field, the basic use of information communication technologies innovation is to assist both instructors and students to engage in interactive educational opportunities across many barriers.

In regard to these reasons, we should combine the online interactive RFID learning system into mobile device. Learners can play the game with others by using smart phone or e-pad. Besides, the learners can utilize their mobile device to practice the skills they received or to learn more information and knowledge in this field.

The other future work is to develop the system related to the educational feedback and reward system. In traditional web-based learning environment, learners usually receive the feedback from the learning module or instructor, such as the ranking of scores, a summary of the learning performance, or provide the information for them to challenge the next difficult level questions or tasks. However, the reward system is not only the one of feedbacks but also is an efficient way to keep learners' motivation and psychological needs.

Acknowledgments Thanks to Chin-Yin Lin and Yen-Ju Tsai, students at Southern Taiwan University, for writing code to establish the system, and Huei-Err Hsu, Si-Yu Su, and Pin-Juin Chen for designing and drawing the user interface.

References

1. Cavanagh S (2008) Playing games in class helps students grasp math. Educ Digest: Essent Readings Condens Quick Rev 74(3):43–46
2. Ke F (2008) Alternative goal structures for computer game-based learning. Int J Comput Support Collaborative Learn 3(4):429–445
3. Ma X, Kishor N (1997) Attitude toward self, social factors, and achievement in mathematics: a meta-analytic review. Educ Psychol Rev 9(2):89–120
4. Miller D, Brown A, Robinson L (2002) Widgets on the Web: using computer-based learning tools. Teach Except Child 35(2-):24–28
5. Moore MG, Kearsley G (1996) Distance education: a systems view. Wadsworth, Belmont
6. National Research Council (2000) How people learn: Brain, mind, experience, and school. National Academy Press, Washington, DC
7. Pearson J, Trinidad S (2005) An instrument for refining the design of e-learning environments. J Comput Assist Learn 21:396–404
8. Schwabe G, Göth C (2005) Mobile learning with a mobile game: design and motivational effects. J Comput Assist Learn 21(3):204–216. doi:10.1111/j.1365-2729.2005.00128.x
9. Sedig K (2008) From play to thoughtful learning: a design strategy to engage children with mathematical representations. J Comput Math Sci Teach 27(1):65–101
10. Thurmond VA, Wambach K, Connors H, Frey B (2002) Evaluation of student satisfaction: determining the impact of a Web-based environment by controlling for student characteristics. Am J Distance Educ 16(3):169–189
11. Trinidad S, Aldridge J, Fraser B (2005) Development and use of an online learning environment survey. J Educ Technol 21(1):60–81

Part VI
Multimedia Technology for Education

Design and Development of an Innovation Product Engineering Process Curriculum at Peking University

Win-Bin Huang, Junjie Shang, Jiang Chen, Yanyi Huang, Ge Li and Haixia Zhang

Abstract Over the past decade, the elements of innovative wisdom are not only a business organization to survive in the dangerous environment, school organizations to enhance the quality of education to meet the needs of the community. For the succession challenge in the continuous impact, an innovation level of a top university covered the administration, curriculum, teaching, equipment, environment and so on is quite extensive. This paper overviews a novel curriculum at Peking University, called Innovation Product Engineering Process, established by six interdisciplinary teachers for school students in various professional fields. The curriculum aims at inspiring students to break through professional limitations for experiencing the innovation process from idea into product. The students are self-organized as a team and construct a prototype collaboratively. Instructors from industrial give a practical perspective lesson and provide market information, funding and technical support. Students in the course are fostered six expected abilities, including creativity, practical, engineering process, team-working, communication and expressiveness. Ideas from students become the topic of a project after competing in three eliminating rounds. All competitions are graded and ranked by the participators (teachers, students, instructors from academic and

W.-B. Huang (✉)
Department of Information Management, Peking University, Beijing 100871, China
e-mail: sebastian.huangwb@gmail.com

J. Shang
Graduate School of Education, Peking University, Beijing 100871, China

J. Chen
Department of Electronics, Peking University, Beijing 100871, China

Y. Huang
College of Engineering, Peking University, Beijing 100871, China

G. Li
Department of Computer Science, Peking University, Beijing 100871, China

H. Zhang
Institute of Microelectronics, Peking University, Beijing 100871, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_25, © Springer Science+Business Media Dordrecht 2014

industry). Finally, six ideas having the opportunity to become prototypes are developed successfully with various properties. Most of students indicated that the curriculum provided them a new training experience, interesting learning style and useful content of courses.

Keywords Workforce development · Product engineering process · Course development

Introduction

Cultivating and training innovative talents is certainly important in order to building an innovative country and improving international competitiveness. School students having high social resilience and entrepreneurial capacity is one of the aims of the long-term education reformation and development program from 2010 to 2020 in China. It should be achieved through education and training of various fundamental science, research and practice. High level education in engineering, besides, in China mainly fosters people translating science and technology knowledge into productive power [1]. Developing students' innovation through engineering practice, integrating technology, humanities, economics and management knowledge, is also an important part of advance engineering education. The interdisciplinary cooperation, however, is few and far not only between academics but also between colleges and enterprises. Advance engineering education, therefore, should be oriented strenuously towards the practice of engineering process. The integrated engineering activities training a student solving complex, comprehensive and interdisciplinary problem collaboratively with different professionals have become the key point of education around the world.

Currently, the reformation of engineering practice is initialed in US and few academics have set up the courses for inspiring students in product processing engineering. The Conceive Design Implement Operate (CDIO) project [2] in Massachusetts Institute of Technology is regarded as the representative in this teaching model and its teaching manner provides students to experience a life cycle of a product process in the real environment. Project-based learning (PBL) in engineering practice is adopted in Canada [3]. Students as a team to completing the production of the specified projects put the theories and methods what they have learned in use. Different from the traditional curriculum, learning assessment is not only dominated teachers but taking a combination of student self-evaluation and peer evaluation by team members. Evaluation system and the way of the reformation of engineering practice courses, as a result, is an advisable merit of the PBL. The reformation in British perform practice course intending to improve the learning experiences of the science and engineering student to be "engineer" instead of "student." Furthermore, the engineers in industry are invited to give lectures or workshops for importing working experiences and instructing interdisciplinary exercises. Academics also connect a two-way interaction with industries in order to understand the demand for graduates. The courses recently increase the number of knowledge in humanities, economics and social sciences [4].

In recent years, the reformation of engineering education and practice are speeded up in China, and a lot of improvements on student participation and engineering practice are achieved. Comparing with the developed countries, however, two problems as follows are still critical. (1) The course of engineering and practice mostly focus on technical contents without much attention on general educations. (2) Teaching proportion of the arrangement in a course is unreasonable and the importance of a teacher in engineering practice education is over-valued. Consequently, students emphasize technical tools and skills too much during their learning in education. Self-innovation of students is probably disappeared under the situation of education. In order to increase the creativity and innovation of students at Peking University, the co-authors from different fields open a novel and interesting course with the following originality in China.

- Inviting engineers and managers from industry to sharing market information, providing finance, importing experiences and so on.
- Providing comprehensive knowledge including social science, engineering and management to students.
- Fostering students to experience product engineering process through completing a project.
- Allowing students to participate in the evaluation system of the course and scoring an idea or a work with the "press" equipment.

According to the outcome of the course and feedback from students, well interaction among teachers and students is observed, and curiosity about engineering and practice capability of students are improved. Students also understand and respect the opinions and works of others. School members comprehend the needs of industries and the discipline of the projects includes chemistry, computer science, communication, signal processing, the design of application on tablet PC or smart phone and so on. At the end of lesson, a company in China even intends to invest its resource in one of them as a product in the future.

Course Description

The course being team taught by the co-authors of this paper and titled, "Innovation Product Engineering Process," at Peking University has the following purposes in creating this course:

- Design a course integrating topics from interdisciplinary content relevant to product engineering process.

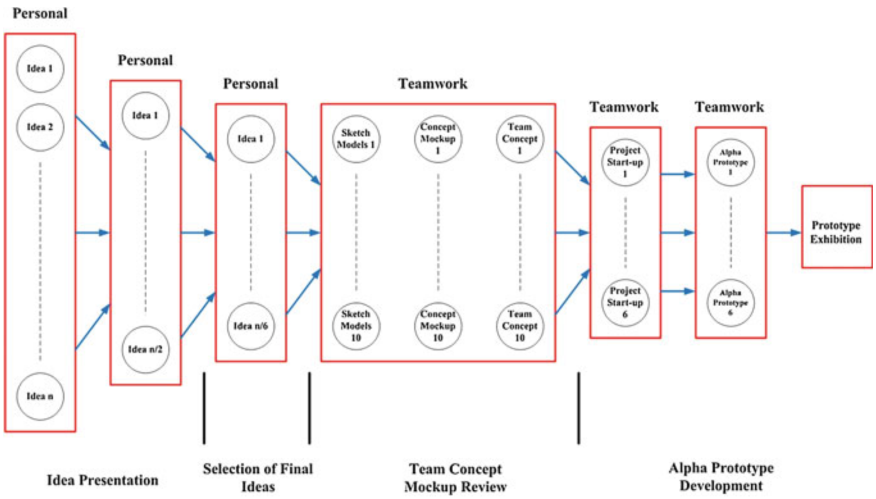


Fig. 1 A high-level schematic of the project workflow in the course is delineated

- Design a course to target audience of undergraduate students of various professional fields.
- Design a course with a teaching team including members in academics and industries.
- Design a course to provide students funding and technical supports to experience from idea creation to prototype making collaboratively.
- Design an evaluation system to allow students participating in.

A high-level schematic of the project workflow is shown in Fig. 1. The process leading to a functional and workable prototype is separated into four major milestones: the ideas presentation; selection of final ideas; team concept mockup review; and prototype development. Every student, at first, in the course creates more than an idea, and these initial ideas are examined in three eliminated round. In the stage, ideas are made under students' brain storming, surveying literature and sharing information. Ten initial ideas then become the major topics of a team project—they work in full cooperation and virtually coordinating resource as appropriate before team concept mockup review. The goal of the team concept mockup review is to inform instructors about the state of the project's functional, the concept of final prototype. This feedback also provides each responsible team prioritize improvements for the final presentation. Course content coverage is described as follows:

- Ideas Presentation
  - Week 1: Overview and introduction to innovation, creativity and entrepreneurship [Lecture].

  - Week 2: All students share personal ideas as more as possible and make a defense to others' interrogation. Here 50 % of the ideas are eliminated.
  - Week 3: Research and design process in an innovated project. [Lecture]
  - Week 4: The students survived in the last competition represent the selected idea in 3 min specifically and also make a defense to others' interrogation. Here 30 % of ideas are collected to next eliminating rounds.
- Selection of Final Ideas
  - Week 5: Fundamentals of proposal writing, and how to make a business plan. [Lecture].
  - Week 6: Each one of the remainder ideas should be represented clearly and specifically in 10 min, including requirement, application and scenario. The students initiated these ideas make a defense to others' interrogation and only ten ideas are selected.
  - Week 7: A special workshop is hold in the class for interaction freely among the original designer and other classmates, and then making a team finally. The designer with his team members, moreover, develops sketch, technical drawing and preliminary plan of the idea as a project.
  - Week 8: Each team has 15 min to make a detail presentation on their preliminary project, and the responses to others' interrogation are also considered. In the end, only six teams are obtained the opportunity to implement their idea with all supports, such as finance, technical, laboratory. Besides, a school teacher is responsible for technical advising, trouble shooting and schedule control of a team selected. Each team also has a budget of near ¥ 3,000, depending on the discussion of all school teachers, to purchase materials, supplies, and resources for the project. The members of the eliminated teams are separated themselves into the succeed teams.
- Team Concept Mockup Review
  - Week 9: The relation between technology and product: why failed? [Lecture].
  - Week 10: A discussion is hold for interaction, exchange and information sharing among all members in the class.
  - Week 11: Management and control of a project. [Lecture].
  - Week 12: A discussion is hold for interaction, exchange and information sharing among all members in the class.
  - Week 13: Leadership, communication and exchange. [Lecture].
- Alpha Prototype Development
  - Week 14: Each team must plan and work to keep their projects on budget and on scheduled. Moreover, they have to report their progress, balance of appropriations and technical detail. The teachers consider budget extensions while a budget overrun for further completing the project well is required.
  - Week 15–18: Each team keeps doing their project and reports to responsible school guidance. Furthermore, a few lectures and discussions of which the speakers are invited from enterprise or industry are held from time to time.

- Week 19: A whole-school exhibition is held for final presentation made by the team's representative and the demonstration of its alpha prototype in interesting, funny, formal or surprising way is necessary.

The lectures with different topics are given by different professional co-authors, engineers, and managers from industry. This course was offered for the first time in spring 2013 and met class once a week with each lecture and workshop being approximately 180-mins. Student's feedback and evaluation of course will be taken into account to improve the course contents and organization in future offerings. Students' performance on idea propagation, teamwork performance, system design, and final presentation is considered as the measure of this course.

Project Budget

Most of financial support in the course is provided by Peking University. In the stage of selection of final ideas, the teachers determine the practical financial needs according to the proposal, scheduling, scenario and development requirements. The final budget of a team is accepted with over 50 % guidance's agreements. On average, each team has a budget of ¥ 3,000 to purchase materials, supplies, and resources for the project. Each team member must participate in planning and keeping their projects on budget and on scheduled. The project budget is not compliant after verdict in the class. The teachers, however, consider budget extensions of a project while a budget overrun for further completion is required. There is only a chance to add their budget up while reviewing the mockup of team concept. Each team member will pay an equal portion of deficit if their budget is overrun. Moreover, each team is allowed to have sponsorship fee from enterprises.

Grading

The overall score of a student is graded based on four parts with its proportion: idea propagation (25 %), teamwork performance (25 %), system design (40 %), final presentation (10 %). Before forming a team, the performance of a student is evaluated as a partial personal grade. All members including teachers and students in the course vote pass or fail to the idea into next eliminated round, and the assistant instructor calculates its score. Here customer needs, thoughtfulness, clarity and quality of the design alternative of the ideas are considered. Once the six ideas are determined as the major content of a team, personal score of a student is graded in the "idea propagation" part. After that, the review contributes to a portion of a shared team-wide grade and members of course or team participate in the rest review process. Key grading critical in teamwork performance are operating, activity, workload and communication in coordination. The score of a student in this part is graded by all team members and guidance. Furthermore, all members in the course participate in grading a team in system design, of which key critical contains mechanical design details, system integration, details of prototype execution and manufacturing. In the final presentation, it takes place in one day of the summer vacation and provides each team with the opportunity to show their works to various audience including academics and industrial visitors. All participators in the exhibition evaluate a team's work based on team's performance in customer data, market information, specifications, or benchmarks for the product.

Prototype Exhibition

The final milestone in the course is a formal presentation which is attended by the overall members in the course, all guidance, sponsors, and guests from academic and industry. A portion of the shared team-wide grade is contributed in the exhibition. Each team has the opportunity to demonstrate their work to all participators. Students may learn how to prepare a complete technical presentation in a life-styled, educated, technical, or business oriented way. A team is also allowed to seek investors for their product to start-up a company. Each team is evaluated and graded by all participators based on its presentation quality, business assessment; technology, the prototype, and overall potential to become a real product.

Summary

A new curriculum combining innovation, learning, co-operation and practice is in development at Peking University in line with the need of interdisciplinary training to product engineering process. This course covers creativity, teamwork, management and practice of a project realization emphasizing pioneering aspects. After taking this course, students are expected to contribute to the start-up aspect of industrial projects related to product engineering. Students will be able to understand vulnerabilities and difficulties to the product engineering process in addition to realizing the basic principles of project workflow. Students are expected to critically analyze the interdependencies of related workflow in product engineering process and apply the interdisciplinary principles that they have learned in starting a practical idea up. After the course, many meaningful and useful prototypes created by the students are impressive.

Acknowledgements Authors would like to thank Peking University for financially supporting the work reported in this paper and the experienced lecturers from academic and industry for a professional speech.

References

1. Chen Jin (2010) Building the innovative country: Theory and practice. Science Press
2. Crawley EF (2001) The CDIO Syllabus. A statement of goals for Undergraduate Engineering Education
3. Michel J (2009) Management of change-implementation of problem-based and project-based learning in engineering. Eur J Eng Educ 34(6):606
4. The Imperial Study Guide [EB/OL] (2011) Imperial, College London pp 7–12

The Design of an Educational Game for Mobile Devices

Daniela Giordano and Francesco Maiorana

Abstract The importance of computing education is well known across different fields from STEM to Computer Science, from Humanities to Social Science. Educating the younger generation to 21st century skills is advocated by many international organizations since these skills can be used across several disciplines. Shifting from educating students to be user of a software tool to be designer of customization of existing tool to their needs or even creators of software artifacts designed around specific needs is deemed the major challenge that educators are facing. This paper describes an educational game that by using modern and appealing technologies such as smartphones and mobile devices presents a game that interleaves ludic and educational aspects. The paper describes the main design goals of an educational game as well as the educational design of the learning path centered around a set of topics organized in different levels. The levels are designed in accordance with the Bloom taxonomy and each level has different stages with increasing grade of difficulties.

Keywords Educational game · Game design · Mobile educational game

Introduction

The importance of acquiring a set of basic competencies and abilities in science, technology, engineering, and mathematics (STEM) education as well as in Computer Science education is internationally recognized. Official documents

D. Giordano (✉) · F. Maiorana
Department of Electrical, Electronic and Computer Engineering, University of Catania, Catania, Italy
e-mail: daniela.giordano@dieei.unict.it

F. Maiorana
e-mail: francesco.maiorana@dieei.unict.it

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_26, © Springer Science+Business Media Dordrecht 2014

such as [1–3] emphasize the importance of acquiring an in-depth knowledge of the fundamental concepts around Computer Science and underlines the necessity of a suitable acquisition of "21st century skills" such as problem solving and critical and creative thinking, as well as communication and cooperation skills.

The "CS principles project" [4], for example, has designed and put into practice, both at school and university level, a curriculum built around seven central ideas:

1. computing is a creative activity,
2. abstraction reduces information and detail to facilitate focus on relevant concepts,
3. data and information facilitate the creation of knowledge,
4. algorithms are used to develop and express solutions to computational problems,
5. programming enables problem solving, human expression and creation of knowledge,
6. the internet pervades modern computing
7. computing has global impacts.

The project aims to teach to a large audience of students, without restricting to small elective courses, the basic concepts of Computer Science in such a way to make these concepts a common basic background that can be used by younger generation in all their further study both in STEM education as well as humanities.

For these reasons there is the necessity to modify both the content of the curricula and the pedagogy and the way of teaching these new materials. These modifications are necessary since the first years of school in such a way to transmit and educate, as soon as possible, not only the ability related to the use of computers and software tools, such as internet, e-mail, text writing or tools to manipulate data, but also creative capacity of designing and implementation of software tools able to resolve a given problem at hand. Once the skill to design and implement a software artifact are acquired, they can be used to customize existing software tools to personal work-related needs and to personal fields of interest.

The experimentation concerning the curriculum is focused on content aiming at developing, through modern instruments, the above mentioned capacities. Modern pedagogies are centered around constructivist theories that put the students at the center, privileging their central and active role. The students should no more be passive listeners of the lessons but have to participate, inside a group of peers, to knowledge creation through the realization of practical projects.

Some pedagogical theories such as the "inverted classroom" [5] banish passive activities of lecturing, by delegating learning to homework and concentrating all class activities on laboratory projects or the development of practical activities.

A necessary tool to customize existing software or to develop new software is represented by a programming language. The knowledge of one or more programming languages and of the techniques to design and implement algorithms, to choose the best and most suitable data structures represents one of the greatest obstacles in the introductory programming course both in high school and in first level University courses.

There is an sample scientific literature dealing with the problems related to teaching an introductory programming course. A recent review can be found in [6] where the authors describe the curricula, the teaching pedagogies, the choice of the programming languages and the tools that can be useful in teaching.

In general, all the introductory programming courses have one of the highest levels of drop out or poor results. In [7] for example, the authors have analyzed the results of 67 international institutions, arriving at the conclusion that, as a general rule, failure percentage is between 30 and 60 % in courses in introductory programming with relatively large classes.

In order to overcome such difficulties one of the possible approaches is to use educational games inside the educational path within a course of an entire curriculum both in school and universities.

As stated in [8] it is natural to combine the content of a learning path with the great motivation and attraction that the educational game have on students and in particular in teenagers. Nevertheless, in a recent literature review the authors point out the need of more quantitative studies on the ability of educational games to foster greater reasoning skills [9]. It is, hence, necessary to collect more quantitative data in order to carry out analytical studies on the validity of educational game as tools supporting education and modern pedagogies.

Another pressure in the direction of technology innovation is represented by the use of smartphones and mobile devices. It is clearly stated in the literature that nowadays the majority of the population in developed countries has access to a mobile network. A recent study [10] reports that since 2010, 90 % of the population has access to a mobile network. The percentage increases if the sample population is restricted to teenagers and younger people, who use mobile devices in many daily activities such e-mail reading, internet access, use of chat, social network access, storing and sharing photos and so on. Younger students, moreover, are used to always bringing a smartphone or a mobile device along with them. These types of technologies have a strong appeal on younger generations

In the light of the above consideration it is natural to use these new technologies in the educational field.

This work has the aim to describe the design and implementation of an educational game that blends entertainment and game play with educational questions. The educational questions have been designed to guide the students in the learning path typical of an introductory programming course both for a major in Computer Science and for non-majors. The game has been designed to be utilized with mobile devices and in particular with smartphones. The game can be adapted for use in STEM courses or in the humanities, or in high school or university, by allowing the personalization of the multiple choice questions. The educational game and the questions have been designed in such a way as to guide the students in their learning path.

This paper is organized as follows: section Game and questions design briefly describes the design considerations both for the game and for the questions, section three describes the game and implementation details, section Game evaluation plan presents an outline of a questionnaire to be administered to the students and the expected outcome from the analysis of both the qualitative and quantitative data gathered from the user answers and from the log data gathered during game play in a client server version of the game that we plan to develop, section Conclusions and further work draws some conclusions and highlights future work.

Game and Questions Design

In designing the game the following aspects were taken into consideration:

- The game should be designed for small mobile devices and in particular for smartphones
- The game should blend recreational aspects with educational activities in the form of questions related to one or more disciplines
- The game should provide, upon request, immediate feedback to the user. In order to avoid abuse of the feedback mechanism and stimulate self-discovery, the feedback should have a cost for the user in terms of scores
- The educational aspects of the game should be embedded in questions carefully designed in order to guide the users, through different level of difficulties, in their learning path.
- The game play should be customizable to the user needs and pace. This customization should account for different levels of initial knowledge and different expectations and goals.
- The game should provide progressive levels of difficulties both in the ludic and educational aspects.

The design of the questions should be organized into topics and each topic in levels, and each level in an increasing grade of difficulty. For example, with an introductory programming course in mind, the topics can be the main part of a procedural language such as: input instructions, variable declarations, arithmetical, relational and logical operators, expressions, input instructions, procedures and functions, conditional instructions, cycles, array, matrices, recursion and linear and non-linear data structure. The level can be, in accordance with the Bloom Taxonomy of cognitive processes [11] organized into the following scale:

1. Find syntactic errors. This type of error is detected by a compiler and is used to test the knowledge of the syntactic structure of the programming language. This type of question represents the lowest level of the Bloom taxonomy.
2. Program understanding: typical question are related to guess the output of a piece of code.
3. Design procedure and functions in terms of their parameters or output value
4. Sort a sequence of instructions in order to obtain a correct algorithm

5. Find semantic errors: given a program description and a piece of code find the semantic errors.
6. Complete a fragment of code with a missing part. This can be framed into the highest level of the Bloom taxonomy.

There are several means to increase the grade of difficulties such as increasing the number of possible answers in a multiple choice game, increasing the number of missing parts or providing more possible answers to choose from, and so on.

Game Implementation

The first prototype of the game was implemented using Eclipse [12], the Android Software Development Kit (SDK) [13] and the Android Development Tools (ADT) plugin, in particular the Android Api level8 and the SqLite [14] database to store the questions.

The game can be framed as a shooting game: two types of characters appear on the screen: good and bad ones, along with other distracting elements. These characters move randomly on the screen. The player has to touch on the screen all the good characters in a fixed amount of time. At increasing levels, the number of characters increases and so the game difficulty. If all the bad characters are not hit in the amount of time the player loses a life.

According to the game level, each time a fixed number of good characters are hit a multiple choice question appears on the screen. The level of the questions is chosen in accordance with the game progression and the above sketched hierarchy.

The game can be played as a complete game from start to end, or the user can choose a level as a starting point in order to customize the learning path and progression. Figure 1 shows a screenshot of the game in the shooting mode.

In Fig. 1 the good and bad characters are also differentiated by a different color of the bounding rectangle.

Figure 2 shows an example of the game in question mode. The game was designed with an object-oriented approach. A database stores the questions. The questions can also be uploaded from a file where the fields of each questions are separated by commas, thus allowing easy customization of the game to different domains.

Game Evaluation Plan

In order to evaluate the game we plan to gather both qualitative and quantitative data. The qualitative data will be gathered through a questionnaire administered to educators and to the game players. The questionnaire will gather data in accordance with the evaluation model presented in Table 1. The table reports the main

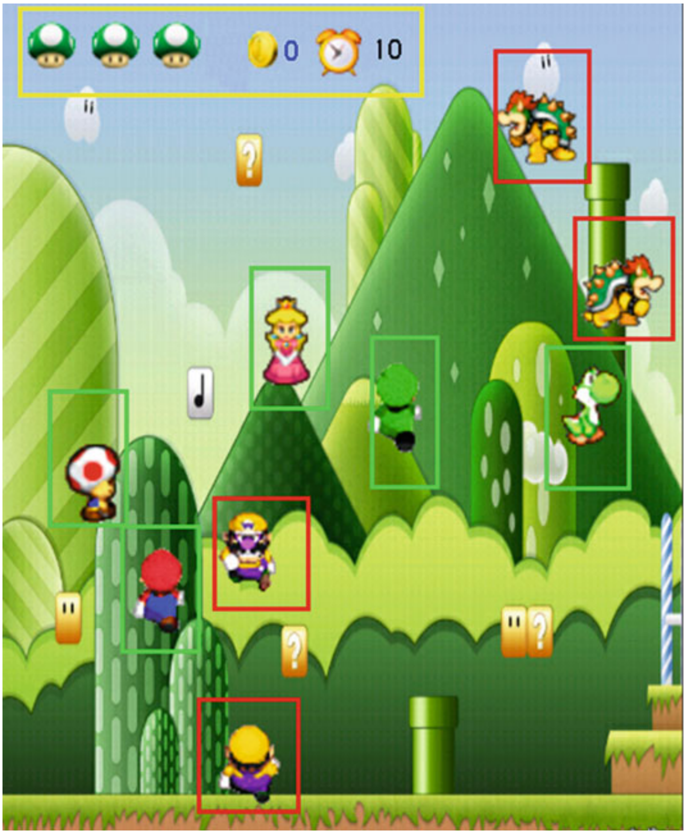
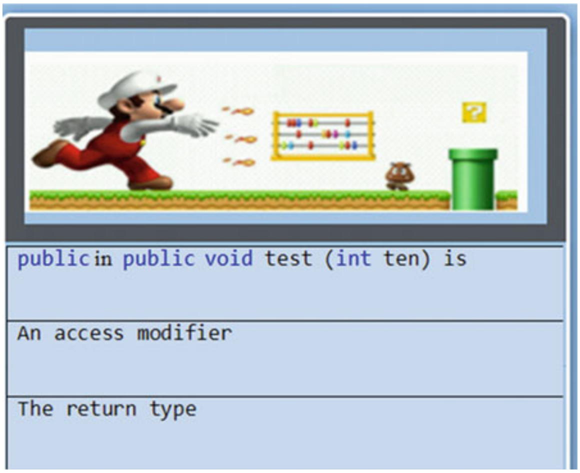


Fig. 1 A screenshot of the game in "game mode"

Fig. 2 A screenshot of the game in "question mode"



public in public void test (int ten) is

An access modifier

The return type

Table 1 Evaluation model

Dimension	Type of data	Expected results
Game design	Questionnaire aimed at collecting data on the game usability, playability and interactivity.	Evaluation of the usability, playability and interactivity.
Question design	Data gathered from student response and log data. Questionnaire aimed at collecting data about question difficulties and cognitive load and perceived effort by the users. Data gathered on the explanation of the cognitive strategies and process followed by the player in answering the questions	Using classical test theory and item response theory to validate both the single item and the overall quality of the questions
Effects on learning	Pre-test post-test and retention test. Questionnaire asking the perceived effectiveness of the learning path followed during the game	Evaluation of the effects on learning and the learning process.

aspects of the investigation, the type of data gathered and the main results expected from the data analysis process. In particular, the quantitative analysis will be based on the log data, collecting information such as the time spent playing the game, the time spent to answer each question, the number of corrections and so on. Both the quantitative and qualitative data can give an indication of the effects on learner motivation; on the quality of the questions; and on the effect on learning.

Conclusions and Further Work

This work has presented the main design principle of an educational game that by posing questions arranged around topics, with each topic divided into different levels in accordance with the Bloom taxonomy, and each level with a different grade of difficulty allows for a customizable environment that, even it has been designed for initial programming courses, can be used, with a careful design of the questions, in different domains ranging from STEM to languages courses or humanities.

As further work we plan to fully develop the database of questions and to use the game as a study aid and as an assessment tool in an introductory programming course for non-majors. At a more advanced level, not only textual but also graphical questions could be posed to the users, so that they may also develop some skills in designing the human-computer interfaces of modern information systems, in which system functionalities should be controlled by suitable multimedia interfaces, as suggested in [15]. Each questionnaire may be stored in a server with a short abstract and keywords so that the educational material may be clustered to give rise to an organizational memory, as envisaged in [16–19], that allows the users to download the game most suitable for their educational needs of the user, and also may facilitate interaction across peers, as in a social network, to select the more engaging games, and eventually, to create a channel for exchanging answers. This approach would also support a pedagogy oriented to social learning, and could be easily implemented by foreseeing in the game also some open questions with no feedback. Analysis of the game usage and answering paths of the students can be used to obtain a deeper insight on the main student difficulties in getting acquainted with programming skills, which is nowadays an important aspect of education.

References

1. ACM & CSTA (2010) Running on empty: The failure to teach K–12 computer science in the digital age
2. ACM-IEEE (2012) Computer science curricula 2013 Strawman Draft
3. Royal Society (2012) Shut down or restart: the way forward for computing in UK schools, January, 12
4. Astrachan O, Briggs A (2012) The CS principle projects. ACM Inroads 3(2):38–42
5. Lage MJ, Platt GJ, Treglia M (2000) Inverting the classroom: a gateway to creating an inclusive learning environment. J Econ Educ 3(1):30–43
6. Pears A, Seidman S, Malmi L, Mannila L, Adams E, Bennedsen J et al (2007) A survey of literature on the teaching of introductory programming. SIGCSE Bull 39(4):204–223
7. Bennedsen J, Caspersen ME (2007) Failure rates in introductory programming. SIGCSE Bulletin 39(2):32–36
8. Prensky M, Prensky M (2003) Digital game-based learning. Mcgraw Hill Book Co
9. Connolly TM, Boyle EA, MacArthur E, Hainey T, Boyle JM (2012) A systematic literature review of empirical evidence on computer games and serious games. [doi: 10.1016/j.compedu.2012.03.004]. Comput Educ 59(2):661–686
10. Johnson L, Smith R, Willis H, Levine A, Haywood K (2011) The 2011 horizon report. The New Media Consortium, Austin, Texas
11. Anderson LW, Krathwohl DR, Airasian PW, Bloom BS, Cruikshank KA, Pintrich PR, Mayer RE (2001) A taxonomy for learning, teaching and assessing, pp 67–68. Addison Wesley Longman, Inc, complete edn
12. Eclipse available at http://www.eclipse.org/
13. Android SDK available at http://developer.android.com/sdk/index.html
14. SqLite database available at http://www.sqlite.org/download.html
15. Faro A, Giordano D (2000) Ontology, esthetics and creativity at the crossroads in information system design. knowledge-based systems, 13(7–8), (1 December 2000), 515–525
16. Faro A, Giordano D (1998) Concept formation from design cases: why reusing experience and why not. Knowl-Based Syst 11(7):437–448
17. Faro A, Giordano D (1998) StoryNet: an evolving network of cases to learn information systems design. In Software, IEE Proc 145(4):119–127, IET
18. Giordano D (2004) Shared values as anchors of a learning community: a case study in information systems design. J Educ Media 29(3):213–227
19. Giordano D Evolution of interactive graphical representations into a design language: a distributed cognition account. Int J Human-Comput Stud 57(4):317–345
20. Faro A, Giordano D (2003) Design memories as evolutionary systems socio-technical architecture and genetics. ProcIEEE Int Conf Syst Man Cybernetics 5:4334–4339

Activating Natural Science Learning by Augmented Reality and Indoor Positioning Technology

Tien-Chi Huang, Yu-Wen Chou, Yu Shu and Ting-Chieh Yeh

Abstract In recent years, with the rapid development of information technology, educational technologies have been used successfully in enriching learning content and improve learning efficiency. This study attempts to develop an assisted learning system for increasing students' motivation by creating flexible learning path. The purpose of the designed system is to bridging the gap between formal and informal learning on natural science subject. Augmented reality and indoor positioning technologies have been implemented in the system to guide learners to construct their knowledge in the informal learning environment, National Museum of Natural Science. Learners can not only receive virtual information left by other learners in such space, but also leave their own learning experience and share with others. Additionally, the system also analyzes individualized learning subject. By doing so, other learners who have the same interests could find an efficient way to learn.

Keywords Educational technology · Augmented reality · Indoor positioning · Self-regulated learning · Formal and informal learning

T.-C. Huang (✉) · Y.-W. Chou · T.-C. Yeh
Department of Information Management, National Taichung University of Science and Technology, Taichung, Taiwan, Republic of China
e-mail: tchuang@nutc.edu.tw

Y.-W. Chou
e-mail: s1801B103@nutc.edu.tw

T.-C. Yeh
e-mail: s13013048@nutc.edu.tw

Y. Shu
Taichung Shinmin Senior High School, Taichung, Taiwan, Republic of China
e-mail: h1257@shinmin.tc.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_27, © Springer Science+Business Media Dordrecht 2014

Introduction

Classroom learning at school is a way of formal learning for students; however abundant resources outside of the classroom cannot be ignored. For example, informal learning in the National Museum of Natural Science (NMNS) combines both formal and informal learning to create a diverse learning, enhancing student's ability of self-regulated learning. Nevertheless, visitors may get lost in the spacious museum so guides are required to lead the route and to explain to the visitors. The same applies to students getting educated at this site. The fact is that not every site has a guide to explain to the visitors, or else a guide may be subject to limited time; also a guide may not be suitable for the learning path for everyone.

Usually when students are in the face of new knowledge, they feel at loss because they know nothing about the topics covered and they can only explore by guessing. Consequently, students spend too much effort during discovery process in search of the topic or may even digress from the topic, resulting in ineffective self-regulated learning. If there is one guideline to enable the students to take less pointless routes and put more effort to explore the knowledge more deeply, they will have better performance on learning. It is not the case that these knowledge are never been discovered, but only that current museum of science has not recorded the history study log systematically to share the experience with others. Hence future generations can only rely on self-learning ability. When knowledge sharing is not available, students may struggle and get lost in searching of the topics, eventually losing interest in learning and then giving up learning.

Therefore, how to establish one knowledge-learning platform to help future generations to follow their predecessors' footsteps for learning, and further aid self-regulated learning is the main goal of this study. A mobile carrier is used to assist students' learning at museum of science. Two additional technologies have been adopted with the mobile carrier. Indoor positioning is adopted to calculate the current position of the learner; AR (augmented reality) technology is then adopted to gradually guide the learning route. The learning route is defined from a learning mode analyzed in accordance with previous learning experience, aiming to reduce time cost in groundless searching. The platform records the learning process of each person with added nodes to further monitor learning progress, as well to share with other learning partners. Through such guiding and sharing, we expect that students' self-regulated learning at the NMNS is assisted.

Literature Review

Augmented Reality

Augmented reality (AR) is a technology combining virtual environment and real world. Past studies have shown that AR supplements inadequacies of the real world in the way that cumulative learning experience triggers learner's thinking skill and understanding of concepts, instead of replacing the real environment [1]. In essence, the educational value of AR is not merely on the use of technology but how to implement AR to formal and informal learning environments. Furthermore, AR application has been demonstrated actually enhances learning motivation [2]. Therefore, this study adopts AR technology on mobile carrier, making the screen displaying auxiliary information and combines indoor positioning technology to display learning process so as to guide the direction of learning theme.

Indoor Positioning

In the recent years, the development of indoor positioning technology has become increasingly mature. At the current developmental stage, indoor positioning technology can already be used in indoor navigation. The majority of navigation application requires specific learning tasks in the past [3]; learners can only passively accept the task and cannot select appropriate learning resources according to their respective needs. In order to tackle this problem, this study implements indoor positioning technology to detect students' current position to provide learning routes. Indoor positioning technology can identify users' indoor position and provide effective information to aid users to orientate resources in a limited indoor environment. NMNS is one informal learning environment for learners to gain extracurricular knowledge. Learners are impelled to enroll in active learning in NMNS and receive help in learning, thereby allowing learners to learn effectively in informal educational environment [4].

Self-Regulated Learning

Self-regulated learning has been one of the main objectives of formal education set by researchers in the past. If learners possess this ability, they are able to determine current learning needs and reflect on learning performance. Therefore, self-regulated learning activities contribute to the mediation of individual, learning context and actual performance [5, 6]. This study explores in depth how to develop one learning platform in which students can gain information and aids effective learning in informal learning environment such as NMNS, Botanical Garden and so on. With the effective learning platform, the study further combines it with adequate learning strategies during learning process to develop learning mode suitable for learning in NMNS.

Methodology

Bridging Formal and Informal Learning

This study proposes a learning mode to be applied in National Museum of Natural Science by integrating AR and indoor positioning technology to illustrate how students can be diverted from the original formal learning environment into informal learning environment. The study adopts the concept of self-regulated learning proposed by Zimmerman et al. in 1996 [8] and the concept of resource management strategies proposed by Pintrich et al. [9] to construct the "self-regulated learning mode in NMNS." The constructed learning mode is combined with learning activities to promote students to effectively seek for learning resources in informal learning environment and to link the concept and practice.

The design has two levels: self-regulated learning and resource management strategies, as shown in Table 1. "Self-regulated learning" is a structure rendered by informal learning mode; self-regulated learning theory is adopted to explore in depth, including self-assessment and monitoring, goal setting, strategic planning, implementation and monitoring of strategies, and monitoring and correction of strategic results. The process of adjustment learning emphasizes on combining with learning resources to set up appropriate learning objectives, formulation and implementation of study plan, effective time management, monitoring environment, and seek for human resources for discussion on specific implementation of learning activities. In other words, resource management strategy helps in learning strategic design to well and truly carry out learning activities. The purpose of this study is to develop a learning system for cultivating students' self-regulated ability, thus the self-regulated learning mode is used to aid learning. Depending on students' learning conditions, the corresponding indicator at every stage of self-regulated learning is a reference for students to adjust their learning pace. Also, learning resources are integrated to learning tasks at every stage for learning.

Learning Activities in the Informal Learning Environment

This study sets natural science and technology for third-graders as the learning subject. The subject of natural science and technology emphasizes on experiments; the required knowledge is complete only when the learners grasp the concept and can verify with experiments. AR technology plays a role to provide visual multimedia information to aid in learning so that students can fully understand the purpose of the experiment with existing equipment and instruments in the NMNS. At the same time, the gapbetween experimental results demonstrated on instruments and the cognitive results can be reduced. Also, the learning route is effectively planned so students can accurately hold onto learning resources without getting lost during the learning process and hence greatly enhances learning

Tableô1 Self-regulated learning model in NMNS

Stage for learning	Teaching mode of self-regulated learning [10]	Resource management strategies [9]	Implementation of learning activities [11]
Performance prior to learning	Self-assessment and monitoring	Set up appropriate learning objectives Rightful achievement attributes to attitude	*Goal Setting* Carry out action learning in the NMNS in accordance with progress of experimental course
	Goal setting and strategic planning	Formulation and implementation of study plan	*Strategy Learning* Set up steps for learning
Performance in learning	Implementation and monitoring of strategies	Effective time management	*Self-management* Effective management of learning time, and control of personal learning condition
Post-learning performance	Monitoring and correction of strategic results	Monitoring environment and seek for human resources	*Strengthening Mutual Aid* Seek for human resources to solve any learning problems *Cooperative Learning* Solve problems with cooperation to meet expected progress *Self-feedback* Complete learning progress and gain experience *Correction on Implementation* Adjust learning steps appropriately; strengthening ability of self-regulated learning

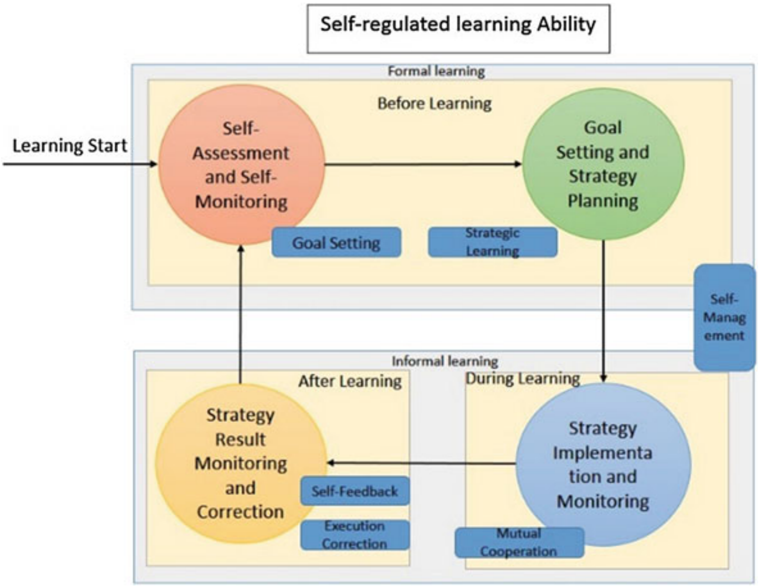


Fig. 1 Self-regulated learning diagram in formal and informal learning contexts

efficiency. This study aims to provide one set of appropriate learning steps (as shown in Fig. 1) designed for learning convergence to guide the students when they encounter the gap between learned knowledge and experimental results when studying natural science and technology subject. For the sake, we provide a solution functioning as bridge between fundamental concept of knowledge and relevant experiments, driving students to have profound experience in learning science and technology as well as to strengthen learning motivation.

Guiding Learning Objectives with AR and Indoor Positioning Technology

As indoor configuration of the NMNS is irregular and with multi-thematic distribution, it is time consuming to look for themes relating to experimental topic corresponding to the curriculum. In order to learn effectively and get the learning resources quickly, this study designs a learning system applicable in the NMNS utilizing the AR and indoor positioning technology.

Augmented Reality

In this study, the use of AR technology assists learners to effectively find the learning resources. The mobile device held by learner will display virtual tag and text prompts to guide the leaner the route to specific learning resources. The problem of time-consuming to find specific learning resource or getting lost along the way during self-regulated learning is thus improved. During the learning process, the system constantly records new learning routes and conditions of usage; such data can be further used in the future to analyze learning effectiveness of the learner.

Indoor Positioning

The usage of indoor positioning technology effectively navigates the learning route and records learning nodes. The technology is implemented to learning routes via learning steps, enabling students to monitor and control their own learning pace and to cultivate self-adjustment learning ability. Students can effectively do self-assessment on their learning performance and the goal of cultivating students' self-adjustment learning ability is therefore achieved.

System Demonstration

This system is designed against the subject of Nature & Life Science for the 3rd-grader. The learning goal, which enables the students to understand the topic learning goal thoroughly, is set up based on the learning requirement of students after ending up with the basic concept program. Moreover, it helps students to find out the learning sources accurately through the assistance of learning platform during the informal learning environment. The virtual arrow and text prompt displayed on the hand-hold device can help the learner not to lose the direction whilst searching for resource indoors and the searching time can also be reduced. Meanwhile, continuous tracking can be executed by using such system and the learning status and usage condition can both be recorded.

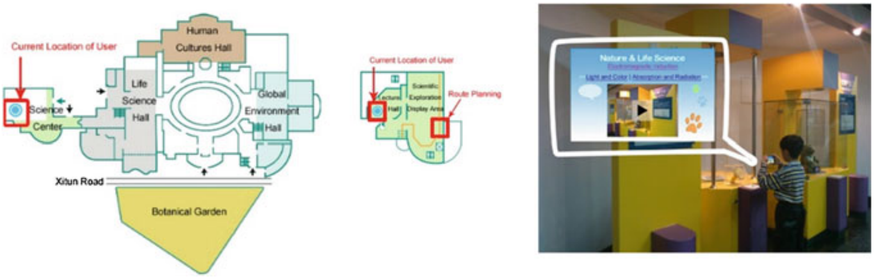


Fig. 2 Simulation screen of interior location / real learning condition of student using such learning platform

The proposed learning system is able to show up the learning path, record the learning nodal point and even present content embedded in learning path effectively by means of the combination of augmented reality technology and indoor positioning technology, as shown in Figs. 2 and 3. In this way, the system not only enriches learning experiences but also draws students' attention and inspire their motivation. During the learning process, self-learning regulation is aroused in order to bring up and regulate the learning habit gradually to achieve the goal of self-regulated learning designed by this study.

Results and Discussion

Augmented Reality is used to Strengthen the Effectiveness of Self-Regulated Learning upon Academic Study

This study will further build up a knowledge learning platform. It will integrate the learning topics in the formal context and learning paths for the learning history of each user through systematic analysis.

When the user is learning in NMNS, the location-based annotation function can be used to hold the learning experience combining with the built-up coordinate system at the current location of learning. Each node can store the learning type of user in NMNS, such as audio visual learning and thinking of some topic. By node learning record, the user can grasp the self-learning progress and share with peers.

Enhance the Learning Efficiency by AR and Indoor Positioning

It is a topic worth concerns if a learner can learn effectively in the environment of informal learning, in which if informal learning is performed in NMNS, the learning status of each learner may not be taken care under the limited human resources. The system are designed to combine the augmented reality and indoor positioning technology to help the learner get the direction whilst searching for resource indoors and reduce searching time. Meanwhile, since the problem of that subject can be provided properly through the system, the learner may not miss the learning point in NMNS. This system can also be used to record the learning history and track the usage condition of system continuously. At last, the learner can not only absorb the knowledge but also manage the self-learning regulation during the learning process and thereupon improve and train the habit of self-learning gradually.

Conclusion and Future Prospect

The purpose to build up this learning system is to combine the environment of formal learning with informal learning. The traditional teaching model of teacher guiding to learn is changed. In a new wave of learning, the students will be guided into a diverse learning environment to enhance the learning experience and knowledge complementary outside class and the self-regulated learning ability in the large-scale environment can be trained. This system provides the assistance of the subject of Nature & Life Science required by the experiment learned by the students, which helps to encourage the students performing expeditionary learning in NMNS, and further strengthen the critical thinking and train the organizational ability of students. The more important point is to be capable of self-regulated learning during the learning process and after it; a set of effective learning model can be obtained. In the future, we will further investigate such system being used in the environment of informal learning and carry out an experiment on the subject of Nature & Life Science learned by the 3rd Grade students of an Elementary School. Meanwhile, we will continuously pay attention to the effect of the self-regulated learning ability developed by the augmented reality and indoor positioning technology.

References

1. Wu HK, Lee SWY, Chang HY, Liang JC (2012) Current status, opportunities and challenges of augmented reality in education. Comput Educ 62, 41–49. doi: 10.1016/j.compedu.2012.10.024
2. DiSerio A, Ibáñez MB, Kloos CD (2012) Impact of an augmented reality system on students' motivation for a visual art course. Comput Educ 1–11. doi:10.1016/j.compedu.2012.03.002
3. Sung YT, Chang KE, Lee YH, Yu WC (2008) Effects of a mobile electronic guidebook on visitors' attention and visiting behaviors. Educ Technol Soc 11(2):67–80
4. Paris SG, Hapgood SE (2002) Children learning with objects in informal learning environments. In: Paris SG (ed) Perspectives on objects-centered learning in museums. Lawrence Erlbaum, Mahwah, NJ, pp 37–54
5. Accurate Mobile Indoor Positioning Industry Alliance, called In-Location, to promote deployment of location-based indoor services and solutions, http://press.nokia.com/2012/08/23/accurate-mobile-indoor-positioning-industry-alliance-called-in-location-to-promote-deployment-of-location-based-indoor-services-and-solutions/ (2012, August 23)
6. Boekaerts M (1997) Self-regulated learning: a new concept embraced by researchers, policy makers, educators, teachers and students. Learning and Instruction. 7(2):161–186. doi: 10.1016/S0959-4752(96)00015-1
7. Pintrich PR (2000) The role of goal orientation in self-regulated learning. In: Boekaert M, Pintrich PR (eds) Handbook of self-regulation Academic Press, San Diego, pp. 13–39
8. Zimmerman BJ (2000) Attaining self-regulation: A social cognitive perspective. In: Boekaerts M, Pintrich PR, Zeidner M (eds) Handbook of self-regulation Academic Press, San Diego, pp. 13–39
9. Pintrich PR, Smith DA, Mckeachie WJ (1989) A manual for the use of the motivated strategies for learning questionnaire (MSLQ). National Center for Research to Improve Postsecondary Teaching and Learning, School of Education, The University Michigan, Mich

10. Zimmerman BJ, Bonner S, Kovach R (1996) Developing self-regulated learners: Beyond achievement to self-efficacy. American Psychological Association, Washington, DC
11. Cherng BL (2001) The Relations Among Motivations, Goal Setting, Action Control, and Learning Strategies: The Construct and Verification of Self-regulated Learning Process Model. J Nat Taiwan Norm Univ 46(1):67–92

Using Particle Swarm Method to Optimize the Proportion of Class Label for Prototype Generation in Nearest Neighbor Classification

Jui-Le Chen, Shih-Pang Tseng and Chu-Sing Yang

Abstract Nearest classification with prototype generation methods would be successful on classification in data mining. In this paper, we modify the encoded form of the individual to combine with the proportion for each class label as the extra attributes in each individual solution, besides the use of the PSO algorithm with the Pittsburgh's encoding method that include the attributes of all of the prototypes and get the perfect accuracy, and then to raise up the rate of prediction accuracy.

Keywords Particle swarm optimization · Prototype generation · Evolutionary algorithms · Classification

Introduction

The nearest neighbor algorithm [1] has a significant effect on classification prediction. To calculate the similarity between the predicted target and the known samples is the way to find the nearest neighbor. This method provides a very high accuracy rate and having the characteristic that the more precise with the more the number of samples. However, there are some drawbacks for the method. The costs of the calculation are too high and the accuracy is susceptible to noise interference.

J.-L. Chen (✉)
Department of Multimedia Design, Tajen University, Tajen, Taiwan, Republic of China
e-mail: reler@mail.tajen.edu.tw

S.-P. Tseng
Department of Computer Science and Information Engineering, National Cheng Kung University, Cheng Kung, Taiwan, Republic of China

J.-L. Chen · C.-S. Yang
The Institute of Computer and Communication Engineering, National Cheng Kung University, Cheng Kung, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_28, © Springer Science+Business Media Dordrecht 2014

In order to solve the above problem, some method can achieve this goal by thinking about reduction of the number of samples. There are two main proposed methods that try to select the reasonable ones from all of the samples and then to perform the nearest neighbor algorithm. These two methods are named prototype selection (PS) [2–7] and the prototype generation (PG) [8–11].

For the purposes of prototype selection to perform the classification, those prototypes are base on the new selection of suitable samples from the training set. There are two main methods for PS problem. One is the concentration method [2], the main idea is to avoid the proportion of certain types of samples are more large than others that make the error decision. For the reason that those samples are eliminated for the properties may be too similar or unrelated. By the second method, the main purpose is to focus on removing those samples would interfere with decision or cause confusion then the follow-up prediction would be more accurate [3]. For prototype generation (PG), not only choose the appropriate samples but also modify the attributes of individual sample. At the result, the decision of classification would be more obvious and distinguished [12, 13].

The main purpose of PG method is to choose or modify the n samples' data from the training set. After that generates a new set, GS, which contains the r prototypes, in which $n > r$. These newly generated prototype can be used for the classification to accelerate the prediction efficiency because of the fewer number of samples would achieve the better accuracy. Find subsets guaranteeing zero errors with N-prototypes for each class when the original data set which is submitted to the Prototype Generation Classifier.

In general, PS and PG problem can be considered as combination and optimization problem, there are many evolutionary search method applied in this problem. PG is regarded as a continuous space search problem. The evolutionary optimization search methods using particle swarm optimization (PSO) and differential evolution (DE) are suitable for continuous spaces. Many schemes are presented on this topic, such as [12–15].

Most of the methods for prototype generation that gives a suggestion for the proportional to classes label is equal to the average, but does not completely arrive at ideal accuracy. The proportion of class for the prototypes in the AMPSO [12] method is uncertain, which is determined by the execution results of each run. In the SFLSDE [13] method, the proportional to the number of all classes for the prototypes is based on the training set. As a result, this method got a good prediction accuracy.

The main contribution of this paper follows a idea to that presented in [13] that is the use of the PSO algorithm with the Pittsburgh's encoding method that include the attributes of all of the prototypes and get the perfect accuracy, but we modify the encoded form of the individual that add the proportion for each class label as the extra attributes in each individual solution and then to raise up the rate of prediction accuracy.

The remainder of the paper is organized as follows. Section "Proposed Method for Prototype Generation" describes in detail the proposed method. Performance evaluation of the proposed method is presented in section “Experimental Framework and Results”. Conclusion is drawn in section “Conclusions”.

Proposed Method for Prototype Generation

In this section, the Particle Swarm Optimization (PSO) method is applied for the prototype generation. The PSO follows the general process of the Evolutionary Algorithm. PSO initializes the population with N members as the candidate solutions (NP), for each solution of NP is named a individual.

Encoding of Prototype

The encoding method will be used the Pittsburgh method, all of the prototype will be encoded into each individual, the size for one individual can be denoted D. The individual with dimension D in DE method can be regarded as a target vector. As the result, $D = r \times n + m$, r is the number of prototypes in the same individual. n is the number of attributes in the prototype. m is the number of type of class label. In addition to those prototypes and it’s attributes as the part of individual solution, the number of distinct class labels is also as a part of solution.

Table 1 describes the structure of an individual. Each prototype p_i has a corresponding class label. Within entire evolutionary cycle of PSO, the value of this class label remains unchanged. It means that by the operation of the PSO, the class labels assigned but still fixed from the initialization phase of each prototype to the end phase, the class labels are not part of the individual. Typically, it is necessary to normalize the values of each attribute before PSO processing. That prevents the attributes in large ranges to influence some attributes in smaller ranges. The normalization means that to transform a value v of a attribute A to v' in the range [−1, 1] by computing with Equation(1) where $\min_A$ and $\max_A$ are the minimum and maximum values of attribute A.

Table 1 Encoding of a set of prototypes

	Prototype 1	Prototype 2	…	Prototype r	Proportion
Attributes	$p_{11}, p_{12},\ldots,p_{1n}$	$p_{21}, p_{22},\ldots,p_{2n}$	…	$p_{r1}, p_{r2},\ldots,p_{rn}$	$d_0, d_1,\ldots,d_m$
Vectors	$x_1, x_2, \ldots, x_n$	$x_{n+1}, x_{n+2}, \ldots, x_{2n}$	…	$x_{(r-1)n+1}, x_{(r-1)n+2}, \ldots, x_{rn}$	$x_{rn+1}, \ldots, x_{rn+m+1}$
Class	d_0	d_1	…	d_m	

r is the number of prototypes
n is the number of attributes
m is the number of class label which denotes as d_0, d_1,…, d_m

$$v' = \left(\frac{v - \min_A}{\max_A - \min_A}\right) \times 2 - 1 \tag{1}$$

Algorithm and Movement

During the initialization process, there is one thing is needed to ensure that each class has at least a prototype to be represented. All of the prototypes are combined and encoded in each individual. These prototypes is proportional to each class with the number of samples in the training set. The individual should include each class with at least one prototype S_i.

Assume an D-dimensional search space S, and a swarm comprising Np particles. The position $X = [x_1, \ldots, x_D] = \left[x_1, x_2, \ldots, x_{rn}, x_{rn+1}, \ldots, x_{rn+(m+1)}\right]$ of a particle in the search space S denotes a candidate solution.

The current position of particle i is an D-dimensional vector $X_i = [x_{i1}, x_{i2}, \ldots, x_{iD}]^t$ belong to S in iteration t.

The velocity of this particle is also an D-dimensional vector $V_i = [v_{i1}, v_{i2}, \ldots, v_{iD}]^t$ belong to S in iteration t, which indicates the displacement for updating the position of each particle in the search space.

The best position encountered by particle i is denoted as $P_i = [p_{i1}, p_{i2}, \ldots, p_{iD}]^t$ belong to S. Assume that g is the index of the particle that attained the best position found by all particles in the neighborhood of particle i.

The swarm is manipulated by the following equations:

$$v_{id}^{t+1} = wv_{id}^{t} + c_1 w_1 \left(p_{id}^{t} - x_{id}^{t}\right) + c_2 w_2 \left(p_{gd}^{t} - x_{gd}^{t}\right) \tag{2}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{3}$$

Where $i = 1,2,\ldots,Np$, is the particles index; $d = 1,2,\ldots,D$, is the dimension index; $t = 1,2,\ldots,T$, is the iteration number; The variable w is a parameter called inertia weight, which balances global and local searches in the PSO. The two positive constants c_1 and c_2 are cognitive and social parameters, respectively. Proper fine tuning of c_1 and c_2 may improve the performance of the PSO. $c_1 = c_2 = 2$ were recommended as default values. The w_1, w_2 generates a random number uniformly distributed within the interval [0, 1].

Finally, if the selected individual obtains the best fitness in the population, and returns the best individual found during the evolutionary process.

Experimental Framework and Results

In this paper, the performance of the proposed algorithm is evaluated by using it to solve the prototype generation in nearest neighbor classification problem. All the experimental results are obtained by running on an IBM X3650 machine with 2.4 GHz Xeon CPU and 16 GB of memory using CentOS 6.0 with Linux 2.6.32. Moreover, all the programs are written in C++ and compiled using GNU C++ compiler.

Parameter Settings and Datasets

We perform experimentation on the problems summarized in Table 2. They are well-known real problems taken from the University of California, Irvine, collection, used for comparison with other classification algorithms.

Table 2 summarizes the properties of the selected data sets. For each data set that include the number of examples, the number of attributes, and the number of classes. For the results of classification, the data sets are using the ten fold cross-validation to perform the prediction.

Experimental Results

In this section, we describe the results of the experiments and perform comparisons between the GA, PSO, AMPSO and proposed method PPGPSO with same population size and iterations. Those parameters for setting. In all experimental results show in this section, we use the following notation: a "(+)" tag to the result means that the average result was significantly better than the result of the other's method. We also use boldface style to highlight the best result (Table 3).

In Table 4, we compare the average success rate of GA, PSO, AMPSO and proposed method PPGPSO. It shows that PPGPSO has more opportunity to do better than others in those problems.

Table 2 Summary description for classification data sets

Data set	#Examples	#Attributes	#Classes
Australian	690	14	2
Breast	286	9	2
German	1000	20	2
Glass	214	9	7
Heart	270	13	2
Iris	150	4	3
Wine	178	13	3

Table 3 Parameter specification for all the methods employed in the experimentation

Algorithm	Parameters
GA	PopulationSize = 50, Iterations = 1,000, reduction rate(r) = 5 %, CR = 0.9, MR = 0.05, CO = TwoPoint
PSO	SwarmSize = 50, Iterations = 1,000, reduction rate(r) = 5 %, C1 = 2, C2 = 2, Vmax = 0.25, Wstart = 1.5, Wend = 0.5

Table 4 Average success rate (in percent) that compared with EA algorithm for prototype generation

Problem	GA	PSO	AMPSO	PPGPSO	Proportion
Australian	64.79	82.43	**87.00** (+)	85.83	(1:1)
Breast	62.31	65.90	**66.16** (+)	65.25	(1:1)
German	70.04	74.54	75.05	**76.49**	(1:1)
Glass	74.43	72.93	82.62	**87.01**	(1:2:1:1:1:3:1)
Heart	95.41	96.28	97.43	**97.54**	(1:1)
Iris	94.31	98.93	**99.99** (+)	99.97	(3:2:1)
Wine	94.43	95.14	96.02	**96.51**	(1:1:1)

Conclusions

In summary, we have proposed differential evolution as a prototype generation for data reduction method. Specifically, it was used to optimize the proportional to the prototypes for the nearest neighbor classification and to perform as a prototype generation method.

The main aim of this paper to modify the encoded form of the individual to plus the proportion for each class label as the extra attributes in each individual solution, besides the use of the DE algorithm with the Pittsburgh's encoding method that include the attributes of all of the prototypes and get the perfect accuracy, and then to raise up the rate of prediction accuracy.

The ongoing work of experimental study would be performed which be allowed us to justify the behavior of DE algorithms when dealing with small and large datasets.

References

1. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
2. Gowda K, Krishna G (1979) The condensed nearest neighbor rule using the concept of mutual nearest neighborhood (corresp.). IEEE Trans Inf Theory 25(4):488–490
3. Wilson DL (1972) Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst, Man and Cybern 3:408–421
4. Brighton H, Mellish C (2002) Advances in instance selection for instance-based learning algorithms. Data Min Knowl Disc 6(2):153–172

5. Marchiori E (2008) Hit miss networks with applications to instance selection. J Mach Learn Res 9:997–1017
6. Fayed HA, Atiya A (2009) A novel template reduction approach for the k-nearest neighbor method. IEEE Trans Neural Networks 20(5):890
7. Marchiori E (2010) Class conditional nearest neighbor for large margin instance selec- tion. IEEE Trans Pattern Anal Mach Intell 32(2):364–370
8. Wilson DR, Martinez TR (2000) Reduction techniques for instance-based learning algorithms. Mach Learn 38(3):257–286
9. Fayed HA, Hashem SR, Atiya AF (2007) Self-generating prototypes for pattern classification. Pattern Recogn 40(5):1498–1509
10. Lam W, Keung C-K, Liu D (2002) Discovering useful concept prototypes for classification based on filtering and abstraction. IEEE Trans Pattern Anal Mach Intell 24(8):1075–1090
11. Bezdek JC, Kuncheva LI (2001) Nearest prototype classi_er designs: an exper-imental study. Int J Intell Syst 16(12):1445–1473
12. Cervantes A, Galván IM, Isasi P (2009) Ampso: a new particle swarm method for nearest neighborhood classification. IEEE Trans Syst, Man, Cybern, Part B: Cybern 39(5):1082–1091
13. Triguero I, Garcia S, Herrera F (2011) Differential evolution for optimizing the positioning of prototypes in nearest neighbor classification. Pattern Recog 44(4):901–916
14. Nanni L, Lumini A (2009) Particle swarm optimization for prototype reduction. Neurocomputing 72(4):1092–1097
15. Triguero I, Garcia S, Herrera F (2010) A preliminary study on the use of differfiential evolution for adjusting the position of examples in nearest neighbor classification. In 2010 IEEE congress on evolutionary computation (CEC). IEEE, pp 1–8

A Novel Genetic Algorithm for Test Sheet Assembling Problem in Learning Cloud

Shih-Pang Tseng, Long-Yeu Chung, Po-Lin Huang, Ming-Chao Chiang and Chu-Sing Yang

Abstract The assessment is the most effectively tool for the teachers to realized the learning status of the learners. The test sheet assembling is an important job in the E-learning. In the future learning cloud environment, the large amount of items would be aggregated into the itembank from various sources. The test sheet assembling algorithm should be with the ability of abstract the needed information directly from the items. This paper proposed an effective method based on genetic algorithm to solve the test sheet assembling problem. The experimental result shows the effectiveness of the proposed method.

Keywords E-learning · Genetic algorithm · Test sheet assembling

S.-P. Tseng (✉) · M.-C. Chiang
Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, Republic of China
e-mail: tsp@mail.tajen.edu.tw

M.-C. Chiang
e-mail: Chungly1@mail.chna.edu.tw

L.-Y. Chung
Department of Applied Informatics and Multimedia, Chia Nan University of Pharmacy and Science, Tainan, Taiwan, Republic of China
e-mail: mcchiang@cse.nsysu.edu.tw

P.-L. Huang
Center of General Education, Kao Yuan University, Kaohsiung, Taiwan, Republic of China
e-mail: tf0155@cc.kyu.edu.tw

C.-S. Yang
Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China
e-mail: csyang@ee.ncku.edu.tw

S.-P. Tseng
Department of Computer Science and Information Engineering, Tajen University, Pingtung, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_29, © Springer Science+Business Media Dordrecht 2014

Introduction

In past 20 years, the development of Internet has changed the modern human living in several domains. The modern learning is also changed by Internet. Learning is a major human activity to accommodate the environment or the society better. Because of the rapid changing of modern society, the learning efficiency becomes more important than the past. The E-learning [1] is proposed and developed in the first decade of the 21st century to enhance the human learning efficiency. In this time, the various kinds of E-learning are widely applied on various types of educations, such as primary education and continuing education.

Cloud computing [2] is the developing trend of information technology. These unprecedented amount of computing and storage resources would be elastically management and organized to support heterogeneous computing needs. The computing applications, such as entertainment and learning, are transformed into services delivered via Internet. The concept of *learning cloud* is proposed to represent the learning service supported by cloud computing. Because of the computing and storage ability, the learning cloud can be used to process and store more learning contents than the traditional learning management system (LMS), and these learning contents can be presented in various different forms and mediums. The learning contents may be from many different sources, such as e-books, wikis and blogs. The integration and organization of learning contents are important issues in learning cloud. Tseng et al. [3] proposed a method to integrate the learning contents from different sources. The large amount of learning content would be prospectively aggregated in the future learning cloud.

Because of the learners with various different knowledge backgrounds and learning experiences, it is almost impossible that the teacher can design a learning plan which can suitable to each student. The individualization [4] of learning tries to provide different learning plans and activities to different students. It is almost impossible in traditional education because of the cost. But it has become an important issue in E-learning domain because the Internet can provide more interactions among the teacher and the students. For the individualization, it is necessary to gather information about the students by assessment process [5]. Computer-based test (CBT), or e-assessment, [6–8] can provide more performance and lower cost than traditional paper–pencil test (PPT) to realize what the students know. The test sheet assembling, which selects the candidate items from the itembank to generate the test sheet, is the basis of E-assessment. The quality of the itembank and the method of assembling are the two main factors which influence the quality of the test sheet. This paper tries to propose a novel method based on genetic algorithm to solve the test sheet assembling problem in learning cloud environment.

The remainder of the paper is organized as follows. The related works are in section Related Works. Section Proposed Method introduced the proposed method. The experimental results are in section Experimental Result. Conclusion is given in section Conclusion.

Related Works

In the last 30 years, the metaheuristics [9] has been successfully applied on various discrete and continuous problems. There are two categories of metaheuristics, single-solution and population based. The single solution based metaheuristics, such as simulating annealing (SA) [10] and tabu search (TS) [11], search the neighbours of the only one solution. The population based metaheuristics, such as genetic algorithm (GA) [12] and particle swarm optimization (PSO) [13], do the parallel searching by a set of solutions and would interchange the information among the solutions.

Genetic algorithm [12] is originally proposed by J. H. Holland. It is inspired by the natural evolutionary process. There are already various variants of GA proposed to different problems, such as travelling salesman problem and numerical function optimization.

Hwang et al. [14] proposed a tabu search method to solve the test sheet assembling problem. Hwangs' work focus on the degree of discrimination and consider the constrains of the expected testing time and the expected ratio of unit concepts. But the weakness of Hwangs' work is that the itembank must be well-organized and can provide all necessary information to the tabu search algorithm for generating the test sheet. It is not usually practical in the learning cloud environment. The items of the itembank may be from several different sources. The organization of the itembank should be usually messy unless a large amount of human work is used to put all items in order. In addition, the information about one item may be incomplete. These situations would restrict the application of Hwangs' work.

Leung et al. [15], propose a personalized genetic algorithm (PGA) for the test sheet assembling problem. The PGA is based on the item response theory and focused on the personalization of assessment. It is the same with the weakness that the itembank must be well-organized.

Proposed Method

In learning cloud environment, the itembank aggregates a large amount of items from different sources. It is not practical to re-organized these items by humans. In one topic, there are many items which contains some redundancy. One item may be identical with another item. Or the item is the variant of another item, the two items do test similar but not the same concepts. Or part of the concepts of one items are in another item. The test sheet assembling in learning cloud environment should reduce the redundancy and maximize the number of concepts in finite items of the test sheet. For this purpose, the Maximum Concepts Genetic Algorithm (MCGA) is proposed in learning cloud environment.

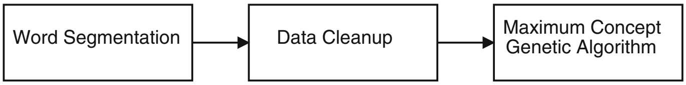


Fig. 1 System architecture of test sheet assembling

Figure 1 shows the system architecture of test sheet assembling. Before the MCGA, there are two stages of pre-processing. At first, all the items of the itembank should be segmented into words. This work focuses on the assessment itembanks on the Taiwan elementary school. Because the characteristic of Chinese, the Chinese item should be segmented into some meaningful words. The Chinese word segmentation is a important research issue in the Chinese natural language processing. We use the Chinese word segmentation service provide by the Chinese Knowledge and Information Processing (CKIP) group [16], Institute of Information Science, Academia Sinica, Taiwan. The accuracy of this Chinese word segmentation is about 96 %. Secondly, the data cleanup stage is responsible for removing all the stopwords in a Chinese item according to the attributes of each word. It is similar to the Stemmers algorithm [17] for English text. But there is no the stop word list which can be used to remove all the useless words. The Chinese word segmentation service provide the attribute of each word. The noun, verb, adjective and adverb is reserved into MCGA, the others is removed.

The solution in MCGAis represented as a bit-string showed as Fig. 2. The length of bit-string, *N,* is the number of total items in the itembank. If the test sheet consists of *n* items; there are only *n* bits of ones in the bit-string, the other bits are zeros.

Algorithm 1 Maximum concepts genetic algorithm

```
Randomly initiate the population
while The terminate condition is not met do
    Selection()
    Crossover()
    Mutation()
end while
Output the result
```

Algorithm 1 shows the outline of MCGA. At first, all the solutions in the population are initiated randomly. The main loop would be terminated after the pre-defined number of iterations. In the main loop, like the general genetic algorithm, there are three steps: *selection*, *crossover* and *mutation*. The tournament selection [18] is used as the selection operator in this work and the tournament size is set to 3. The selected solutions would be store a temporary population as the parents of crossover.

The simple 2-points crossover is used to reproduced the next generation solutions. Figure 3 shows an example which is used to illustrate the 2-points crossover in MCGA. There are total 10 items in the itembank of the example, and the

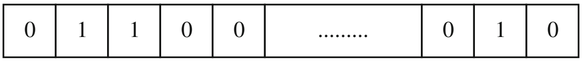


Fig. 2 The representation of solutions

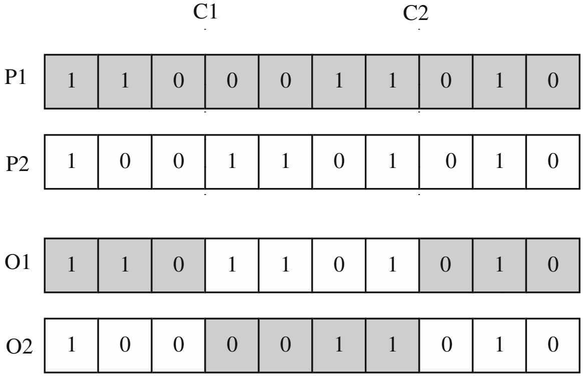


Fig. 3 The example of crossover in MCGA

testsheet consists of 5 items. The parents, *P*1 and *P*2, are legal solution, but the ofspring, *O*1 and *O*2, are both illegal. There are 6 ones in *O*1, and only 4 ones in *O*2. It is necessary to repair the offspring solutions.

In the binary bit-string representation of solutions, the simplest mutation is flipping one arbitrary bit. Because the repairing may be needed after the crossover step, we designed two kinds of mutation, *increment mutation* and *decrement mutation*. Figure 4 shows the example of mutations. If the number of the solution's one bits is greater than *n*, the increment mutation would be applied. On the other hand, If the number of the solution's one bits is less than *n*, the decrement mutation would be applied. The increment mutation would choose the zero-bits to flip until the solution become a legal solution. In addition, the decrement mutation would choose the one-bits to flip until the solution become a legal solution. In Fig. 4, the solution *O*1 is applied the decrement mutation, the 2nd bit is changed from one to zero. And the solution *O*2 is applied the increment mutation, the 4th bit is chosen to flip from zero to one. There are two strategies used to choose the flip bits in the

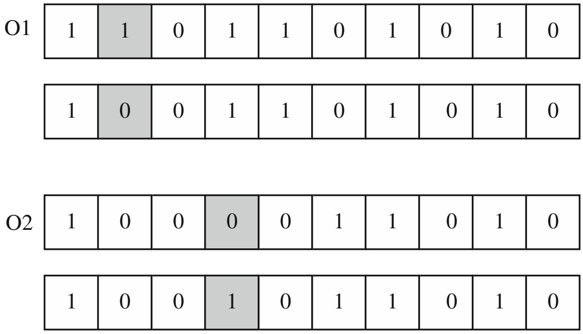


Fig. 4 The representation of solutions

both increment and decrement mutations. The first flip-bit choosing strategy is *random choosing*. This is simple and easy to implement. The second flip-bit choosing strategy is *heuristic choosing*. The heuristic choosing strategy on the increment mutation would prefer the item which can increase the maximum fitness value. The heuristic choosing strategy on the decrement mutation would prefer the item which would decrease the minimum fitness value.

Experimental Result

For illustrating the effectiveness of MCGA, we have implemented the MCGA on a HP DL165 G7 machine with 2.6 GHz AMD Opteron CPU and 12 GB of memory using Ubuntu 12.04. Moreover, all the programs are written in C++ and compiled using g++ (GNU C++ compiler). The two variants of MCGA are implemented. The first one is the MCGA with *random choosing* mutation, denoted by $\mathrm{MCGA_R}$. The other one is the MCGA with *heuristic choosing* mutation, denoted by $\mathrm{MCGA_H}$. In this paper, the population size is set to 40, and the crossover rate is 0.8. The number of iterations is equal to 100. The initial solutions are generated by randomizing. Each experiment repeated 30 trials and all results shown in this paper are the average of 30 trials.

As shown in Table 1, ten itembanks—denoted DS1-10—are used to measure the performance of the MCGA. The DS1 itembank is from the textbook published by the Han Lin Publishing [19]. And the DS2 itembank is from the textbook published by the Kang Hsuan Publishing [20]. These two books are for the grade 4 Society course of the elementary school in Taiwan. These two books are edited according to the same standard of Society textbook; so the contents of these two books are eventually identical, but with the different schemas. The DS1 itembank has 6 chapters, 13 sections, and 697 items. The DS2 itembank has 3 chapters, 15 sections, and 464 items. The DS3-DS6 itembanks are the subsets of DS1. The DS3 contains the first 3 chapters of DS1, DS4 contains only the 2nd chapter. The items of the last 3 chapters are in DS5, and the DS6 contains only the 5th chapter. The DS7-10 are based on DS3-6 and integrated the items from DS2 by using the method [3]. All of the itembanks are segmented by using the service of the Chinese word segmentation system [16].

In this work, *coverage* are chosen as the fitness value and the performance measure. The coverage evaluation function could be described as Eq. (1). We assume that the concepts in one item can be represented by its keywords. And all identical keywords of the itembank can be as the domain of the itembank. The coverage means that the ratio of the itembank's domain is covered by the test sheet.

$$MSE = \frac{\text{The identical keywords in the test sheet}}{\text{The identical keywords in the itembank}} \tag{1}$$

Table 1 Itemsets

	Item#	Keyword#	Identical keywords
DS1	697	7,268	1,874
DS2	464	3,908	1,255
DS3	340	3,395	1,058
DS4	123	1,230	443
DS5	357	3,873	1,144
DS6	157	1,707	611
DS7	551	5,137	1,481
DS8	182	1,702	616
DS9	516	5,248	1,525
DS10	254	2,555	887

Table 2 Experimental result of MCGA, the test sheet size is 50

	Random	MCGAR		MCGAH	
	Coverage	Coverage	Time	Coverage	Time
DS1	0.19	0.36	0.78	0.39	5.81
DS2	0.23	0.37	0.50	0.39	1.53
DS3	0.30	0.54	0.42	0.57	1.16
DS4	0.58	0.86	0.16	0.88	0.24
DS5	0.31	0.51	0.46	0.54	1.54
DS6	0.50	0.76	0.23	0.78	0.38
DS7	0.22	0.39	0.62	0.42	3.77
DS8	0.44	0.70	0.23	0.72	0.36
DS9	0.23	0.38	0.63	0.41	3.30
DS10	0.34	0.56	0.34	0.58	0.86

Table 3 Experimental result of MCGA, the test sheet size is 100

	Random	MCGAR		MCGAH	
	Coverage	Coverage	Time	Coverage	Time
DS1	0.33	0.54	1.56	0.59	10.02
DS2	0.38	0.58	0.97	0.61	3.23
DS3	0.50	0.76	0.80	0.80	1.83
DS4	0.89	1.00	0.30	1.00	0.45
DS5	0.50	0.74	0.87	0.78	2.19
DS6	0.77	0.96	0.42	0.97	0.67
DS7	0.36	0.58	1.20	0.63	5.46
DS8	0.70	0.93	0.43	0.94	0.46
DS9	0.37	0.58	1.24	0.62	5.48
DS10	0.57	0.80	0.65	0.82	1.43

The size of test sheet is set to 50 and 100. The Tables 2 and 3 show the comparison of *random*, MCGA_R and MCGA_H. Both MCGA_R and MCGA_H are dramatically better than the random method in all itembanks. In addition, the coverage of the $\mathrm{MCGA_H}$ is better than the $\mathrm{MCGA_R}$. But $\mathrm{MCGA_H}$ is slower than the $\mathrm{MCGA_R}$.

Conclusion

The test sheet assembling is an important job in the E-learning. In the future learning cloud environment, the large amount of items would be aggregated into the itembank from various sources. In this situation, the well-organized itembank is not practical. The test sheet assembling algorithm should be with the ability of abstract the needed information directly from the items. The proposed MCGAis based on genetic algorithm and incorporated with the domain-specific heuristic. The experimental result shows the MCGA can effectively this problem. In the future, we try to use the multi-objective optimization to assemble the test sheet for more different test requirements.

Acknowledgments The authors would also like to thank Kang Hsuan Publishing and Han Lin Publishing for providing their itembanks and the Chinese Knowledge and Information Processing (CKIP) group, Institute of Information Science, Academia Sinica for providing their Chinese Word Segmentation System to support this research.

References

1. Zhang D, Zhao JL, Zhou L, Nunamaker JF Jr (2004) Can e-learning replace classroom learning? Commun ACM 47(5):75–79
2. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zaharia M (2010) A view of cloud computing. Commun ACM 53(4):50–58
3. Tseng SP, Chiang MC, Yang CS, Tsai CW (2010) An efficient algorithm for integrating heterogeneous itembanks. Int J Innovative Comput, Inf Control 6(10):4319–4334
4. Bork A, Gunnarsdottir S (2001) Individualization and interaction. Tutorial Distance Learning, vol 12 of Innovations in Science Education and Technology. Springer, Netherlands, pp 47–62
5. Pellegrino JW, Chudowsky N, Glaser R (2001) Knowing what students know: the science and design of educational assessment. The National Academies Press, USA
6. Chua YP (2012) Effects of computer-based testing on test performance and testing motivation. Comput Hum Behav 28(5):1580–1586
7. Llamas-Nistal M, Fernndez-Iglesias MJ, Gonzlez-Tato J, Mikic-Fonte FA (2013) Blended e-assessment: migrating classical exams to the digital world. Comput Educ 62:72–87
8. JISC (2007) Effective practice with e-assessment: an overview of technologies, policies and practice in further and higher education
9. Blum C, Roli A (2003) Metaheuristics in combinatorial optimization: overview and conceptual comparison. ACM Comput Surv 35(3):268–308
10. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680
11. Glover F, Laguna M (1997) Tabu search. Kluwer Academic Publishers, Heidelberg

12. Holland JH (1992) Adaptation in Natural and Artificial Systems. MIT Press, Boston
13. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proceedings of the IEEE international conference on neural networks. pp 1942–1948
14. Hwang GJ, Yin PY, Yeh SH (2006) A tabu search approach to generating test sheets for multiple assessment criteria. Education, IEEE Transactions on 49(1):88–97
15. Gu P, Niu Z, Chen X, Chen W (2011) A personalized genetic algorithm approach for test sheet assembling. In Leung H, Popescu E, Cao Y, Lau R, Nejdl W (eds) Advances in web-based learning—ICWL 2011. Vol 7048 of Lecture notes in computer science, Springer, Berlin, pp 164–173
16. Chinese Document Segmentation (2008). http://ckipsvr.iis.sinica.edu.tw/
17. Porter MF (1997) Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco, pp 313–316
18. Miller BL, Miller BL, Goldberg DE, Goldberg DE (1995) Genetic algorithms, tournament selection, and the effects of noise. Complex Systems 9:193–212
19. Han Lin Publishing (2002). http://www.hle.com.tw/
20. Kang Hsuan Publishing (2002). http://www.knsh.com.tw/

Part V
Modern Learning Technologies and Applications with Smart Mobile Devices

Investigation of Google+ with Mobile Device Implement into Public Health Nursing Practice Course

Ting-Ting Wu and Shu-Hsien Huang

Abstract In recent years, mobile device assisted clinical practice learning is popular for the nursing school students. The introduction of mobile devices not only saves manpower and avoids the errors, but also enhances nursing school students' professional knowledge and skills. In order to respond to the demands of different learning strategies and reinforce the maintenance of learning system, new Cloud Learning is gradually introduced to instructional environment. This study introduces mobile devices and Cloud Learning in public health nursing practice with their advantages and adopts Google+ integrating different application tools as the learning platform. The users can save and use the data by all devices with wireless internet. According to findings of this study, learning effectiveness of the learners adopting Google+ is higher than that in traditional learning. Most of the students and nursing educator have positive attitude and are satisfied with the innovative learning method.

Keywords Mobile device · Public health nursing practice · Google+

Introduction

Rise of internet and information technology has changed the learning style and learning becomes more diverse [1, 2]. In highly flexible medical and nursing environment, the immediateness and continuity of mobile devices allow the users

T.-T. Wu (✉)
Department of Information Management, Chia-Nan University of Pharmacy and Science,
Tainan, Taiwan, Republic of China
e-mail: wutt0331@mail.chna.edu.tw

S.-H. Huang
Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan,
Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing,* Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_30, © Springer Science+Business Media Dordrecht 2014

to immediately acquire the resources needed and solve the problems [3]. Besides general hospitals, nursing related courses gradually introduce mobile devices in nursing school students' practices, such as clinical practice [4], tracking medication administration [5], outpatient clinics [6] and nursing homes [7], etc. Introduction of mobile devices in nursing practice saves the manpower, avoids the errors and provides rapid and immediate information searching [8]. In this study treats mobile devices as main learning tools and introduces them in public health care nursing practice. By the immediate support and service of mobile devices, the learners can acquire critical learning guidance and content at the right time and construct theoretical and practical connection in learning process to achieve the purposes of nursing practice.

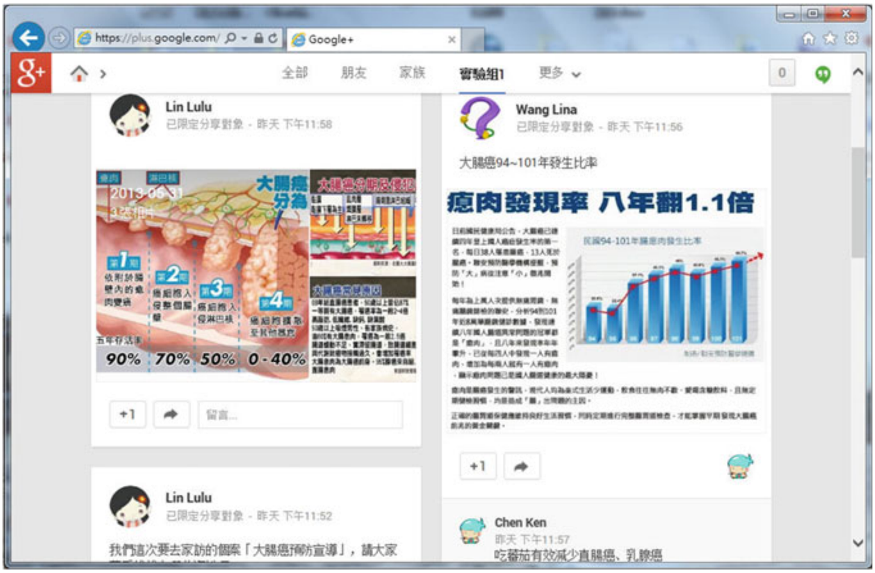
Besides, in order to enhance the maintenance of learning system, reduce the problems of program upgrading and increase the reuse of instructional materials [9], Cloud Computing is used to design and plan the system. On Cloud platform, according to the users' different needs, the system can dynamically and immediately arrange and reorganize the support and functions to provide substantial or virtual services [10, 11]. Cloud Learning resembles a virtual office or campus. The learners and the nursing educator can acquire the resources, instructional materials, programs and services needed freely or by low cost. The function of high calculation and saving capacities allows the users to save all learning data and instructional materials on Cloud. By internet, the users can access and operate the data through internet. High flow of Cloud can enhance the reuse of instructional materials and avoid the incompatibility between data or programs [12]. In the highly flexible and convenient Cloud Learning environment, the learners having different types of mobile devices can acquire the learning content and instructional materials needed through internet at any time and in any place. Thus, this study suggests constructing learning management system on Cloud and the users can upload and access learning content and materials by mobile devices and internet. With the attributes and characteristics of Cloud Learning, instructional strategy and application of collaborative learning can be accomplished. Google+ is the learning management system of this study and it is currently integrated with the programs related to Google. With Google+, the learners can experience the effectiveness of Cloud Learning by varied application tools.

The purpose of this study is to help nursing learners connect the theories and practices by the introduction of new information and transform the abstract concept into specific operation, enhance the learners' knowledge structure and increase the learners' application of public health care theories. In addition, with the assistance of technology, the learners' obstacles in home visit will be reduced to enhance the interest in public health care practice. Besides, learning system can record all learning portfolio and the data can be the criteria for the teachers' instruction after class and adjustment of learning content.

Cloud Learning System for Public Health Practice

This study treats Google+ as the learning management system of Cloud environment (Fig. 1). Google+ integrates the application tools related to Google (Google Picasa、Google Docs、Google Map、Google Location、Hangouts…etc.). The learners can apply for the accounts of Google to use varied services of tools through internet. They can rapidly post, access and use the information, avoid the incompatibility of programs and increase the flow and use of information. Through the setting of high degree of privacy on Google+, the learners can classify different types of social circles and they can construct the public groups in information sharing. It increases the authority of personal privacy. In addition, besides the basic operation of words, pictures and videos, Google+ can position the locations of the data uploaded and offer the map. Besides, Google+ also functions as multi-user word chat room and video conference. In any places, the users can exchange the information. All behaviors on Google+ will be saved and recorded on Cloud and the nursing educator can download all activities of the learners, analyze and recognize the learners' learning situation and provide proper feedback and assistance.

Introduction of Cloud Learning allows nursing practice students to access and browse Google+ by wireless computers and mobile devices. The learners can use the functions on the platform for the practice of home visit. Use interface of the



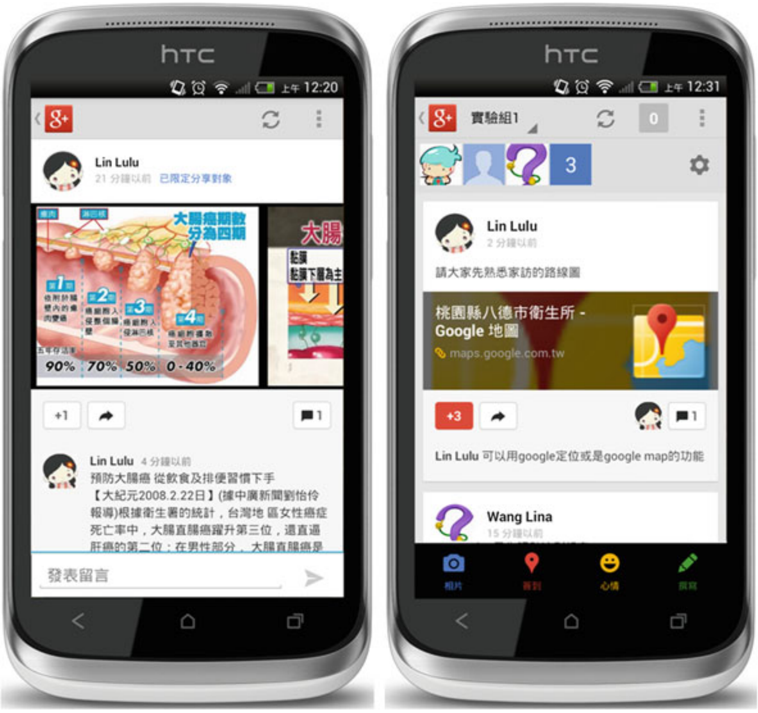
Lin Lulu
已限定分享對象 - 昨天 下午11:58

Wang Lina
已限定分享對象 - 昨天 下午11:56
大腸癌94~101年發生比率

Chen Ken
昨天 下午11:57
吃蕃茄有效減少直腸癌、乳腺癌

Lin Lulu
已限定分享對象 - 昨天 下午11:52
我們這次要去家訪的個案「大腸癌預防宣導」，請大家

Fig. 1 Interface of cloud learning system (Google+)



Lin Lulu
21 分鐘以前 已限定分享對象

Lin Lulu 4 分鐘以前
預防大腸癌 從飲食及排便習慣下手
【大紀元2008.2.22日】(據中廣新聞劉怡伶報導)根據衛生署的統計，台灣地 區女性癌症死亡率中，大腸直腸癌躍升第三位，還直逼肝癌的第二位；在男性部分，大腸直腸癌是

實驗組1

Lin Lulu
2 分鐘以前
請大家先熟悉家訪的路線圖

桃園縣八德市衛生所 - Google 地圖
maps.google.com.tw

Lin Lulu 可以用google定位或是google map的功能

Wang Lina
15 分鐘以前

Fig. 2 User’s interface showing on mobile device

learners is shown in Fig. 2. Before home visit, the learners of different groups can plan the visit routes by Google Map and post the planned routes and maps on Google+. In the process of home visit, they can access learning platform by mobile devices and read the planned routes of home visit. The learners can even search for the locations of the cases with the guidance of Google Map to avoid the confusion with the routes. In addition, Google+ learning platform is the supporter in the practice. Through the platform, the nursing educator can control the home visit practice of the students in different groups at the same time and enhance the interactive learning between the nursing educator and the students. The nursing educator can actively care about and guide the passive learners and passively assist with active learners. Thus, in the home visit, the guidance information will support the practice. The learners will realize the contribution of basic medical care institutions to overall medical care system and it will effectively fulfill the application of knowledge related to health education.

Research Design

Participants

Two classes participated in this study and they were the grade 4 students in five-year nursing college. The students have taken the course of public health care. There were totally 66 subjects and they voluntarily participated in the experiment. The researcher randomly divided the subjects into experimental group (34 students) and control group (32 students). In the practice of home visit, experimental group students treated Google as learning management system and used smart mobile phones as the mobile devices of home visit. Through Cloud Learning, the learners could operate, use and save the functions and information on learning platform and rapidly share and exchange the information. Control group students received traditional instruction and in home visit, they recorded the process and exchanged with each other by paper. The researcher tried to find the difference of learning effectiveness of the students in two groups after the introduction of Cloud Learning.

The nursing educator of two classes was the same person who had rich instructional experience and was considerably interested in the course of nursing practice upon technology. Thus, the said nursing educator voluntarily participated in the experiment. In addition, in order to make sure that each subject had positive execution ability and learning process in experiment, the scores would be included in the subjects' academic performance in this semester.

Experiment Procedure

Public health care nursing practice of the experiment lasted for 4 weeks and the course took 7 h every day. One day was designed for the practice of home visit every week. When the students did not go out for home visit, the nursing educator would introduce basic public health care in health office to allow the learners to approach the functions and jobs in basic medical health care institutions and recognize the roles and contribution of public health care nurses. Before the practice of home visit, the nursing educator introduced the activity, divided the students into groups (two students in a group) and distributed the areas of home visit; in addition, for experimental group students, the nursing educator would arrange one day for explaining the learning system and allow the students to use it. In the process, the technical personnel would help the learners be familiar with the interface. Besides making the nursing educator and the learners to approach the system, this study also distributed the scale of computer use to the learners to recognize the learners' past experience and cognition of information use.

In the practice of home visit, experimental group learners could share, save and discuss the related information on Google+ by any devices. In home visit, through smart mobile phones, the learners could immediately check the information, position the locations, upload data, pose the questions and discuss the related information. With the platform, the nursing educator could control the practice of the students in different groups and offer feedback and assistance at the right time. Control group students received traditional instruction and they recorded and planned pre-planning, practice and even discussion after class by paper. The nursing educator controlled the students' practice by phone. Overall experiment process is shown in Fig. 3.

Result and Discussion

As to evaluation of learning effectiveness, *t* test is used to analyze the scores of experimental group and control group, and descriptive statistics is used to explain the effect of immediate and convenient Cloud Learning in public health care nursing practice on learning effectiveness.

Statistical result of scores of experimental group and control group is shown in Table 1. Experimental group and control group are treated as independent variables and the scores of practice are dependent variables for t-test of independent

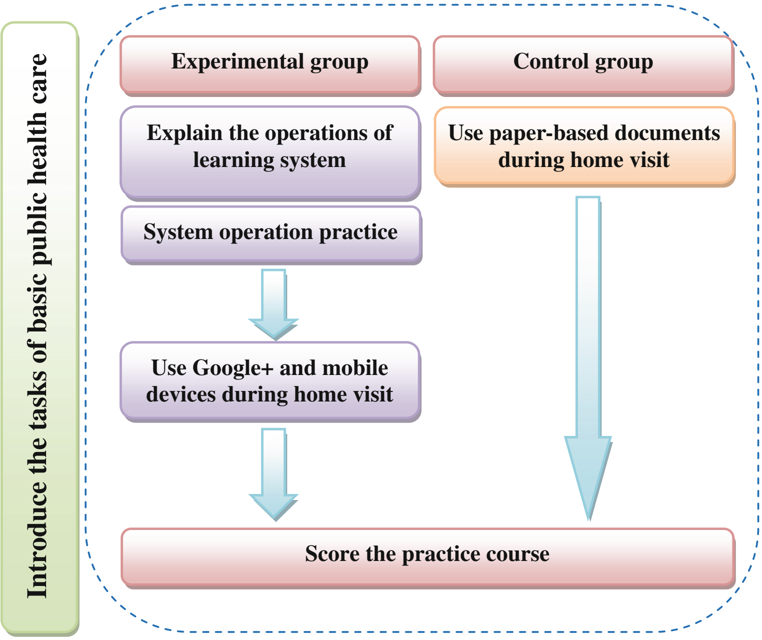

Introduce the tasks of basic public health care
Experimental group
Control group
Explain the operations of learning system
Use paper-based documents during home visit
System operation practice
Use Google+ and mobile devices during home visit
Score the practice course

Fig. 3 The flow of experiment process

Table 1 t-test result of the learning effectiveness

	Mean	N	Std. Deviation	*t*	df	*p*
Experimental	85.63	34	2.34	4.26	64	.02*
Control	79.84	32	3.17			

*$p < 0.05$

samples in order to analyze the difference of learning effectiveness of experimental group learners with Google+ from control group with traditional learning.

Before judging the statistical result, it is necessary to test the homogeneity of two groups of variables. According to analytical result, the value is more than 0.05. It means that homogeneity of two groups of variables is supported. Besides, according to statistical result, *p* value is below 0.05. It means that the practice scores of the learners in two groups are significantly different. According to mean of the practice scores of two groups, experimental group learners' learning effectiveness of home visit with Google+ and Cloud Learning is significantly higher than that with traditional paper instructional strategy. In addition, standard deviation of experimental group is significantly lower than control group. Therefore, Cloud and information system can effectively reduce the learners' gap. The statistical result shows that on Google+ as the platform of Cloud Learning, the learners with mobile devices connect with the practice content and situations at any time. It can effectively reduce the abstract concept and increase practical experience. The varied functions on Cloud allow the learners to effectively internalize the learning content and develop comprehension, operation and integration abilities of home visit.

Conclusion and Future Work

This study tries to introduce Google+ and Cloud Learning in nursing practice course; thus, it only probes into learning effectiveness The application tools related to Google+ are diverse and rich. Future studies can analyze different practice courses and probe into the tools suitable for different practices. In addition, community function of Google+ can be analyzed by collaborative learning or explored by different learning methods in the courses, such as PBL, Concept Map, Jigsaw, etc. In addition, future studies can explore the learners' perceived loading in learning process. This study is still ongoing and the researcher expects that the findings can serve as references for related researches and provide more complete and convenient learning environment and strategy for nursing learners and nursing educator.

Acknowledgments This work was supported in part by the National Science Council (NSC), Taiwan, and ROC, under Grant NSC 102-2511-S-041-002, NSC 101-3113-P-006 023 and NSC 100-2511-S-006-015-MY 3.

References

1. Welsh ET, Wanberg CR, Brown KG, Simmering MJ (2003) E-learning: emerging uses, empirical results and future directions. Int J Training Dev 7(4):245–258
2. Eklund J, Kay M, Lynch HM (2003) E-learning: emerging issues and key trends: a discussion paper. Australian National Training Authority
3. Hwang GJ, Wu TT, Chen YJ (2007) Ubiquitous computing technologies in education. J Distance Educ Technol 5(4):1–4
4. Wu PH, Hwang GJ, Tsai CC, Chen YC, Huang YM (2011) A pilot study on conducting mobile learning activities for clinical nursing courses based on the repertory grid approach. Nurse Educ Today 31:8–15
5. Brian J, Jamieson S (2002) Post-surgical cardiac patients receive new level of care. Caring 21 (3):28–29
6. Edwards DJ (2001) Help is at hand (held): meet the new personal digital assistants that will keep you informed. Nurs Homes Long Term Care Manage 50(4):52–54
7. Hassett M (2002) PDAs: leading the information revolution. Nursing 32(2):66
8. Miller J, Shaw-Kokot JR, Arnold MS, Boggin T, Crowell KE, Allegri F et al (2005) A study of personal digital assistants to enhance undergraduate clinical nursing education. J Nurs Educ 44:19–26
9. Wang Y, Hong A (2010) The construction of virtual learning community of teachers in the times of cloud learning. Chin Educ Technol 1:118–122
10. Raman T (2008) Cloud computing and equal access for all. W4A2008 Keynote, Beijing, China, pp 21–22
11. Sedayao J (2008) Implementing and operating an internet scale distributed application using service oriented architecture principles and cloud computing infrastructure. iiWAS2008, Austria, pp 417–421
12. Al-Zoube M (2009) E-Learning on the cloud. Int Arab J e-Technol 1(2):58

Cognitive Diffusion Model with User-Oriented Context-to-Text Recognition for Learning to Promote High Level Cognitive Processes

Wu-Yuin Hwang, Rustam Shadiev and Yueh-Min Huang

Abstract This study proposed Cognitive Diffusion Model to investigate the diffusion and transition of students' cognitive processes in different learning periods (i.e. pre-schooling, after-schooling, crossing the chasm, and high cognitive processes). In order to enable majority of students crossing the chasm, i.e. bridge lower and higher levels of cognitive processes such as from understanding the knowledge that students learn in class to applying it to solve daily-life problems, this study proposes User-Oriented Context-to-Text Recognition for Learning (U-CTRL). Students participating at learning activities can capture learning objects and then recognize them into text by using U-CTRL. Finally, this study presents a case that shows how to facilitate students' cognition in English through applying the knowledge to solve daily-life problems with U-CTRL and how to evaluate the case.

Keywords Cognitive diffusion model · User-oriented context-to-text recognition for learning · Cognitive processes · EFL learning

W.-Y. Hwang
Graduate Institute of Network Learning Technology, National Central University, No. 300, Jhongda Road, Jhongli 32001, Taiwan, Republic of China
e-mail: wyhwang@cc.ncu.edu.tw

R. Shadiev (✉) · Y.-M. Huang
Department of Engineering Science, National Cheng Kung University, No. 1, University Road, Tainan 70101, Taiwan, Republic of China
e-mail: rustamsh@gmail.com

Y.-M. Huang
e-mail: huang@mail.ncku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_31, © Springer Science+Business Media Dordrecht 2014

Introduction

After learning at school, most students usually remember and understand knowledge taught by the teacher and just few of them can apply it to solve daily-life problems. Remember and understand are low level and apply, analyze, evaluate, and create are high-level cognitive processes [1]. Apply level of cognitive processes plays an important role as it separates cognitive domain into high and low level processes. After students remember and understand the knowledge, a very important goal for an instruction becomes enabling students to apply that knowledge to solve daily-life problems. Furthermore, cognitive processes, such as Apply, need to be cultivated in students as they promote other cognitive processes of higher level.

Accordingly, it is important that students learn the knowledge at school as well as apply it outside of school, therefore, learning is not confined to the classroom anymore but takes place in a range of situations. Students will be able to apply the knowledge to solve daily-life problems in context of school district environment (i.e. outside of school or at home) after class time, through exploration and verification of knowledge. In this way, students will learn really useful knowledge and utilize it in different daily-life situations, e.g. paper-based PISA assessment [2].

Cognitive Diffusion Model

Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system [3]. There are a total of five categories of the members who adopt an innovation: innovators, early adopters, early majority, late majority, and laggards. To achieve popularity and acceptance of a technology, first, it needs to be utilized by 16 % of members, such as innovators and early adapters. Following early adopters are early majority (about 34 %) who usually want to be sure a technology works and is useful before adopting it. According to Moore [4], there is a chasm between the early adopters and the early majority as they have very different expectations. Crossing the chasm is a very difficult task that any innovation or innovative company must successfully accomplish to reach wide market success.

In order to promote students' cognitive processes from low to higher level, this study proposed the cognitive diffusion model. In the proposed model (Fig. 1), students' cognitive processes distributed into six levels [1] based on data reported by Kocakaya and Gönen [5]. The first (and the highest) level of the model is Create and the last (and the lowest) level is Remember or do not remember.

Crossing the chasm concept [4] was adopted in Cognitive diffusion model. It locates between Apply and Understand levels (Fig. 1); that is crossing the chasm imply that students' cognitive processes reach higher level from low level. It is very important for educators to find a way to cross the chasm. However, it cannot

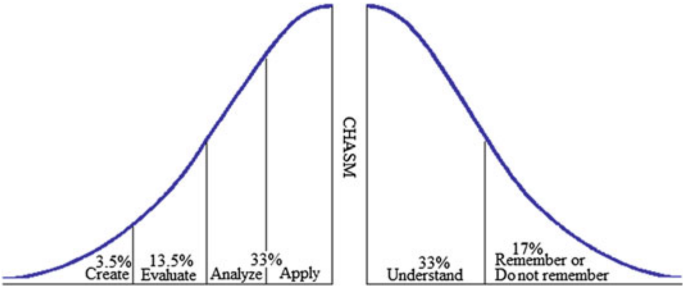


Fig. 1 Cognitive diffusion model chasm

be achieved through paper and pencil tests or exercises. Thus, some effort from educators is required to assist students to reach at least Apply level. The figure shows that the half of students (i.e. 50 %) has crossed the chasm. Cognitive processes of 3.5 % of these students are on Create level, of 13.5 % on Evaluate level, and of 33 % on Apply and Analyze levels. Cognitive processes of the other 33 % of students are on Understand level and of the left 17 % of students are on Remember or do not remember level. However, the percentage of distribution can be different and skewed for different domain knowledge.

Furthermore, this study explores the distribution of students' cognitive processes in cognitive diffusion model according to four learning periods (i.e. pre-schooling, after schooling, crossing the chasm, and high cognitive processes).

In the first period (pre-schooling), most students do or do not remember certain knowledge and only a small number of students can understand it. Therefore, cognitive processes of students in this period are only on the lowest level. In the second period (after schooling) students were instructed about the knowledge and they carried out some related exercises and assignments. In this period, level of cognitive processes of students increased so that most students not only remember the knowledge but also understand it. With further practice, few students even can apply knowledge to solve daily-life problems. The third period (crossing the chasm) is a critical period as it enables most students (at least 50 %) cross from the low level to the high level of cognitive processes (i.e. at least apply). The fourth period is high cognitive processes and cognitive processes of most students (70–80 %) reach highest level, i.e. equal or higher then apply.

In order to better understand why there are four learning stages, an example is given related to learning English as a foreign language (EFL). Most students know English words, how to spell them and their phonetic, but only a small part of them can apply these words to solve daily-life problems, e.g. communicate with someone else. Therefore, even after school learning, most students are still in the second stage (after school learning) and level of their cognitive processes cannot be high. On a contrary, most students have knowledge of native language and they can easily use it for daily conversation. Thus, we may conclude that instruction of native language in elementary school can enable crossing the chasm so that most students reach high level of cognitive processes.

Why there is such a big difference between learning native language and English? In fact, current educational system puts too much emphasis on concept learning and acquisition of knowledge. A little attention is paid on application of knowledge to solve daily-life problems. In most classes, students are requested to do assignments or answer test questions to test their knowledge. No focus is made to enable students to apply knowledge of English to solve daily-life problems. As for native language, obviously students got used to apply it in daily life conversation.

User-Oriented Context-to-Text Recognition for Learning

This study proposes User-Oriented Context-to-Text Recognition for Learning (U-CTRL) mechanism. Students capture objects in the real-life situation then technology converts information into learning text. To enhance cognition students this study suggests extending boundaries of learning environment to outside of classroom so that learning activities are no longer confined to the classroom teaching, instead, students learn anytime and anywhere. U-CTRL provides students with the context of real-life learning situations; that is, students can construct meaningful knowledge and apply it to solve daily-life problems. Students actively capture various learning objects they are interested in (active learning) by using U-CTRL (e.g. tablet computers) and accumulate information in cloud computing database. Therefore, this approach promotes students motivation and interest, and makes learning context richer. Portability of tablet computers will help students to carry out various learning activities outdoor. Information can be inputted through screen keyboard, voice or handwriting, digital camera and etc. in order to carry out individual and collaborative learning. Global positioning system (GPS) will obtain and display students' location in school district. Finally, students will share learning information with peers and it will enhance students' capabilities to search for information, process, and analyzing it. Moreover, it will demonstrate students' ability to apply knowledge to solve daily-life problems. U-CTRL approach is superior comparing to use of RFID or QR code technologies, as the instructor needs to set up learning environment for the latter in advance (e.g. to locate barcode labels or tags). Furthermore, U-CTRL is more context-aware richer comparing to information provided by experts to students.

How to assist students to cross the chasm with U-CTRL? This study proposes four phases (Fig. 2), each of them should have incentive that encourages students to become familiar with U-CTRL and then use it for learning. Phase 1: Training potential students (around 3.5 %) about how to apply knowledge to solve daily-life problems by using U-CTRL. Phase 2: Students (around 3.5 %) with high level (at least apply) tutor students (13.5 %) with lower level of cognitive processes (at least Understand) about how to apply knowledge to solve daily-life problems by using U-CTRL; approximate proportion will be 1 student with higher level to 4 students with lower level. Phase 3: Students (17 %) with at least Apply level tutor students with at least Understand level (33 %) about how to apply knowledge to solve daily-life problems by using U-CTRL; approximate proportion will be 1 student with at least Apply level to 2 students with Understand level. After completion of the three phases, proportion of students to cross the chasm will reach 50 %. That is, it can be said that after schooling reached crossing the chasm, as shown in Fig. 3. Phase four: Students (50 %) with at least Apply level tutor the rest students with Remember level (50 %) about how to apply knowledge to solve daily-life problems by using U-CTRL to promote level of their cognitive processes to, at least, Apply level. After crossing the chasm, to further enhance students' high-level cognitive processes, learning activity and challenges that requires students' higher level of cognitive processes need to be introduced, but slowly.

Applications of U-CTRL for Learning English as a Foreign Language

This study will design English learning activity to provide students with opportunity to learn and practice English in real-life situation. It will include parent-teacher interaction and parental involvement in order to improve students learning achievement [6]. The design will be tested against its potential to contribute to academic achievement of students.

The first step of learning activity includes remember and understand new vocabulary. Students freely explore surrounding context in school district. They learn new vocabulary through capturing learning objects in authentic contexts then the system identify objects and provide their names in English and in native language. In order to better understand pronunciation and meaning of new vocabulary, students will interaction with the teacher and other students about new vocabulary.

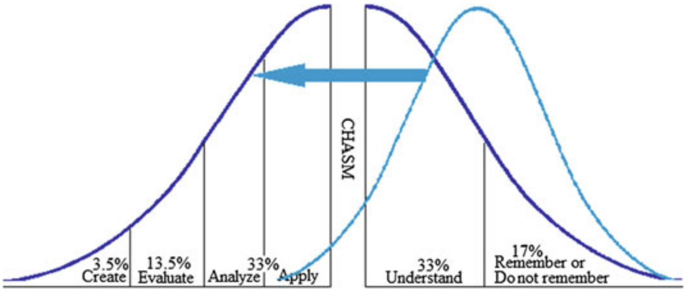


Fig. 2 Crossing the chasm

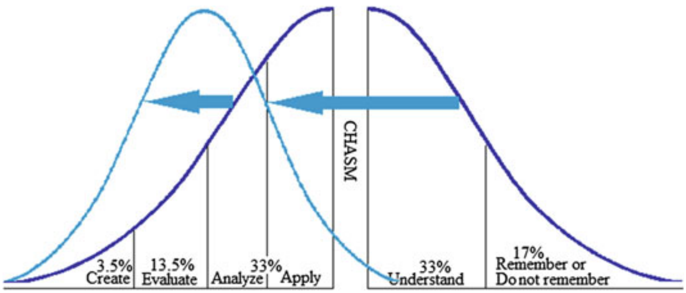


Fig. 3 After crossing the chasm

Capturing learning objects can be combined with specific strategies (meaning discovery strategy and memory reinforcement strategy). Students will collect different objects in school district show how each object relate to a specific situation in the district (meaning discovery strategy). Besides, students will be engaged in vocabulary learning and memorizing new words (memory reinforcement strategy). System identifies captured learning objects and provides their names in English, meaning, pronunciation, different sentence patterns including these words. Students practice new vocabulary repeatedly to facilitate their recall and understanding of new words and the system provides students with self-assessment feature so that students can monitor their learning process and progress. Words that students could not remember or understand during the test will be reminded by the system later so that students can practice them more. The following are some potential sub-activities: (1) Words exercise: Provide students with pictures, text, pronunciation, and meaning of words so that students can exercise words; (2) Vocabulary test: Provide students with pictures and name of words and then students need to give correct meaning of words; (3) Listening test: Provide students with pictures and pronunciation of words and then students need to give correct meaning of words; (4) Pronunciation test: Provide students with pictures of words and name recognized by the system and then students need to speak them out; (5) Spelling test: Provide students with pictures and pronunciation of words and then students need to spell out words; (6) Matching: Provide students with five pictures and five names of words and then then students need to match pictures and their names; (7) Using name of objects in sentences: Students capture objects in learning environment first, the system recognize name of these objects as text, and then students need to use name of objects in sentences.

The second step includes application of new vocabulary to solve daily-life problems. This study suggests using e-books with multimedia annotation tool [7, 8]; it can be carried everywhere to achieve seamless learning (e.g. take home so that parents may engage in learning process by helping their children). This study designed two sub-activities: (1) Introducing family members and (2) Introducing a menu of the dinner today. In the activities students expected to use simple English sentences to introduce family members and a menu of the dinner. Students can use e-books to record voices of family members, describe the dinner at home as well as to interact with their family members during the process of introduction. Teachers will still teach in traditional way and assign regular homework.

Evaluation

This study will carry out an empirical research to evaluate effects of U-CTRL on cognitive processes, analyze learning behavior of students to study with U-CTRL, and investigate students' perceptions and acceptance of the innovative approach.

A quasi-experimental design will be used in this study. Two classes of high grade elementary school students will be invited to participate in the experiment. One class with around thirty randomly assigned students will be the control group (with no treatment) and the other class with around thirty randomly assigned students will be the experimental group (with treatment). This study will administer a pre-test at the beginning of the experiment to assess a prior knowledge and prior cognitive processes of students. At the end of the experiment, after the treatment, this study will administer post-test to assess learning achievement and cognitive processes of students. A pretest–intervention–posttest design will allow evaluating effects of U-CTRL on learning achievement and cognitive processes of students. The targets of pre-test and post-test will focus on evaluating students' cognitive level, rather than scores, by designing test items based on the six levels of Bloom cognition. Therefore, the transition of students' cognitive processes could be analyzed to validate whether U-CTRL could facilitate them to cross the chasm and reach higher cognitive levels.

Meanwhile, regarding evaluating learning behaviors during U-CTRL activities, students will be motivated to capture learning objects, recognize them into text, and use both images and text for learning. All such learning behavior will be recorded and accumulated by students in learning portfolio. This study will explore students' learning behavior and analyze their portfolios to evaluate the transition of cognitive processes and their learning performance throughout the experiment.

Finally, this study will conduct a questionnaire survey and interviews with students to investigate students' perceptions, acceptance, and potential effectiveness of the innovative approach for learning.

Conclusion

This study proposed Cognitive Diffusion Model that distinguishes students' distribution of cognitive processes according pre-schooling, after-schooling, crossing the chasm, and high cognitive processes learning periods. Crossing the chasm is a very critical period as it promotes cognitive processes of most students from low to higher level. To facilitate crossing the chasm period, this study proposed four phases supported by User-Oriented Context-to-Text Recognition for Learning (U-CTRL). Students participate at learning activities by applying the knowledge they learnt in class to solve daily-life problems through capturing learning context, recognizing it into text, and employing for learning by using U-CTRL. Furthermore, one case related to learning English was proposed by using U-CTRL, and how to evaluate effectiveness of U-CTRL approach on cognitive processes of students was discussed.

References

1. Anderson LW, Krathwohl DR (eds) (2001) A taxonomy for learning, teaching and assessing: a revision of Bloom's Taxonomy of educational objectives, Complete edn. Longman, New York
2. PISA (2013) OECD program for international student assessment. Retrieved from http://www.oecd.org/pisa/aboutpisa/ on May 29, 2013
3. Rogers EM (2003) Diffusion of innovations, 5th edn. Free Press, New York
4. Moore GA (1999) Crossing the chasm: marketing and selling high-tech products to mainstream customers, Rev edn. HarperBusiness, New York
5. Kocakaya S, Gönen S (2010) Analysis of Turkish high-school physics-examination questions according to Bloom's taxonomy. Asia-Pacific Forum Sci Learn Teach 11(1)
6. Ho ESC, Kwong WM (2013) Effects of parental involvement and investment on student learning. In Parental Involvement on Children's Education, Springer, Singapore, pp 131–148
7. Hwang WY, Chen NS, Shadiev R, Li JS (2011) Effects of reviewing annotations and homework solutions on math learning achievement. British J Educ Technol 42(6):1016–1028
8. Hwang WY, Shadiev R, Huang SM (2011) A study of a multimedia web annotation system and its effect on the EFL writing and speaking performance of junior high school students. ReCALL 23(2):160–180

A Study of the Reader Recommendation Service Based on Learning Commons for Satisfaction of ePortfolio Users

Yu-Qing Huang, Cheng-Hsu Huang, Jen-Hua Yang and Tien-Wen Sung

Abstract This study implements co-interest recommending service embedded in ePortfolio for managing and sharing reader reading book experience from library records. We also propose five hypotheses based on proposed research model. Two hundred eleven Taiwanese graduate and undergraduate students participated in this research. The experimental results have shown that five hypotheses were supported.

Keywords ePortfolio · Recommendation · Virtual community

Introduction

The learning commons provides a range of programs and services to support learners' learning tasks. The learning commons not only assisting students in managing information, but also helping users manage learning. Recently, constructivism education has become new trend of teaching in universities [6]. On the constructivism learning theory, learning is a process of cooperative construction of the

Y.-Q. Huang (✉) · J.-H. Yang
Department of Computer Science and Information Engineering, National Central University, Jhongli, Taoyuan, Taiwan, Republic of China
e-mail: u9115903@ccms.nkfust.edu.tw

J.-H. Yang
e-mail: jhyang@csie.ncu.edu.tw

C.-H. Huang
Department of Computer Science and Information Engineering, Hwa Hsia Institute of Technology, Taipei, Taiwan, Republic of China
e-mail: Jeff@cc.hwh.edu.tw

T.-W. Sung
Department of Network Multimedia Design, Hsing Kuo University, Taipei, Taiwan, Republic of China
e-mail: kevin@mail.hku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_32, © Springer Science+Business Media Dordrecht 2014

knowledge by teachers and students. The learning commons can powerful support collaborative learning and learning activities through integrating of various resources and smart services. Learners in constructivism education acquire knowledge by a social process of learners' involvement into knowledge community [7].

Virtual learning community (VLC) is a virtual space of interactive collaboration which limited to network environments. According to constructivism learning theory, the learners of learning commons and VLC must actively probe into meaning of information in a social collaborative learning process. VLC is a new network learning organization established by web technologies to acquire knowledge and perform learning task through communicating with each other. The main distinguishing characteristic of VLC is its virtual space which limited to network environments. Li [6] evaluated VLC as an integral part of the learning commons and constructed a VLC in the learning commons. With the rapid development of web technology, ePortfolio has gradually developed become as a web-based information system for demonstrating the evidence of students' learning process over time in school education [10]. In order to provide more advanced and enhanced learning services, ePortfolio is used to manage the virtual learning community in grid environment [3].

One disadvantage of conventional virtual learning community is that learners are responsible for finding the right people to interact with. Therefore, a mechanism which can seamlessly integrate learning services, collaborators, and learning content is needed to enhance knowledge sharing in VLC and improve learning performance. As a result, finding the right collaborators whom they can derive tacit knowledge in an efficient manner is the most concerned issue of researches at present. The recommendation mechanism in the network is to recommend the information or products to users. Recommendation mechanisms based on social relations among people have become emerging fields. Recommendation people service may be considered an effective mechanism for finding common interest partners. In this study, a school-wide ePortfolios was chose as a platform to develop the co-intertest reader recommendation service. The main goal of recommendation co-interest reader service is to help students to find co-interest partners quickly to construct virtual learning community in ePortfolio.

Literatures Review

With the techniques of Web 2.0 initiated a new era of interactive cooperation and sharing knowledge, the concept of learning commons with the characteristic of cooperation learning was proposed. Li [6] asserted that the learning commons and virtual learning community (VLC) share the same basic concept, which emphasize that learners must not only actively explore the social meaning of information but also regard learning as a social collaborative process. In order to share knowledge in learning commons, finding common interest readers was the important task. Huang et al. [4] have demonstrated the recommending co-interest people service was a useful mechanism to build mobile learning networks. Therefore, recommendation people service may be considered an effective mechanism for finding co-interest partners. An ePortfolio is a digital electronic platform that stores visual and auditory content to demonstrate competencies and reflections in a field of knowledge to a teacher, a colleague, a professional, or a community. With the advent of new web technologies, the ePortfolio is also useful as a personal online virtual learning environment [1]. This study embedded recommending co-interest reader service to construct virtual learning community in ePortfolio. The goal of proposed recommending co-interest people service is to find right partners according to borrowing books from library's database.

Methodology

System Architecture

Finding common interest people was the important task in virtual learning community. Therefore, this study has proposed co-interest reader recommending service to construct virtual community in ePortfolio. The proposed co-intertest reader recommendation service finds co-interest readers according to the library resources of students. The library resources include reading statistics and the reader portfolio of reading authorized students, based on the functionalities provided by open source software called Mahara ePortfolio, which includes student portfolios, virtual community, and Web2.0 tools. The database contains library records, book categories, tables, and student portfolios of borrowing books. Figure 1 illustrates the proposed system architecture.

Instrument and Participants

An ePortfolio served as a virtual learning environment or knowledge-management system which is devised for students to collect, store and manage their learning

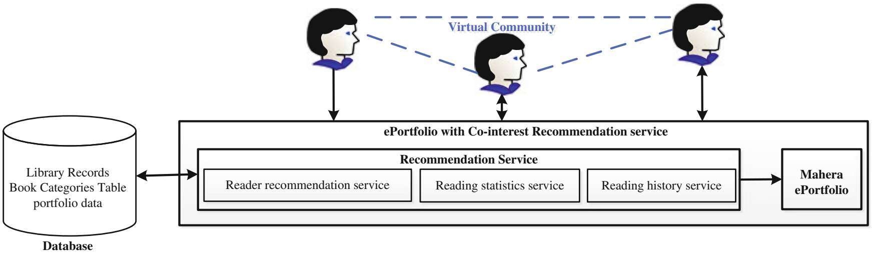


Fig. 1 System architecture

artifacts to demonstrate their competency [1, 5]. The proposed reader recommendation service provides an innovative service to find co-interest reader to construct virtual community in the Mehara ePortfolio system. Thus, this study not only considered ePortfolio as a platform for readers to share their reading experience in efficient manner but also applied recommendation co-interest reader service to help readers find co-interest readers quickly in virtual community. According to [12] KMS success model, the proposed reader recommendation service was used to enhance Mehara ePortfolio platform's system quality, information quality, system use, and learner satisfaction.

This study uses paper-based questionnaires to collect data. A seven-point Likert scale was developed for the measurement. System quality was measured with a 5-item developed scale consistent with [8]. Information quality was measured with a 5-item scale developed by [2]. Perceived system benefits were adapted from [12] with a 4-item scale. Fanally, student satisfaction was measured with a 4-item scale developed by Seddon and Kiew [9] and Wu and Wang [12].

Eighteen items relevant to the four factors of the proposed research model were adopted from existing literature. Data for this study were collected using an online questionnaire to survey universities in Taiwan. All 289 prospective participants were university graduate and undergraduate students. Of the 211 responses received, 189 were considered useful and used for analysis. The rate of useful response was 96.43 %.

Research Model and Hypotheses

Based on the results of a review in existing literature, this study included four factors in the research model (Fig. 2), which are system quality (SQ), information quality (IQ), perceived benefit (PB) and student satisfaction (SAT). The arrows in the research model are causal paths that represent the causal relationships between factors. Five hypotheses are proposed as follow:

H1: System quality positively influences perceived benefits
H2: Information quality positively influences perceived benefits
H3: System quality positively influences student satisfaction

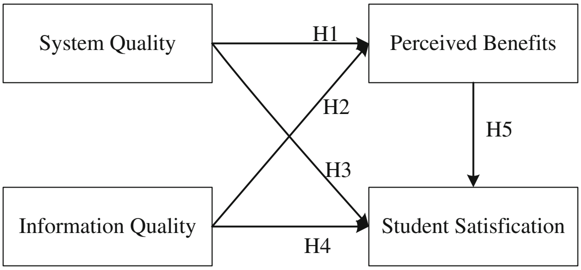


Fig. 2 Research model

Table 1 Results of multiple regression analysis of perceived system benefits (n = 189)

Independent variable	Dependent variable: perceived system benefits		
	β	t-value	Sig.
System quality	0.37	5.59	0.000***
Information quality	0.53	8.22	0.000***
F (2,186)	213.79***		
R^2	0.70		
Adjusted R^2	0.69		

$^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

H4: Information quality positively influences student satisfaction
H5: Perceived benefits positively influences student satisfaction

Results

A multiple regression analysis was conducted to test the hypotheses, shown in Tables 1 and 2. Hypotheses 1–2 examined the relationship between system quality, information quality and perceived system benefits. The result of regression analysis is shown in Table 1. From Table 1, we can know that system quality was a predictor of perceived system benefits ($\beta = .37, p < 0.001$), and that information quality was also a predictor of perceived system benefits ($\beta = .53$, $p < 0.001$). Therefore, system quality and information quality have positive impact on perceived system benefits. The results supported Hypothesis 1 and 2.

Hypotheses 3–5 examined the relationship between system quality, information quality, perceived system benefits and student satisfaction. The result of regression analysis is shown in Table 2. The results of Table 2 showed that system quality was a predictor of student satisfaction ($\beta = .11, p < 0.05$), and that information quality was also a predictor of student satisfaction ($\beta = .21, p < 0.01$), and that perceived system benefits was also a predictor of student satisfaction ($\beta = .57$, $p < 0.001$).

Table 2 Results of multiple regression analysis of student satisfaction (n = 189)

Independent variable	Dependent variable: student satisfaction		
	β	t-value	Sig.
System quality	0.11	2.08	0.039*
Information quality	0.21	3.48	0.001**
Perceived system benefits	0.57	8.89	0.000***
F (5,183)	192.32***		
R^2	0.84		
Adjusted R^2	0.84		

$^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

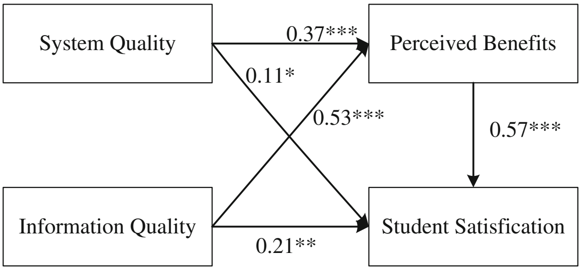


Fig. 3 Final model of multiple regression analysis (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$)

Therefore, system quality, information quality and perceived system benefits have positive impact on student satisfaction. The results supported Hypothesis 3–5.

Based on the results of previous studies, system quality, information quality and perceived benefits have positive affect on student satisfaction [2, 8, 11]. From a management information system perspective, perceived benefits have positive impact on satisfaction [9]. Wu and Wang [12] pointed out that perceived benefits have positive impact on satisfaction in view of knowledge information systems. In our study, system quality, information quality, and perceived benefits are all supported, similar to existing studies. (Fig. 3).

Conclusion

This study developed a research model for explaining and predicting factors influencing reader satisfaction in the context of higher education. Based on the concept of learning commons, the co-interest reader recommendation service was implemented to recommend co-interest partners according to library reading records to construct virtual community in ePortfolio. The experimental results have shown that system quality and information quality have positive impact on perceived system benefits. Readers also feel satisfy with system depend on system quality, information quality and perceived benefits. From these experimental results we can know that the co-interest reader recommendation service was benefit to improve the opportunity for sharing knowledge or experience on virtual community of ePortfolio.

Acknowledgments This work is supported by National Science Council, Taiwan under grants NSC 101-2511-S-146 -001.

References

1. Barrett HC (2009) Online personal learning environments: Structuring electronic portfolios for lifelong and life wide learning. On Horiz 17(2):142–152
2. Bliemel B, Hassanein K (2007) Consumer satisfaction with online health information retrieval: a model and empirical study. E-Serv J 5(2):53–81

3. Gouardères G, Conté E (2006) E-Portfolio to promote the virtual learning group communities on the grid. Int J Inf Technol Web Eng 1(2):25–42
4. Huang JJS, Yang SJH, Huang Y-M, Hsiao IYT (2010) Social learning networks: build mobile learning networks based on collaborative services. Educ Technol Soc 13(3):78–92
5. Kim P, Olaciregui C (2008) The effects of electronic portfolio-based learning space on science education. Br J Educ Technol 39(4):700–714
6. Li MJ (2009) Construction of virtual learning community in learning commons of academic libraries. 2009 IEEE international symposium on IT in medicine and education (ITME 2009) vol 1, pp 565–569
7. McMahon M (1997) Social constructivism and the world wide web—a paradigm for learning. 1997 ASCILITE Conference, Curtin University, Perth
8. McKinney V, Kanghyun Y, Fatemeh Z (2002) The measurement of web-customer satisfaction: an expectation and disconfirmation approach. Inf Syst Rev 13(3):296–315
9. Seddon PB, Kiew MY (1996) A partial test and development of the DeLone and McLean's model of IS success. Aust J Inf Syst 4(1):90–109
10. Tosh D, Penny Light T, Fleming K, Haywood J (2005) Engagement with electronic portfolios: challenges from the student perspective. Can J Learn Technol 31:89–110
11. Wixom BH, Todd PA (2005) A theoretical integration of user satisfaction and technology acceptance. Inf Syst Res 16(1):85–102
12. Wu JH, Wang YM (2006) Measuring KMS success: a respecification of the DeLone and McLean's model. Inf Manage 43(2):728–739

A Study of the Wikipedia Knowledge Recommendation Service for Satisfaction of ePortfolio Users

Cheng-Hsu Huang, Yu-Qing Huang, Jen-Hua Yang and Wen-Yen Wang

Abstract This study extended a conventional ePortfolio by proposed Wikipedia knowledge recommendation service (WKRS). Participants included 100 students taking courses at National Central University which were divided into experimental group and control group. The control group students and experimental group students have created their learning portfilios by using ePortfolio with WKRS and conventional ePortfolio without WKRS, respectively. The data for this study was collected over 3 months. The experimental results have shown that the learners' satisfaction, system use, system quality and information/knowledge quality of experimental group students have significant progress than control group students.

Keywords ePortfolio · Wikipedia knowledge · Satisfaction

C.-H. Huang (✉)
Department of Computer Science and Information Engineering, Hwa Hsia Institute of Technology, New Taipei, Taiwan, Republic of China
e-mail: Jeff@cc.hwh.edu.tw

Y.-Q. Huang · J.-H. Yang
Department of Computer Science & Information Engineering, National Central University, Jhongli, Taiwan, Republic of China
e-mail: u9115903@ccms.nkfust.edu.tw

J.-H. Yang
e-mail: jhyang@csie.ncu.edu.tw

W.-Y. Wang
Department of Information Science, Kun Shan University, Tainan, Taiwan, Republic of China
e-mail: wwang@mail.ksu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_33, © Springer Science+Business Media Dordrecht 2014

Introduction

An ePortfolio serves as a personal virtual space in which students can construct, collect and organize digital artifacts as well as present their effort, progress and achievements in learning process. An ePortfolio not only provides students with opportunities for self-reflection, but also allows teachers to assess and understand their students' current learning status [1, 2]. Many schools and universities have developed ePortfolios to improve the quality of education and training. During the learning process of school education, students collect, self-reflect, evaluate and connect knowledge artifacts in ePortfolios to achieve their learning goals.

In recent years, many teachers have used ePortfolio to guide their students to deeper learning. However, the benefits of using ePortfolio are weakened because most students do not spend time and effort to maintain their learning portfolios unless they are forced to do so. Jafari [3] has proposed that the "advanced feature", an interactive intelligent service for helping students to maintain their learning portfolios, is one of seven essential attributes for successful ePortfolios. For enhancing the quality of students' learning, advanced features offer a convenient environment for students' reflections in ePortfolio. From this viewpoint, one may say that advanced features can contribute to the stickiness and success of ePortfolios.

Research on ePortfolios in education has been conducted for more than a decade. Gorbunovs [4] classified ePortfolios into three levels: the simplest ePortfolio, higher-level ePortfolio, and modern ePortfolio. The simplest ePortfolio works with MS Office to present students' achievements. Higher-level ePortfolio offer some interactivity, allowing students to communicate with others by using GoogleDocs and Web 2.0. Modern ePortfolio provide interactivity, data management, and reporting systems for assessment. An ePortfolio with more functions and tools can better facilitate students' learning process. In the future, ePortfolios are expected to be capable of offering advanced features.

During learning process for constructing knowledge, the relevant information of knowledge can help students improve the quality and efficiency of learned knowledge. In the recent years, there are many studies have developed searching and recommendation methods for helping knowledge workers to acquire relevant information [5]; (Chen and Chung 2008). For helping students to efficiently construct knowledge in conventional ePortfolios, this study proposes a WKRS based on searching and recommendation technologies to provide relevant information of knowledge. The experimental results have shown that the students' satisfaction, system use, system quality and knowledge/information quality in the experimental group were significantly higher than the control group. This is meaning that the embedded WKRS in ePortfolio can provide suitable knowledge to help students to organize their learning portfolios in specific knowledge domains without any additional system burden.

Literatures Review

Improving Students' Learning in ePortfolio by Searching and Recommendation Technologies

To investigate how to successfully develop ePortfolios, Jafari [3] proposed seven essential attributes that affect their success. These seven essential attributes include ease of use, sustainable business plan, advanced features, robust integrated technology architecture, lifelong support, standards and transportability, and the x-or other-attribute. For the advanced features attribute, ePortfolio should provide attractive, unique, flexible, and interactive intelligent services to help students to complete and maintain their learning portfolios. Sequentially, students can be guide to deeper learning through the advanced intelligent features embedded in ePortfolio. It will be seen from this that the value of ePortfolio is closely with embedded intelligent features. Therefore, this study embedded an intelligent service WKRS in conventional ePortfolio to provide relevant knowledge during students' learning process.

ePortfolios have recently been viewed as knowledge management systems that provide new opportunities for learning [6]. Promoting the knowledge creation and sharing can enhance the effectiveness of individual lifelong learning [7]. Many studies have developed searching methods for improving the effectiveness of sharing and creating knowledge. For example, Liaw [5] indicated that search engines are important knowledge construction tools for creating individual knowledge. The recommendation function also plays an important role for helping students to construct personal knowledge by reading recommended relevant information. The main purpose of existing recommendation systems is to determine the learning materials to be recommended by analyzing the students' learning portfolios [8–10]. For example, Chen and Hsu [9] have implemented a Personalized Intelligent Mobile Learning System for recommending English news articles to students. Therefore, this study develops the WKRS based on searching and recommendation technologies to assist students' knowledge construction in ePortfolio.

Measuring Success in ePortfolios

In e-learning field, the information systems success model can be utilized in examining e-learning systems success [11]. In this model, the effect factors of system success include of system quality, information quality, system use, and user satisfaction. In the KMS success model, the information quality is replaced by knowledge/information quality. Kim and Olaciregui [6] indicated that ePortfolios can be viewed as a knowledge management system (KMS) for students to demonstrate their learning. This study uses the effect factors of system quality, knowledge/information quality, system use, and learners' satisfaction to measure the successful of the proposed ePortfolio with WKRS.

Procedure, Instruments and Hypotheses

To investigate the influences of provided ePortfolio features on students' learning satisfaction, we have develop a conventional ePortfolios platform and extended the conventional ePortfolios platform by proposed intelligent service. This study proposes a WKRS intelligent service that collects Wikipedia content in certain domains (more than 20,000 articles) to provide relevant knowledge in students' learning process. The WKRS provides two functions: (1) Keyword annotation (2) Article recommendation. Figure 1 shows the system architecture of embedding WKRS ePortfolio. The "Keyword annotation" function would annotate keywords of the students' learning articles and sent the relevant knowledge of annotated keywords to students. The relevant knowledge was acquired from google or yahoo searching engine according to students' query decisions. When the learning portfolio was created by students, the "Article recommendation" function would annotate keywords of the created learning portfolio and recommend the relevant Wikipedia knowledge articles of annotated keywords to students.

One hundred students from two classes at National Central University participated in this experiment from February 2010 to April 2010. One class, consisting of 49 students using the embedded WKRS ePortfolio, served as the experimental group. Another class, consisting of 51 students using the conventional ePortfolio, served as the control group. Students received ePortfolio training in March 2009. Every week, students spent three hours defining the goal of their portfolio, collecting information related to the course, recording their reflection and learning experience, and organizing curriculum files on ePortfolio after class. Teachers did not participate in students' creation of ePortfolios or give any comment during the experiment.

This study uses paper-based questionnaires to collect data. A seven-point Likert scale was developed for the measurement. This study included system quality scale, knowledge/information quality scale, ePortfolio system use scale, and ePortfolio learner satisfaction scale. The system quality (7 items, $\alpha = 0.90$) was assessed with scales adopted from Wang et al. [12]. The knowledge/information quality (11 items, $\alpha = 0.88$) and system use (5 items, $\alpha = 0.91$) were assessed

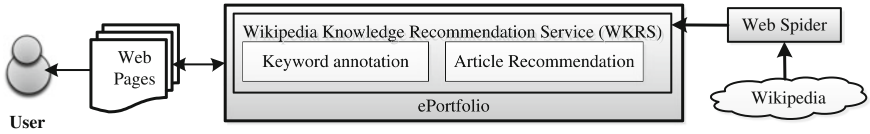


Fig. 1 System architecture of embedded WKRS ePortfolio

with scales adopted from Wu and Wang [13]. Learner satisfaction (4 items, $\alpha = 0.92$) was assessed with scales adopted from Seddon and Kiew [14]. The research question of this study is to investigate the influence of proposed WKRS intelligent service on ePortfolio system quality, knowledge/information quality, system use, and learner satisfaction. So, this study proposes the following hypotheses:

- Hypothesis 1. The WKRS has a positive effect on the system quality of the ePortfolio.
- Hypothesis 2. The WKRS has a positive effect on the knowledge/information quality of the ePortfolio.
- Hypothesis 3. The WKRS has a positive effect on the system use of the ePortfolio.
- Hypothesis 4. The WKRS has a positive effect on the learner satisfaction of the ePortfolio.

Results

Hypotheses 1–4 examine the effects of the WKRS on ePortfolio system quality, knowledge/information quality, system use, and learners' satisfaction. The Independent Samples *t* test was used to measure the difference in each variable between the experimental and control groups. Table 1 shows that students' satisfaction in the experimental group was significantly higher than the control group ($t = 12.56$, $p < 0.01$). The knowledge/information quality of the experimental group was significantly higher than that of the control group ($t = 9.91$, $p < 0.01$). Learners' satisfaction in the experimental group was also significantly higher than that of the control group ($t = 12.56$, $p < 0.01$), and system use was significantly higher in the experimental group than the control group ($t = 9.96$, $p < 0.01$). These results support Hypotheses 1–4.

Table 1 Independent *t*-test of prerequisites between experimental and control groups

Prerequisites	Experimental group		Control group		*t*	Sig.
	Mean	SD	Mean	SD		
System quality	5.65	0.61	4.32	0.59	11.12	0.00**
Knowledge/information Quality	5.55	0.58	4.19	0.77	9.91	0.00**
Students' satisfaction	5.59	0.64	4.00	0.62	12.56	0.00**
System use	5.02	0.81	3.30	0.95	9.96	0.00**

Note *$p < 0.05$, **$p < 0.01$

Conclusion

Intelligent services for managing knowledge can promote the personal core competitiveness of students through deeper learning in ePortfolio. To assist students construct knowledge in ePortfolio, this study proposes the WKRS intelligent service embedded in ePortfolio to automatically provide annotated keywords and recommended articles from Wikipedia after information is collected, chose, and reflected on by students for enhancing knowledge or information quality. Compared with the conventional ePortfolio without WKRS, the extended ePortfolio with WKRS provides more knowledge relevance and better search efficiency. The results of a series of experiments on this ePortfolio show that the WKRS can enhance the system quality, knowledge/information quality, system use, and learner satisfaction. These experimental results show that the WKRS can provide suitable knowledge to help learners organize their ePortfolio in specific knowledge domains with no additional system burden. Students also feel satisfied with the system and effectively use it to meet their needs.

In recent years, the applications of mobile devices have gradually increased in the field of education. The proposed WKRS intelligent service can strengthen the learning feedback of an ePortfolio and enhance learner satisfaction. In the future, we hope to apply the WKRS to other devices, such as smart phones or e-books, to facilitate students' learning in ubiquitous learning (u-learning) environments.

Acknowledgments This work is supported by National Science Council, Taiwan under grants NSC 101-2511-S-146 -001.

References

1. Barrett H (2000) Create your own electronic portfolio. Learn Lead Technol 27(7):14–21
2. Barker KC (2006) Environmental scan: overview of the ePortfolio in general and in the workplace specifically. Retrieved on 20 Oct 2006 from http://www.FuturEd.com
3. Jafari A (2004) The "sticky" ePortfolio system: tackling challenges and identifying attributes. J EDUCAUSE Rev 39(4):38–49
4. Gorbunovs A (2011) Prospective propulsions to embed artificial intelligence into the e-portfolio systems, In Al-Dahoud A et al (eds) Advances in information technology from artificial intelligence to virtual reality. UbiCC Publishers, pp 44–59
5. Liaw S–S (2005) Developing a web assisted knowledge construction system based on the approach of constructivist knowledge analysis of tasks. J Comput Human Behav 21(1):29–44
6. Kim P, Olaciregui C (2008) The effects of electronic portfolio-based learning space on science education. Br J Educ Technol 39(4):700–714
7. Li W, Liu Y (2008) Personal knowledge management in e-learning era. Technol e-learning Digital Entertainment Lect Notes Comput Sci 5093:200–205
8. Aleksandra K-M, Boban V, Mirjana I, Zoran B (2011) E-Learning personalization based on hybrid recommendation strategy and learning style identification. Comput Educ 56(3):885–899
9. Chen CM, Hsu SH (2008) Personalized intelligent mobile learning system for supporting effective english learning. J Educ Technol Soc 11(3):153–180

10. Hsu C-K, Hwang G-J, Chang C-K (2010) Development of a reading material recommendation system based on a knowledge engineering approach. Comput Educ 55(1):76–83
11. Freeze RD, Alshare KA, Lane PL, Wen HJ (2010) IS success model in e-learning context based on students' perceptions. J Inf Syst Educ 21(2):173–184
12. Wang Y-S, Wang H-Y, Shee DY (2007) Measuring e-learning systems success in an organizational context: scale development and validation. Comput Hum Behav 23(1):1792–1808
13. Wu J-H, Wang Y-M (2006) Measuring KMS success: a respecification of the DeLone and McLean's model. J Inf Manage 43(6):728–739
14. Seddon PB, Kiew MY (1994) A partial test and development of the DeLone and McLean model of IS success. In: International conference on information systems. Vancouver, Canada, pp 99–110

Using Personal Smart Devices as User Clients in a Classroom Response System

Tien-Wen Sung, Chu-Sing Yang and Ting-Ting Wu

Abstract This study proposed a classroom response system (CRS) different from existing commercial product. Modern and widespread used personal smart devices are utilized as the teacher-side controller and student-side response devices in the CRS instead of early infrared or radio frequency-based remote control. A prototype was developed for the proposed CRS, and it will be kept developing for further full functionality in CRS with the advantages and features of smart devices and modern network technologies.

Keywords Classroom response system · Smart device · E-learning

Introduction

A Classroom Response System (CRS) [1] is a technology-enabled learning system that allows instructors to project questions onto the screen, collects students' responses immediately and reports the feedback results in a class. It is also called an Interactive Response System (IRS) [2], an Audience Response System (ARS) [3], or a Student Response System (SRS) [4]. A CRS consists of three major parts: hardware, software, and communications. As mobile technology advances, there are various smartphone or pad products appeared on the consumer electronics

T.-W. Sung (✉) · C.-S. Yang
Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China
e-mail: tienwen.sung@gmail.com

C.-S. Yang
e-mail: csyang@ee.ncku.edu.tw

T.-T. Wu
Department of Information Management, Chia-Nan University of Pharmacy and Science, Tainan, Taiwan, Republic of China
e-mail: wutt0331@mail.chna.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_34, © Springer Science+Business Media Dordrecht 2014

market and used in people's daily life in recent years. These smart mobile devices are fully capable of serving as all-in-one computing devices and can offer a lot of potential applications, especially educational ones. In this paper, a prototype of smart device-based CRS is proposed. The objective is to develop a CRS with modern technologies and personal smart devices instead of earlier ones, and to make CRS more convenient by using smartphones or pads, now in widespread use, as user clients in the classroom.

CRS-Related Works

Earlier CRS systems used specific controllers and infrared signals for user response transmission. For example, EduClick II [2] consists of infrared signal transmitters and a corresponding receiver. The transmission is one-way and works with direct line of sight. A radio frequency-based CRS [5] uses RF signals instead of infrared for response transmission. It has less limitations and disadvantages but still has limited signal transmission range. A web-based CRS [6] utilizes Internet infrastructures to transmit user response data, and uses web technologies instead of specialist software and hardware devices of IR and RF-based CRS. A mobile CRS is another type. In [7], a non-smartphone Java application was presented for enhancing lecture interaction between a teacher and the students. In addition, a short message service (SMS) based classroom interaction system was proposed in [8]. Regarding to the evaluation, Kay et al. [9] have examined and summarized the benefits and challenges when using a CRS by a review of the literatures, and provided suggestions for further investigation. And Lin et al. [10] proposed results of evaluating students' perceptions of an SMS-based e-learning CRS before and after its implementation by applying a postexposure model called the information technology (IT) continuance model.

Proposed Smart Device-Based CRS

Learning Environment

As shown in the Fig. 1, this study proposed a smart device-based classroom response system, SD-CRS for short, which works on a WiFi or 3G enabled campus that provides a wireless data communication network. In the classroom, the teacher uses a smart device (smartphone or pad) held in the hand to control the learning content presentation. Students use hand-held smart devices to make responses. The CRS server can collect the responses and immediately generate a feedback report for showing on the screen and providing important information about learning effects and interaction outcomes.

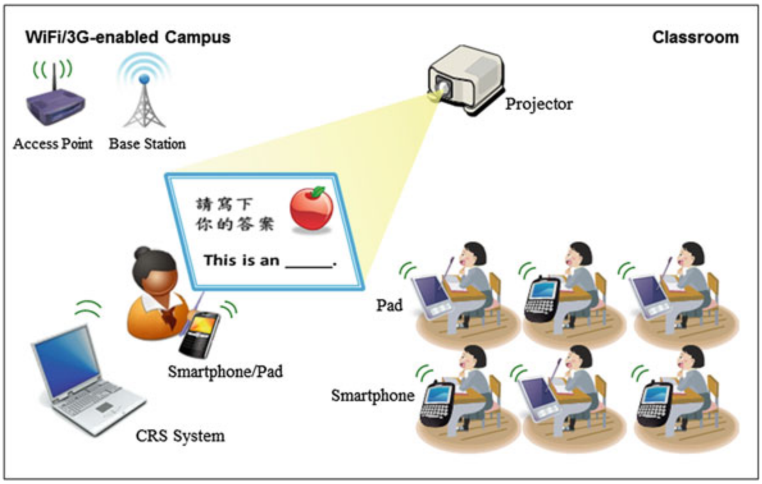


Fig. 1 SD-CRS learning environment and scenario

Design

Although there are some existing commercial CRS systems and applications used in practice, the system still can be promoted to a more convenient level. The key point is the type of the user device. The existing CRS products use specific remote controller with infrared (IR) or radio frequency (RF) signal transmission technology as user client devices. In this study, widespread used smartphones and pads are used instead. The comparison and differences between these two types of CRS are shown as below (Table 1).

Using modern smart devices as client devices has many advantages. A smart device is not merely used as a controller but also able to develop more special applications. For example, the feature of full keyboard can be utilized to input

Table 1 Comparison between general commercial CRS and the proposed SD-CRS

	General commercial CRS	Proposed SD-CRS
Client device	Specific remote controller	Smartphone or pad
Receiver	Specific receiver device	Infrastructure
Communication	Infrared (IR) or radio frequency	WiFi/3G network
Transmission angle	Limited/directional (using IR)	Omni-directional
Transmission range	Shorter	Farther
Obstacle effect	Significant (using IR)	Not significant
Keyboard	Simple buttons	Full keyboard
Screen	No	Yes
Music sound	No	Yes
Computing power	No	Yes
Consumer electronics	Not	Yes

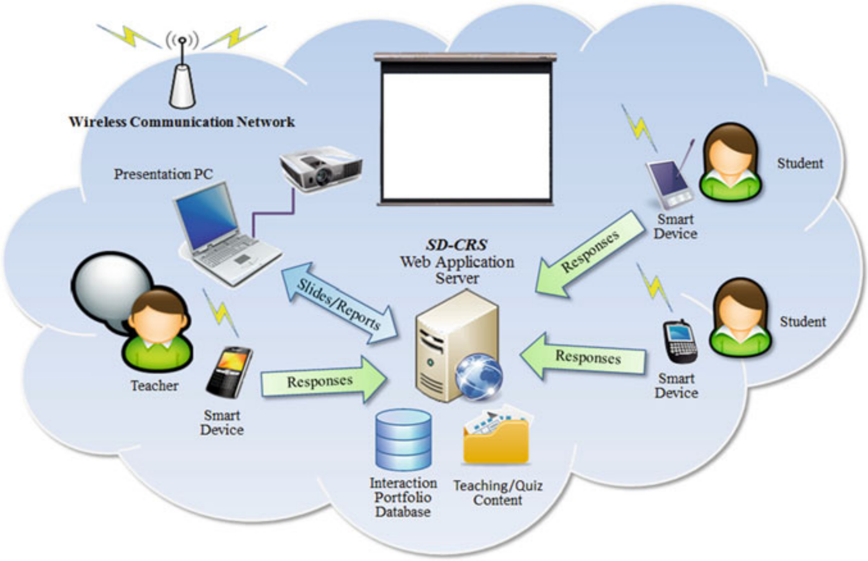
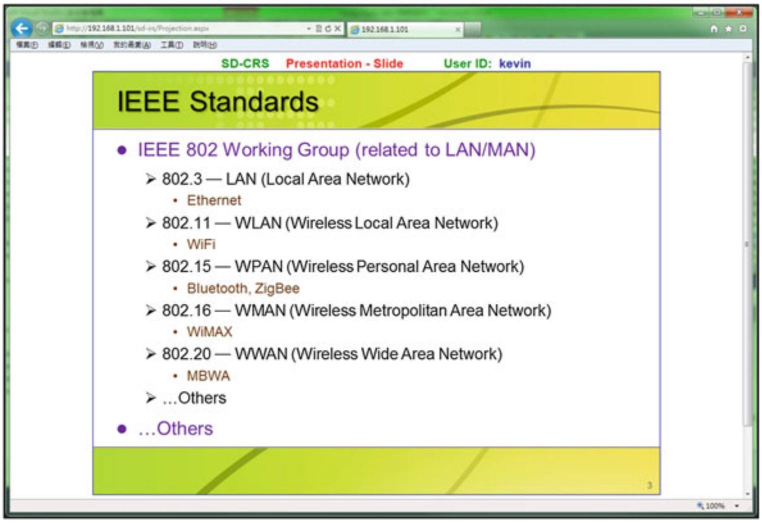


Fig. 2 SD-CRS system architecture

words as a response while simple buttons of traditional controller only can be pressed to select an option. For another example, the screen of a smart device also can be utilized to present additional information or contents within a CRS application.

Figure 2 shows the system architecture of the proposed SD-CRS prototype. The system uses a web-based design, which has three major advantages: (1) The system can work well on various personal smart devices with different OS platforms such as Android, iOS, and Windows Phone. (2) It has no need to install or update SD-CRS APP on the smart devices. The application is on a web server and any modification of the web application should be only made on the SD-CRS server. (3) Even a laptop or desktop PC can be used as a client device to make classroom responses. As shown in the figure, there are teaching and quiz contents stored in the SD-CRS sever, the teacher uses a presentation PC to show the contents by browsing SD-CRS pages and projecting to the large screen. The presentation can be controlled by the teacher's personal smart device. Students can make responses to the contents by the personal smart devices, too. Both teacher-side and student-side devices use web-based application to complete the functions. The interaction portfolio database (IPD) shown in the figure is used to record users' actions and response results in the SD-CRS. The IPD is further used to gather statistics and generate feedback reports regarding the students' responses. These reports are shown on the projection screen immediately. Then, it may cause students a valuable discussion and introspection in the classroom.


SD-CRS Presentation - Slide User ID: kevin

IEEE Standards

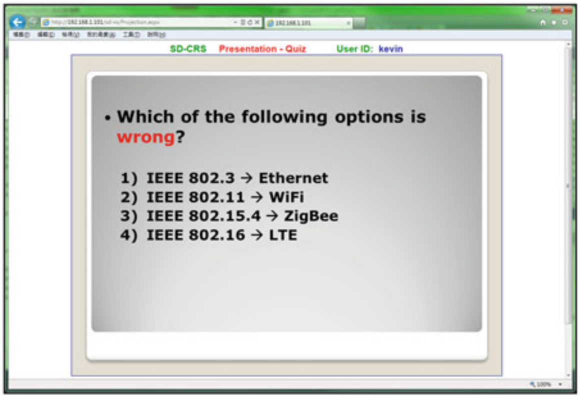
- IEEE 802 Working Group (related to LAN/MAN)
  - 802.3 — LAN (Local Area Network)
    - Ethernet
  - 802.11 — WLAN (Wireless Local Area Network)
    - WiFi
  - 802.15 — WPAN (Wireless Personal Area Network)
    - Bluetooth, ZigBee
  - 802.16 — WMAN (Wireless Metropolitan Area Network)
    - WiMAX
  - 802.20 — WWAN (Wireless Wide Area Network)
    - MBWA
  - ...Others
- ...Others

Fig. 3 Web-based teacher-side presentation screen (Slide)

Implementation

The prototype of SD-CRS was implemented by ASP.NET with C# and AJAX. Teaching slides and corresponding quizzes will be shown on the projection screen (Figs. 3 and 4). They can be switched and controlled by the function buttons on the interface of teacher-side smart-device (Fig. 5). A student can response to a quiz question by clicking a button or sending words (Fig. 5). The real-time feedback report then will be shown (Fig. 6) when the teacher clicks 'Report' button.

Fig. 4 Web-based teacher-side presentation screen (quiz)


SD-CRS Presentation - Quiz User ID: kevin

- Which of the following options is wrong?

1) IEEE 802.3 → Ethernet
2) IEEE 802.11 → WiFi
3) IEEE 802.15.4 → ZigBee
4) IEEE 802.16 → LTE

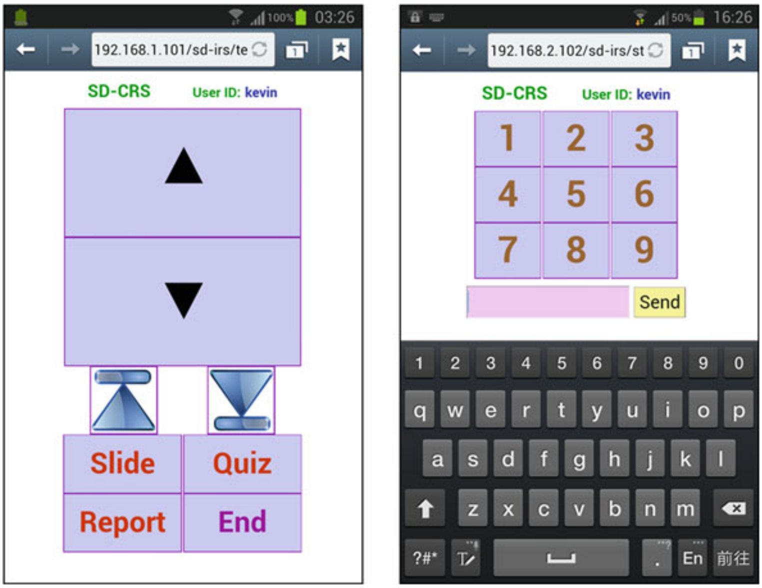
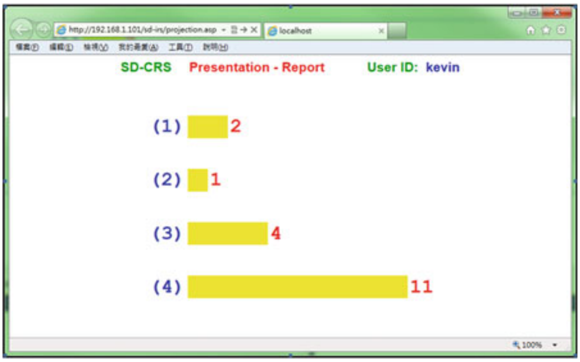


Fig. 5 Smartphone interfaces of teacher-side control (*left*) and student-side response (*right*)

Fig. 6 Web-based teacher-side presentation screen (report)



Conclusion and Future Works

Various smartphone and pad products of consumer electronics are more and more popular and in widespread use. And the infrastructure of wireless communication networks such as WiFi and HSPA + make people easy to use Internet-based applications. This study utilizes the features and advantages of these modern devices and technologies to develop a classroom response system for further assistance in learning.

A prototype is proposed in this paper, and further full functionality of the smart device-based CRS will be kept developing. A learning experiment and evaluation of the completed system will be made in the near future.

Acknowledgments This work is sponsored by National Science Council under grant number NSC 102-2511-S-041-002, Taiwan.

References

1. Siau K, Sheng H, Nah FF (2006) Use of a classroom response system to enhance classroom interactivity. IEEE Trans Educ 49(3):398–403
2. Liu TC, Liang JK, Wang HY, Chan TW, Wei LH (2003) Embedding EduClick in classroom to enhance interaction. In: 11th international conference on computers in education, Hong Kong, pp. 117–125
3. Petr DW (2005) Experience with a multiple-choice audience response system in an engineering classroom. In: 35th annual conference on frontiers in education, Indianapolis, pp S3G-1–S3G-6
4. Hall RH, Collier HL, Thomas ML, Hilgers MG (2005) A student response system for increasing engagement, motivation, and learning in high enrollment lectures. In: 7th Americas conference on information systems, Omaha, pp 1–7
5. Nelson M, Hauck RV (2008) Clicking to learn: a case study of embedding radio-frequency based clickers in an introductory management information systems course. J Inf Syst Educ 19(1):55–64
6. Mantoro T, Ayu MA, Habul E, Khasanah AU (2010) Survnvote: a free web based audience response system to support interactivity in the classroom. In: IEEE conference on open systems, Malaysia, pp 34–39
7. Costa JC, Ojala T, Korhonen J (2008) Mobile lecture interaction: making technology and learning click. In: IADIS international conference mobile learning, Portugal, pp 119–124
8. Scornavacca E, Marshall S (2007) TXT-2-LRN: improving students' learning experience in the classroom through interactive SMS. In: The 40th annual Hawaii international conference on system sciences, Hawaiipp, pp 1–8
9. Kay RH, LeSage A (2009) Examining the benefits and challenges of using audience response systems: a review of the literature. Comput Educ 53:819–827
10. Lin J, Rivera-Sanchez M (2012) Testing the information technology continuance model on a mandatory sms-based student response system. Commun Educ 61(2):89–110

Part VI
Embedded Computing

3D Bidirectional-Channel Routing Algorithm for Network-Based Many-Core Embedded Systems

Wen-Chung Tsai, Yi-Yao Weng, Chun-Jen Wei, Sao-Jie Chen and Yu-Hen Hu

Abstract Network-on-Chip (NoC) is an emerging technology designed for the communication of IPs in an embedded system. This paper proposes a 3D (Three-Dimensional) model for a Bi-directional NoC (BiNoC). This three-dimensional model inspires the development of a new routing algorithm for BiNoC, called Bidirectional Routing (Bi-Routing). Bi-Routing is a fully adaptive routing algorithm using different layers in the proposed three-dimensional model to avoid deadlock without prohibiting the use of any path. As such, Bi-Routing can improve the load balance and reduce the packet latency of an NoC. Experimental simulation results demonstrated superior performance compared with existing routing methods.

Keywords Three-dimensional (3D) · Network-on-chip (NoC) · Bidirectional channel · Routing algorithm

W.-C. Tsai (✉)
Department of Information and Communication Engineering, Chaoyang University of Technology, Taichung, Taiwan, Republic of China
e-mail: azongtsai@gmail.com

Y.-Y. Weng · S.-J. Chen
Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan, ROC
e-mail: niles90221@gmail.com

S.-J. Chen
e-mail: csj@cc.ee.ntu.edu.tw

C.-J. Wei · S.-J. Chen
Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, Republic of China
e-mail: d92921022@ntu.edu.tw

Y.-H. Hu
Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI, USA
e-mail: hu@engr.wisc.edu

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_35, © Springer Science+Business Media Dordrecht 2014

Introduction

System-on-Chip (SoC) uses numerous kinds of Intellectual Properties (IPs) and interconnections to form an embedded system in a single chip. As the technology progresses, the number and operating frequency of IPs are increasing. The bottleneck has transferred from IPs to interconnections. For example, with the deep sub-micron integrated circuit technology, crossing a chip with a highly optimized interconnects takes between six to ten clock cycles and only one set of IPs can use the traditional bus-based interconnection to transact data, such that the rest numerous IPs are waiting for the using right. Therefore, a new approach to designing the communication subsystem between IPs, Network-on-Chip (NoC), has been proposed in the past years to meet the design productivity and signal integrity challenges of next-generation system designs [1–3].

Routing is to decide which path a packet is to deliver. In other words, given a source and a destination, routing directs a packet where to go. A bad routing algorithm will let numerous packets pass through the same path or choose a longer path. We realize that the same route will lead to the lack of path diversity, and it creates a large load imbalance in the network. So path diversity provided by the adopted routing algorithm determines the performance of an NoC greatly. Most important of all, a deadlock will cause the on-chip interconnection crashed. Thus, routing algorithms must be deadlock-free.

A Bidirectional-channel Network-on-Chip (BiNoC) architecture was proposed to enhance the performance, quality-of-service, and fault-tolerance of on-chip communications [4–7]. BiNoC allows each communication channel to be dynamically self-configured to transmit flits in either direction in order to better utilize on-chip hardware resources. However, the conventional routing methods adopted in these BiNoC studies cannot fully exploit the path diversity of the BiNoC architectures. Accordingly, we present a three-dimensional model of BiNoC and a new routing algorithm for BiNoC called Bidirectional Routing (Bi-Routing). Bi-Routing can reduce packet latency and achieve higher bandwidth utilization due to its high path diversity by conditionally making channel bidirectional. Moreover, deadlock-freedom is provided with Bi-Routing which will be introduced in detail in this paper.

The rest of this paper is organized as follows. In Sect. “Background”, we will first introduce the background about BiNoC and some deadlock-free routing algorithms. Section “Methodology” will describe a 3-dimensional model of BiNoC and a routing algorithm based on the 3-dimensional model will be presented. In Sect. “Experimental Results”, we will show the experimental simulation results. Finally, Sect. “Conclusion” will draw a conclusion.

Background

First, we will introduce Bidirectional-channel Network-on-Chip (BiNoC) in Sect. "Bidirectional-Channel Network-on-Chip". Next, Sect. "Related Routing Algorithms" will compare several deadlock-free routing algorithms.

Bidirectional-Channel Network-on-Chip

In a conventional router, all channels are unidirectional, thus it may lead to the following scenario where one output channel is busy or in congestion and another channel is idle because the direction of the idle channel is an input channel. BiNoC was proposed to overcome this problem by make all channels bidirectional and to allow each communication channel to be dynamically self-configured to transmit flits in either direction. For example, as shown in Fig. 1a, every vertex represents a task with a value t_j of its computation time, and every edge represents the computing dependence with a value of communication volume. A mesh NoC with most optimized mapping is shown in Fig. 1b. We can find that the NoC only use three unidirectional channels with the other directional channels being idle. However, if we make the same mapping solution on BiNoC which can dynamically change the direction of each channel between each pair of routers as shown in Fig. 1c, the bandwidth utilization will be improved and the total execution time can be reduced.

Related Routing Algorithms

Considering load balance, we prefer to use an adaptive routing algorithm that has more paths to choose for a packet routing delivery. Glass and Ni presented an

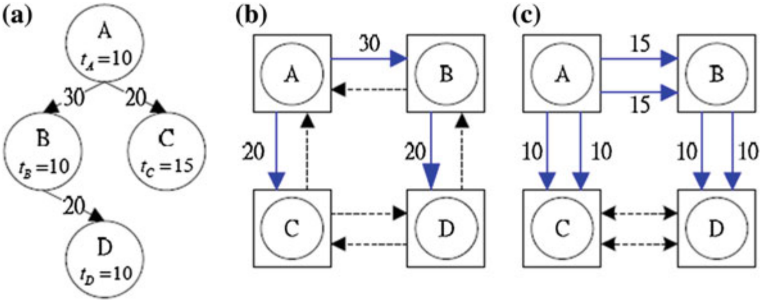


Fig. 1 Example of (**a**) task graph mapping to (**b**) a conventional NoC, and (**c**) a BiNoC

elegant concept of *turn model* [8]. The basic idea of turn model is to prohibit the minimum number of turns that break all of the deadlock cycles such that routing algorithms based on turn model can be deadlock-free. Three adaptive routing algorithms, namely west-first, north-last, and negative-first were designed based on turn model. We show that the four cases of prohibited turns of the three routing algorithms in Fig. 2. Note that the solid lines indicate the allowed turns and the dash line indicate the prohibited turns. For example, Case two uses the turn model that prohibits S–W turn and N–W turn. According to this turn model, west-first routing delivers all the packets to west first if packets need to be delivered to west. Similar with the west-first routing, negative-first routing and north-last routing were designed according to their own turn models. Turn model provides a simple way to design a deadlock-free adaptive routing. Nevertheless, there is a highly uneven routing path use problem in a global view. That is at least half of the source–destination pairs are limited to having only one minimal path, while full adaptive is provided for the rest of the pairs.

To solve the uneven routing path use problem, an odd–even turn model was presented by Chiu in [9]. With the odd–even turn model, any packet is not allowed to take an E-N turn or an E-S turn at any nodes located in an even column, and any packet is not allowed to take an N–W turn or an S–W turn at any nodes located in an odd column. This odd–even turn model restricts certain turns based on the locations such that none of the turns are eliminated in an NoC. Although the odd–even still restricts some turns for a packet to use, these restricted turns are unobvious in a global view. Therefore, the odd–even turn model has higher path diversity than other turn models. Based on the odd–even turn model, we can design an OE-Routing algorithm, which will compare with our proposed Bi-Routing algorithm in Sect. "Experimental Results".

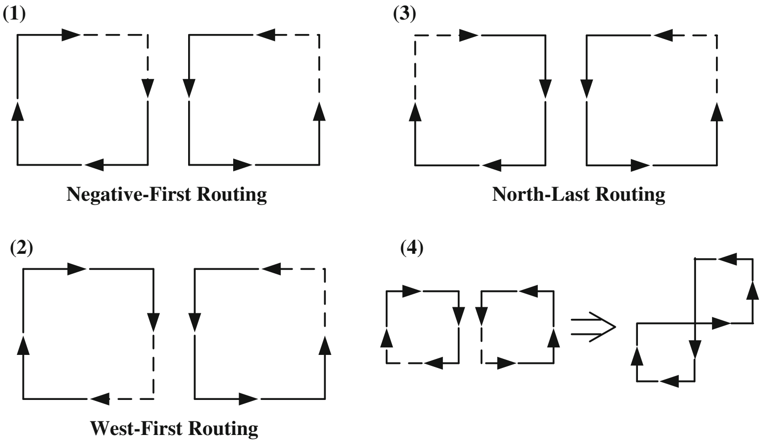


Fig. 2 Four cases of turn models

Methodology

In this section, we present the design methodology of our proposed Bi-Routing routing algorithm to exploit the characteristics of bidirectional channels and provide higher path diversity.

Three-Dimensional Model of BiNoC

Since the original model of mesh NoC cannot show the behavior of BiNoC, we have to represent these four kinds of bidirectional channel patterns as a three-dimensional model in Fig. 3a. The new Z-dimension is time related, which shows the channel diversity during time changed. The three-dimensional graph as shown in Fig. 3a is not a physical three-dimensional IC, but a conceptual model to represent the behavior of a BiNoC. Moreover, as shown in Fig. 3a, the odd–even turn model in BiNoC can also be represented in our three-dimensional model.

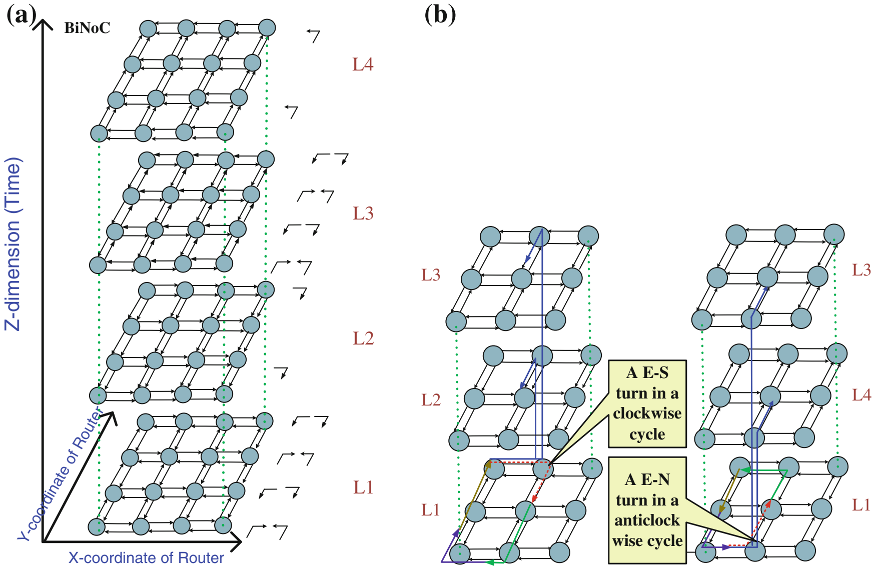


Fig. 3 **a** BiNoC three-dimensional model and **b** Cycles breaking example (rule 1)

Bidirectional Routing Algorithm

The three-dimensional model of BiNoC mentioned in Sect. "Three-Dimensional Model of BiNoC" indicates that BiNoC has higher path diversity than the original unidirectional NoC. We use this path diversity to develop a Bi-Routing algorithm for BiNoC in this section. The Bi-Routing idea is shown in Fig. 3b. On a unidirectional NoC, a deadlock cycle formed by the paths on the same layer can be broken by using another layer of channel (in the Z-dimension). Therefore, we need not prohibit any turn and all paths can be included in the feasible routing set of Bi-Routing. We develop the Bi-Routing based on Theorem 1 brought up in [10].

Theorem 1 A connected and adaptive routing function R for an interconnection network I is deadlock-free, if there are no cycles in its channel dependency graph.

A channel dependency graph D for a given interconnection network I and routing function R is a directed graph, D = G(C, E). The vertices of D are the channels of I. An arcs in D is a pair of channels (c_i, c_j) where there exists a direct dependency from c_i to c_j. The meaning of connected routing function is that for any packet, the connected routing function can find a path to deliver the packet to the destination. Therefore, from Theorem 1, if we can break the cycle in a channel dependency graph, the routing algorithm is deadlock-free. Hence, three rules are brought up for our Bi-Routing algorithm.

Rule 1: Packets use reverse channel at the E-S turn and the E-N turn.

To escape from deadlock in a BiNoC by using another layer to route, as shown in Fig. 3b, we choose E-S turn and E-N turn as a breaking position (using reverse channel in another layer) in clockwise and counter-clockwise cycles.

Rule 2:. Packets from south (north) reverse channel and delivered to north (south) must use reverse channel.

An inter-layer deadlock will appear without Rule 2. Rule 2 indicates that packets should keep using a reserve channel in south or north such that an inter-

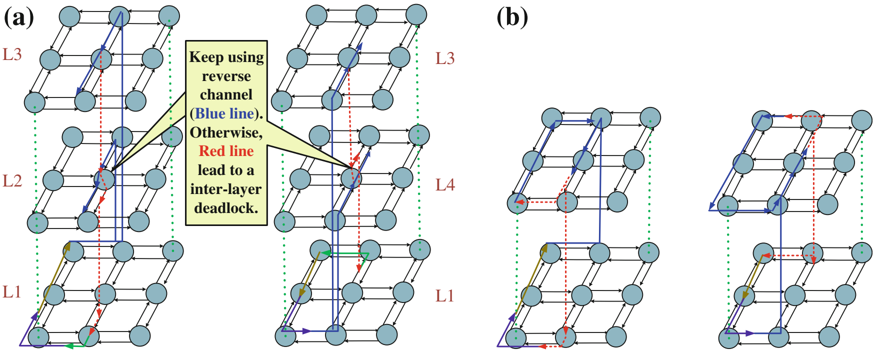

(a)
L3
L2
L1
Keep using reverse channel (Blue line). Otherwise, Red line lead to a inter-layer deadlock.
L3
L4
L1
(b)

Fig. 4 **a** Example of rule 2 and **b** Example of rule 3 layer deadlock can be removed as shown in Fig. 4a. In which, red dotted lines represent paths violating Rule 2 and lead to inter-layer deadlock conditions.

Rule 3: Packet form reverse channel cannot take S–W or N–W turn.

The essence of Rule 3 is similar to the conventional turn model; it eliminates a turn in just one layer. In other words, reverse channels will make up a cycle, if we do not prohibit packet be routed to a lower layer. With Rule 3, in three-dimensional model, packets cannot take S–W turn and N–W turn when packets are not in the L1 layer as shown in Fig. 4b. Where the red dotted lines represent prohibiting turns in a higher layer and lead to inter-layer deadlock conditions.

With the three rules, Bi-Routing can we provide a fully adaptive routing, which can spread traffic loads to the whole network instead of keeping some parts of network in heavy congestion.

Experimental Results

Our simulation environment comprised an 8 × 8 mesh. Three traffic patterns were used, including uniform, transpose, and hotspot traffics. In the uniform traffic, a node receives a packet from any other node with equal probability. Every node transmits packets to a randomized destination with a probability based on the injection rate. In the transpose traffic, a node at a source with coordinate (*i, j*) will sent a packet to a destination with coordinate (*j,i*). In the hotspot traffic, 20 % of the packets change their destination to some selected hotspots while the remaining 80 % of the traffic keep uniform. In this work, we chose (3, 3), (3, 2), (3, 1), (3, 0) as hotspots.

We simulated XY-Routing, west-first routing algorithm (WF-Routing), odd–even routing (OE-Routing), and our proposed Bi-Routing algorithm. The packets in our experiments were composed of 16 flits with one header flit and one tail flit. The capacity of the buffer in each of the 5 directions of channels was 8 flits using wormhole switching. We simulated our network by injecting loads, from 20 flits per clock cycle to 500 flits per clock, at each node. For each injection rate, the simulation time was 25,000 clock cycles. The results of latency in three traffic patterns are shown in Fig. 5a, and the results of throughput are shown in Fig. 5b.

The simulation results show that our bidirectional routing, Bi-Routing, has the best performance among the four algorithms. XY-Routing outperforms OE-Routing and WF-Routing because XY-Routing can distribute packets evenly in the uniform traffic condition. This part of results is the same as in [8, 9]. Our bidirectional routing algorithm still had better saturation throughput than XY-Routing, about a 6.9 % improvement, as shown in Fig. 5a, b. However, the throughput of bidirectional routing decreases much more than XY-Routing in high injection rate.

The transpose and hotspot traffic patterns are close to the real-case embedded system traffics, because in a SoC most IPs have communications with the main CPU core. Adaptive routing algorithms perform better than XY-Routing in transpose and hotspot traffic patterns, because adaptive routing algorithms have

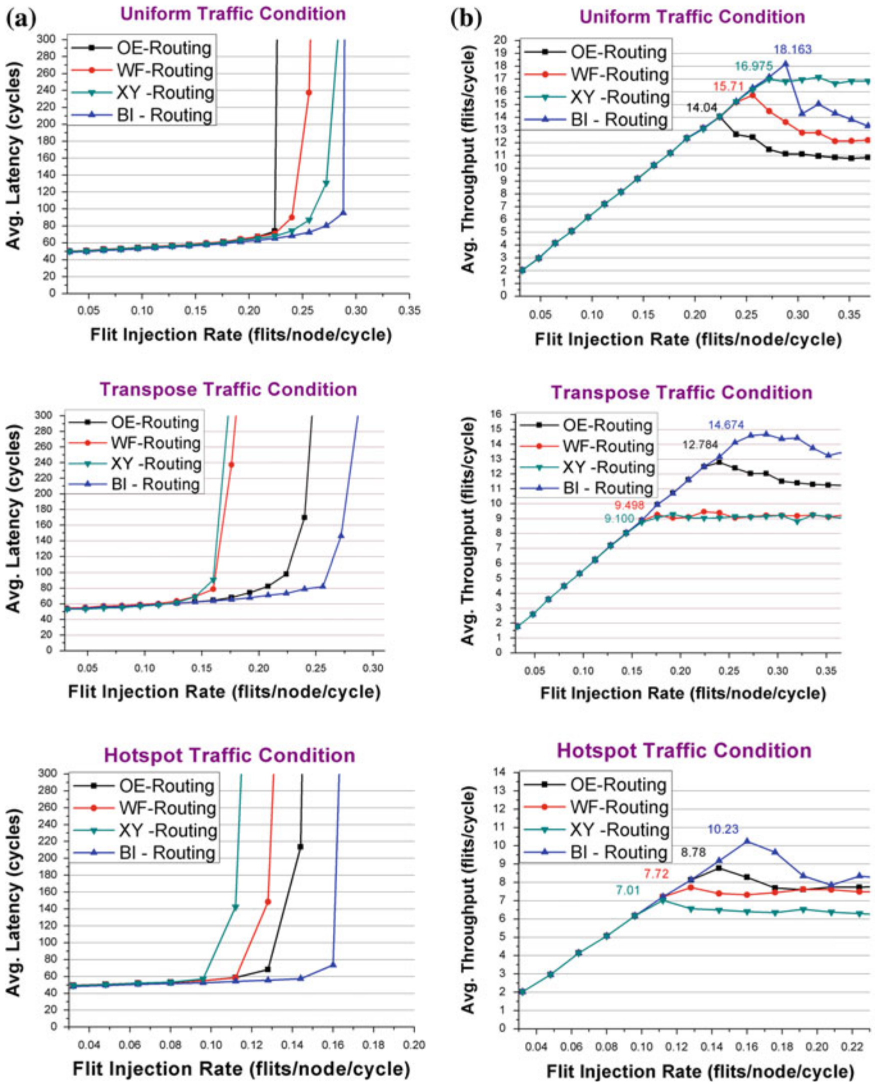


Fig. 5 **a** Latency and **b** Throughput versus injection rate under OE-routing, WF-routing, XY-routing, and Bi-routing

more paths to route. The Bi-Routing method had 14.78 and 16.51 % improvements over the OE-Routing one, in transpose traffic and hotspot traffic, respectively. The reason is our proposed Bi-Routing algorithm can spread traffic loads to relief local traffic congestion.

Conclusion

In this paper, we proposed a three-dimensional (3D) model of BiNoC. Based on this 3D model, we developed a new routing algorithm for BiNoC called Bi-Routing. Bi-Routing used the reversed channel to break the deadlock cycle in BiNoC, if any. Experimental results showed that our proposed Bi-Routing delivers better performance over the original BiNoC with OE-Routing because of the increased path diversity and the enhanced load balance provided by Bi-Routing.

Acknowledgments This work was partially supported by National Science Council, ROC, under grant NSC-101-2220-E-002-008.

References

1. Dally WJ, Towles B (2011) Route packets, not wires: on-chip interconnection networks. In: Proceedings of the design automation conference, pp 684–689
2. Benini L, DeMicheli G (2002) Networks in chips: a new SoC paradigm. IEEE Comput 35(1):70–78
3. Jantsch A, Tenhunen H, Ebrary I (2003) Networks on chip. Kluwer Academic Publishers, Dordrecht
4. Lan YC, Lo SH, Hu YH, Chen SJ (2009) BiNoC: a bidirectional NoC architecture with dynamic self-reconfigurable channel. In: Proceedings of the 3rd ACM/IEEE international symposium on network-on-chip, San Diego, pp 266–275
5. Lan YC, Lin HA, Lo SH, Hu YH, Chen SJ (2011) A bidirectional NoC (BiNoC) architecture with dynamic self-reconfigurable channel. IEEE Trans Comput Aided Des Integr Circuits Syst 20(3):427–440
6. Lo SH, Lan YC, Yeh HH, Tsai WC, Hu YH, Chen SJ (2010) QoS aware BiNoC architecture. In: Proceedings of the 24th IEEE international parallel & distributed processing symposium, Atlanta, pp 1–10
7. Tsai WC, Zheng DY, Chen SJ, Hu YH (2001) A fault-tolerant NoC scheme using bidirectional channel. In: Proceedings of the 48th design automation conference, San Diego, pp 918–923
8. Glass CJ, Ni LM (1994) The turn model for adaptive routing. J ACM 41(5):874–902
9. Chiu GM (2000) The odd–even turn model for adaptive routing. IEEE Trans Parallel Distrib Syst 11(7):729–738
10. Dally WJ, Seitz CL (1987) Deadlock-free message routing in multiprocessor interconnection networks. IEEE Trans Comput C-36(5):547–553

An Energy-Aware Routing Protocol Using Cat Swarm Optimization for Wireless Sensor Networks

Lingping Kong, Chien-Ming Chen, Hong-Chi Shih, Chun-Wei Lin, Bing-Zhe He and Jeng-Shyang Pan

Abstract In this paper, we propose an energy-aware routing protocol for wireless sensor networks. Our design is based on the ladder diffusion algorithm and cat swarm optimization algorithm. With the properties of ladder diffusion algorithm, our protocol can avoid the generation of circle routes and provide the backup routes. Besides, integrating cat swarm optimization can effectively provide better efficiency than previous works. Experimental results demonstrate that our design reduces the execution time for finding the routing path by 57.88 % compared with a very recent research named LD.

Keywords WSN · Routing · Cat swarm optimization

L. Kong · C.-M. Chen (✉) · C.-W. Lin · J.-S. Pan
Innovative Information Industry Research Center, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
e-mail: chienming.taiwan@gmail.com

L. Kong
e-mail: konglingping@utsz.edu.cn

C.-W. Lin
e-mail: jerrylin@ieee.org

J.-S. Pan
e-mail: jengshyangpan@gmail.com

C.-M. Chen · C.-W. Lin · J.-S. Pan
Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen 518055, China

H.-C. Shih
Department of Electronics Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan, Republic of China
e-mail: hqshi@bit.kuas.edu.tw

B.-Z. He
Department of Computer Science National, Tsing Hua University, Hsinchu, Taiwan, Republic of China
e-mail: ckshjerho@is.cs.nthu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_36, © Springer Science+Business Media Dordrecht 2014

Introduction

Recently, wireless sensor networks (WSN) attract considerable research attention since they have been deployed in various applications, such as military, environmental monitor, industry automation and smart space. A WSN is composed of a large number of sensor nodes which collaborate with each other. Each sensor node detects a target within its detection range, gathers useful date, performs simple computations and sends the package to the sink node. In fact, sensor nodes are constrained in battery power and energy capability; therefore, energy saving is necessary to be considered carefully when constructing WSN.

Several routing protocols [1–8] for WSN have been proposed. One well-known protocol named DD (Directed Diffusion) was proposed by C. Intanagonwiwat et al. [2] in 1999. DD attempts to achieve better power consumption by reducing the data relay. More specifically, in DD, the collected data is transmitted only when sensor nodes fit the query from the sink node. However, in a large-sized WSN, the waste of power consumption and storage becomes worse. Besides, the circle route problem also becomes more serious. In 2012, Ho et al. proposed a protocol named LD [3]. LD describes a ladder diffusion algorithm to solve the circle route problem in DD and utilizes ant colony optimization [9–13] to select the most suitable routing path. According to their simulation results, LD indeed reduces power consumption and increases data forwarding efficiency compared with DD. However, LD algorithm still can be improved. In order to obtain the most suitable path, LD requires to perform various iteration operations. It will cause slow performance.

In this paper, we propose an energy-aware routing protocol based on ladder diffusion [3] algorithm and cat swarm optimization [14–17] for WSN. With the properties of ladder diffusion algorithm, our protocol can avoid the generation of circle routes and provide the backup routes if some sensor nodes are captured or unavailable. On the other hand, integrating cat swarm optimization can effectively overcome the weakness of LD that performs various iteration operations. Experiment results demonstrate that our protocol reduces the execution time for finding the routing path by 57.88 % compared with LD. Besides, the energy consumption of LD is about three times greater than our protocol. As a result, the lifetime of whole WSN can effectively be extended.

The remainder of the paper is organized as follows. Section "Related Works" describes a brief review of some closely related works. The proposed routing protocol is illustrated in Sect. "The Proposed Routing Protocol". Simulation results are shown in Sect. "Simulations". Finally, we conclude this paper in Sect. "Conclusion".

Related Works

In this section, we first review the routing protocol LD proposed by Ho et al. [3]. Then, Cat Swarm Optimization (CSO) is described.

LD Algorithm

In 2011, Ho et al. [3] proposed a routing algorithm named LD based on ladder diffusion and ant colony optimization [9–13] for WSN. LD shows better performance in reducing energy consumption and increasing data forwarding efficiency compared with several previous well-known protocols [1, 2].

LD contains two phases, ladder diffusion and ant colony optimization. The purpose of the ladder diffusion phase is to identify routes from sensor nodes to the sink node and avoid the generation of circle routes. This phase also generates a *ladder table* for each sensor node. With the *ladder table*, LD can provide the back-up path and avoid redundant relay.

In the ant colony optimization phase, each sensor node can construct the most suitable routing path from itself to the sink node with the ladder table.

CSO

Chu et al. [14] proposed a new optimization algorithm named cat swarm optimization (CSO) in 2006. CSO imitates the natural behaviors of cats. Although cats spend lots of time in resting, they always remain alert. Once cats sense the presence of a prey, they chase it very quickly spending high energy. This two major behaviors of cats can be modeled into two modes, seeking mode and tracing mode. The seeking mode represents the behavior of cats which is resting and seeking the next position to move to. One the other hand, the tracing mode represents the behavior of cats which trace the target while spending large amount of energy. The detailed steps of CSO can be found in [14–17].

The Proposed Routing Protocol

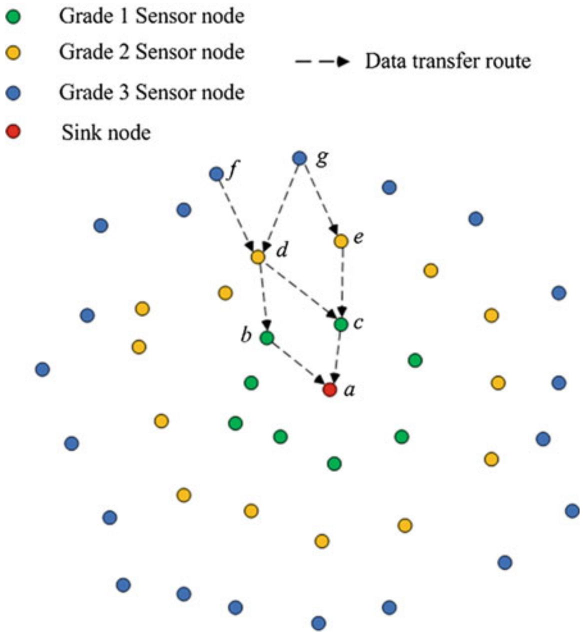
In this section, we propose an energy-aware routing protocol. Our design has two phases, ladder diffusion phase and cat swarm optimization phase.

Ladder Diffusion Phase

In our protocol, we utilize the ladder diffusion algorithm described in [3]. Here we use a simple example to explain how the ladder diffusion works. Take Fig. 1 as an example. The sink node *a* first broadcasts a ladder-creating package with the grade value of one. Sensor nodes which receive this package are identified as grade one. It means that these grade one sensor nodes transmit data to the sink node require only one hop. In Fig. 1, sensor nodes with green color are categorized into grade one. Then, grade one sensor nodes, for example, *b* and *c*, broadcast another ladder-creating package with the grade value of two. Sensor nodes that receive this package (colored yellow in Fig. 1) are categorized into grade two. These grade two sensor nodes send data to the sink node require two hops count. Similarly, sensor nodes colored blue are categorized into grade three. After finishing the above procedures, all sensor nodes have been classified into a grade.

When a sensor node desires to send data to the sink node, the route is dynamically created by starting with nodes of high grade value and ending with nodes of low grade value. Normally, each sensor node will have more than one route to the sink node. As an example shown in Fig. 1, sensor node *g* can transmit data to the sink node *a* through two candidate nodes *d* and *e*.

Fig. 1 An example of ladder diffusion algorithm



Cat Swarm Optimization Phase

Since each sensor node has more than one route path to the sink node, in this phase, we integrate the cat swarm optimization (CSO) to obtain the most suitable route path.

Fitness function. In our design, each cat means a routing path from the source node to the sink node. In order to evaluating the energy consumption of each cat, we define a fitness function F_m for Cat_m which is shown in Eq. (1).

$$\mathrm{F}_m = r(\alpha_1 + \alpha_2 \times D_\mathrm{m}) + E_m \tag{1}$$

In Eq. (1), D_m means the summation of the Euclidean distance between any two adjacent nodes for Cat_m; E_m denotes the summation of the remainder energy of sensor nodes for Cat_m; r is the data transfer rate of sensor nodes; α_1 is the distance-independent parameter; α_2 the distance-dependent parameter.

Initialization. Here we use Fig. 2 as an example to describe this phase. Assume that sensor node *tar* (grade six) desires to send a package to sink node *sink*. The configuration of the initialization parameters is as follows.

1. *tar* generates k cats. In this example, we assume that *tar* generates 10 cats. Among these 10 cats, 7 cats are in seeking mode and 3 cats are in tracing mode.
2. Each cat has its own position composed of 5 dimensions and each dimension has its velocity. In Fig. 2, the position of Cat_1 (the path colored red) is $(\chi_{1,1}, \chi_{1,2}, \chi_{1,3}, \chi_{1,4}, \chi_{1,5})$ and the initial velocity for these dimension are $(v_{1,1}, v_{1,2}, v_{1,3}, v_{1,4}, v_{1,5})$. Note that these initial velocities are randomly selected.
3. *tar* utilizes the fitness function (Eq. 1) to calculate the fitness value. Then, *tar* sets the best fitness value as the local best.

Seeking mode. In seeking mode, we first define three essential factors, seeking range of selected dimension (SRD), counts of dimension to change (CDC) and seeking memory pool (SMP). SRD and CDC denote the mutative ratio for the selected dimensions and how many dimensions will be varied respectively. Besides, SMP means the number of copies of a cat in seeking mode.

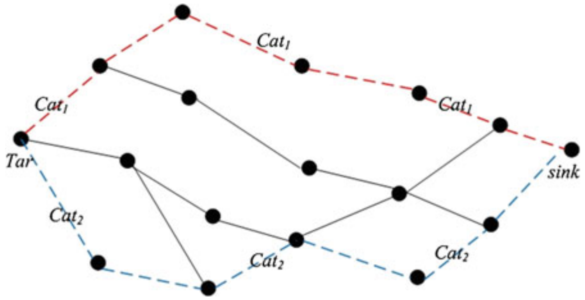


Fig. 2 Routing path from *tar* to *sink*

Assume Cat_2 in Fig. 3 is categorized into seeking mode. Besides, SRD, CDC and SMP are set 0.2, 0.4 and 6 respectively. The procedure of seeking mode is listed as follows

1. Make 6 copies of Cat_2, notes $cat_{2:1}$, $cat_{2:2}$, $cat_{2:3}$, $cat_{2:4}$, $cat_{2:5}$, $cat_{2:6}$.
2. For $Cat_{2:1}$, suppose the mutative nodes are b and d. According to SRD (20 %), b is varied to g and d is varied to s. Consequently, $Cat_{2:1}$ is varied from ($a; b; c; d; e$) to ($a; g; c; s; e$).

Similarly, other five copies of Cat_2 perform the same procedure as $Cat_{2:1}$.After that, tar calculates the fitness values of these 6 copies of Cat_2 through Eq.1 and then replaces the Cat_2 with the best fitness value copy cat.

Of course, other cats classified into seeking mode do the same procedure as Cat_2. Finally, we get 7 updated cats.

Tracing mode. Assume that Cat_1 shown in Fig. 3 is categorized into tracing mode. It first updates the velocities according to Eq. (2).

$$v_{1,d} = v_{1,d} + r_1 \times c_1 \times \left(x_{best,d} - x_{1,d}\right) \quad where\ d = 1, 2, \ldots, 5 \tag{2}$$

$x_{\text{best;d}}$ is the dth dimension position of Cat_1 who has the best fitness value; $x_{1;d}$ is the dth dimension position of Cat_1; c_1 is a constant and r_1 is a random value in the range of [0,1].

Then, $x_{1;d}$ is updated according to $v_{1;d}$ (see Eq. 3).

$$\mathrm{x}_{1,d} = x_{1,d} + v_{1,d} \quad where\ d = 1, 2, \ldots, 5 \tag{3}$$

Similarly, other two cats classified into tracing mode also move to a new place. Thus, we get three new cats.

Global Updating

tar evaluates the fitness values of these ten new cats and sets the cat of best fitness value as Cat_{11}. After that, it re-picks 3 cats and sets them into tracing mode and

Fig. 3 Seeking mode and tracking mode

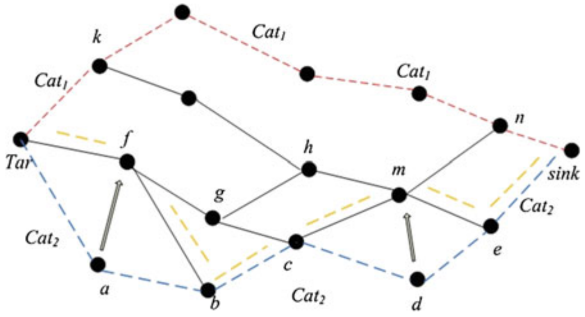
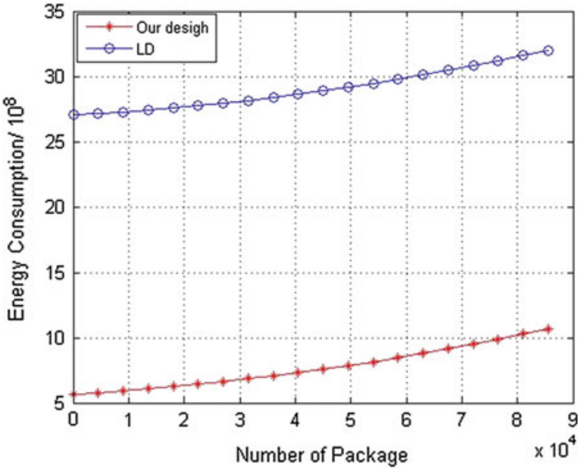


Table 1 Average time consumption

	Average time consumption (ms)
LD	25.313
Our design	10.683

Fig. 4 Energy consumption



sets the rest 7 cats into seeking mode. Then, *tar* performs the above procedures of seeking mode and tracing mode repeatedly. The final path would be the best one since CSO keeps the best path position till it reaches the end of iterations.

Simulations

The simulation results are shown in this section. We implemented LD [3] and our protocol and then compared the performance of these two protocols. In our simulation environment, 3,000 sensor nodes are deployed in an 100 × 100 × 100 unit of three-dimensional space. The transmission range of the sensor nodes are set to 15 units and the Euclidean distance is at least 2 units between any two sensor nodes. The sink node is distributed in the center of the area.

In our simulation, the simulator randomly selects one sensor node and calculates the execution time for this sensor node to find the most suitable routing path. The simulator performs this calculation 90,000 times and then obtains the average execution time. The results are shown in Table 1. Obviously, our work reduces the execution time by 75.84 %.

We also evaluated the energy consumption of our design and LD. The result is shown if Fig. 4. As shown in Fig. 4, the energy consumption of LD is about three times greater than our design.

Conclusion

In this paper, we present an energy-aware routing protocol based on ladder diffusion and cat swarm optimization. Similar to LD, our work can avoid the generation of circle route, reduce the energy consumption and provide back-up routes. Moreover, the proposed protocol further reduces the execution time and energy consumption compared with LD.

References

1. Perkins CE, Royer EM (1999) Ad-hoc on-demand distance vector routing. In: Proceedings of 2nd IEEE workshop on mobile computing systems and applications, pp 90–100
2. Intanagonwiwat C, Govindan R, Estrin D, Heidemann J, Silva F (2003) Directed diffusion for wireless sensor networking. IEEE/ACM Trans Network 11(1):2–16
3. Ho JH, Shih HC, Liao BY, Chu SC (2012) A ladder diffusion algorithm using ant colony optimization for wireless sensor networks. Inf Sci 192:204–212
4. Carballido JA, Ponzoni I, Brignole NB (2007) Cgd-ga: a graph-based genetic algorithm for sensor network design. Inf Sci 177(22):5091–5102
5. He S, Dai Y, Zhou R, Zhao S (2012) A clustering routing protocol for energy balance of wsn based on genetic clustering algorithm. IERI Procedia 2:788–793
6. Nayak P, Ramamurthy G, et al (2012) A novel approach to an energy aware routing protocol for mobile wsn: Qos provision. In: Proceedings of international conference on advances in computing and communications, IEEE, pp 38–41
7. Chen CM, Lin YH, Chen YH, Sun HM (2013) SASHIMI: secure aggregation via successively hierarchical inspecting of message integrity on WSN. J Inf Hiding Multimedia Signal Process 4(1):57–72
8. Chen CM, Lin YH, Lin YC, Sun HM (2012) RCDA: recoverable concealed data aggregation for data integrity in wireless sensor networks. IEEE Trans Parallel Distrib Syst 23(4):727–734
9. Chu SC, Huang HC, Shi Y, Wu SY, Shieh CS (2008) Genetic watermarking for zerotree-based applications. Circuits Syst Signal Process 27(2):171–182
10. Chu SC, Roddick JF, Pan JS (2004) Ant colony system with communication strategies. Inf Sci 167(1):63–76
11. Dorigo M, Gambardella LM (1997) Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans Evol Comput 1(1):53–66
12. Dorigo M, Maniezzo V, Colorni A (1996) Ant system: optimization by a colony of cooperating agents. IEEE Trans Syst Man Cybern B Cybern 26(1):29–41
13. Misra R, Mandal C (2006) Ant-aggregation: ant colony algorithm for optimal data aggregation in wireless sensor networks. In: In Proceedings of IFIP international conference on wireless and optical communications networks, IEEE, p. 5
14. Chu SC, Tsai PW, Pan JS (2006) Cat swarm optimization. In: PRICAI 2006: Trends in artificial intelligence, pp 854–858
15. Wang ZH, Chang CC, Li MC (2012) Optimizing least-significant-bit substitution using cat swarm optimization strategy. Inf Sci 192:98–108
16. Panda G, Pradhan PM, Majhi B (2011) Iir system identification using cat swarm optimization. Expert Syst Appl 38(10):12671–12683
17. Pradhan PM, Panda G (2012) Solving multi-objective problems using cat swarm optimization. Expert Syst Appl 39(3):2956–2964

Software Baseband Optimization Except Channel Decoding for PC-Based DVB-T Software Radio Receiver

Shu-Ming Tseng, Yao-Teng Hsu, Yen-Yu Chang and Tseng-Chun Lee

Abstract The software radio has the advantages of flexibility, low cost and multimode ability, but the major disadvantage is the slower speed. To increase the speed of PC-based software DVB-T receiver, we need implement the baseband signal processing algorithms much faster. In this thesis, we discuss faster implementation of (1) 16QAM de-mapper to bits, (2) expanding_channel coefficient value in Viterbi decoder, and (3) inner de-interleaver and depuncturer. The speed of these three blocks is 6.64x, 0.57x and 0.88x faster.

Keywords Software radio · DVB-T · SIMD

Introduction

Recently, PC-based software radio systems have been developed, such as software global position system (GPS) [1]. The software GPS reduce the cost and time to modify the hardware. Another PC-based software GPS is in [2]. The new

S.-M. Tseng (✉) · T.-C. Lee
Department of Electronic Engineering, National Taipei University of Technology,
Taipei 106, Taiwan, Republic of China
e-mail: shuming@ntut.edu.tw

T.-C. Lee
e-mail: jerryli0527@gmail.com

Y.-T. Hsu
Liteon Corp, Taipei, Taiwan, Republic of China
e-mail: Matthew.Hsu@liteon.com

Y.-Y. Chang
College of Electrical Engineering and Computer Science, National Taipei University
of Technology, Taipei 106, Taiwan, Republic of China
e-mail: yeychang6614@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_37, © Springer Science+Business Media Dordrecht 2014

acquisition and tracking method is used to increase operation speed and it can deal with the 12 channels computing requirement. A software digital audio broadcasting (DAB) receiver is implemented in [3]. It is implemented in a notebook PC platform. Besides, a PC-based multimode digital audio broadcasting system includes digital radio mondiale (DRM), DAB in Europe, and HD radio in United States is implemented in [4].

The PC-based software radio has the following advantage. It's easy to modify the signal processing algorithms to enhance the channel capacity and reduce the impact of multipath fading. In addition, we are more familiar with the PC platform than digital signal processing (DSP) or field programmable gate array (FPGA) platforms. Nowadays, the biggest challenge of the software radio is to process massive radio data fast enough. In order to make the speed of the software digital video broadcasting-terrestrial (DVB-T) receiver be fast enough to achieve real-time playing of TV programs, we need to optimize the time-consuming functional blocks of the software DVB-T receiver.

Currently, we have completed the optimization of Viterbi decoder [5], Reed Solomon (RS) decoder [6] and sampling frequency offset compensation [7]. To achieve real-time reception, we also propose a hard iterative channel decoding to enhance the error-correcting ability [8]. We still have to optimize the remaining functional blocks.

The single instruction multiple data stream (SIMD) instructions is often used in H.264 decoder [9] and wavelet transform [10], but not used in communication systems. Now we used SIMD instructions in the PC-based software DVB-T receiver. We also remove branches (IF command, for example), data re-arrangement, and rewrite the program by Assembly to improve performance.

The remaining of the paper is organized as follows: In section "Software and Hardware Environment", we describe the software and hardware environment. In section "Optimization of 16QAM De-mapper", we discuss the optimization of 16 quadrature amplitude modulation (16QAM) de-mapper. In "Optimization of Expanding_Channel Coefficient Value", we discuss the optimization of expanding_channel coefficient value in Viterbi decoder. In section "Optimization of Inner De-interleaver and Depuncturer", we discuss how to speed up the inner de-interleaver and depuncturer operations. And the conclusion of the paper is presented in section "Conclusion".

Software and Hardware Environment

We use SIMD instructions to improve performance, so we choose the operating system and CPU of using 128 bit registers. And the hardware platform is listed in Table 1. Our program takes 215.5 Mbytes in memory, so PC memory can be as low as 1 GB or 2 GB. We use the performance explorer in Microsoft Visual Studio 2010 to measure the performance (speed). The field test data comes from the DVB-T specification currently used in Taiwan (Parameters are shown in Table 2),

Table 1 Hardware list

Item	Model
CPU	Intel Core 2 Quad Q8200(2.33 GHz)
memory	DDR2 800 8 GB
main board	GIGABYTE GA-EP43-DS3L
Graphic card	NVIDIA GeForce 6200 TurboCache(TM)

Table 2 Parameter of Taiwan DVB-T

Arameter	Value
Frame	68 OFDM symbols
Super frame	4 frames
Mmodulation	16 QAM
Transmission modes	8 K
Subcarriers	6,817
Subcarriers (without pilots)	6,048
Guard interval(GI)	1/4
Code rate	2/3

we use a customized USB 2.0 dongle (including RF front end and A/D converter) to record data. Performance calculation is based on the requiring time of decoding 28 frames.

The functional blocks are shown in Fig. 1. In this paper, we discuss faster implementation of the following three blocks: (1) 16QAM de-mapper to bits, (2) expanding_channel coefficient value in Viterbi decoder, and (3) inner de-interleaver and depuncturer.

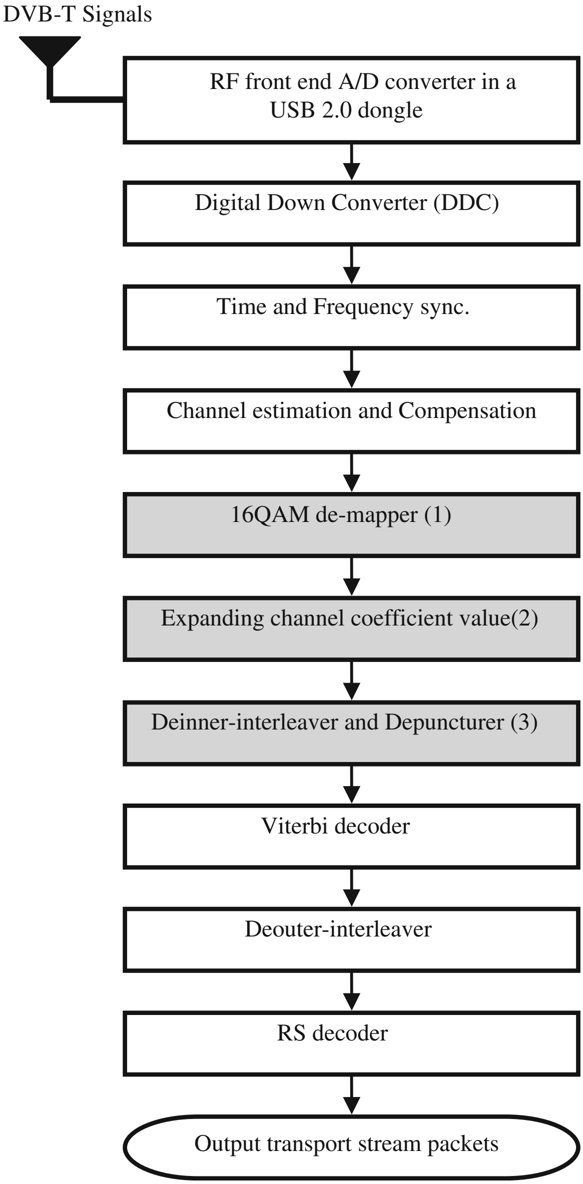
Optimization of 16QAM De-mapper

We consider 16QAM de-mapper (to 4 bits) in the non-hierarchical mode. For each DVB-T frame, we have 68 orthogonal frequency division multiplexing (OFDM) symbols and each symbol has 6,048 sub-carriers. There are 411,264 de-mapper operations for each frame. This computation is large and speedup is necessary.

For the 16 QAM in Fig. 2, the 4 soft bits ($i1$, $q1$, $i2$, $q2$) are as follows:

$$\begin{aligned} i1 &= -I \\ q1 &= -Q \\ i2 &= \begin{cases} 2D - I, I > 0 \\ 2D + I, I < 0 \end{cases} \\ q2 &= \begin{cases} 2D - Q, Q > 0 \\ 2D + Q, Q < 0 \end{cases} \end{aligned} \tag{1}$$

Fig. 1 The block diagram of the software DVB-T receiver


DVB-T Signals
RF front end A/D converter in a USB 2.0 dongle
Digital Down Converter (DDC)
Time and Frequency sync.
Channel estimation and Compensation
16QAM de-mapper (1)
Expanding channel coefficient value(2)
Deinner-interleaver and Depuncturer (3)
Viterbi decoder
Deouter-interleaver
RS decoder
Output transport stream packets

Where I is the magnitude of the in-phase carrier. Q is the magnitude of the quadrature phase carrier. $D = 1$ and is shown in ETSI EN 300 744 [11].

For i2 and q2, we need to check if the input value (I, Q) is positive or negative. The old scheme to do so is shown in Fig. 2a. It can be seen that we have a branch (IF) in this scheme. We use two steps to improve the performance of 16QAM de-mapper, as shown in Fig. 2b. First, we use SIMD instruction set of packed

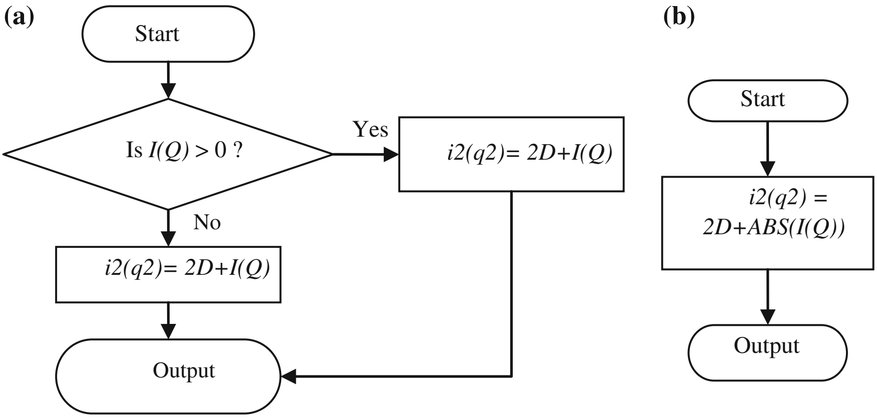
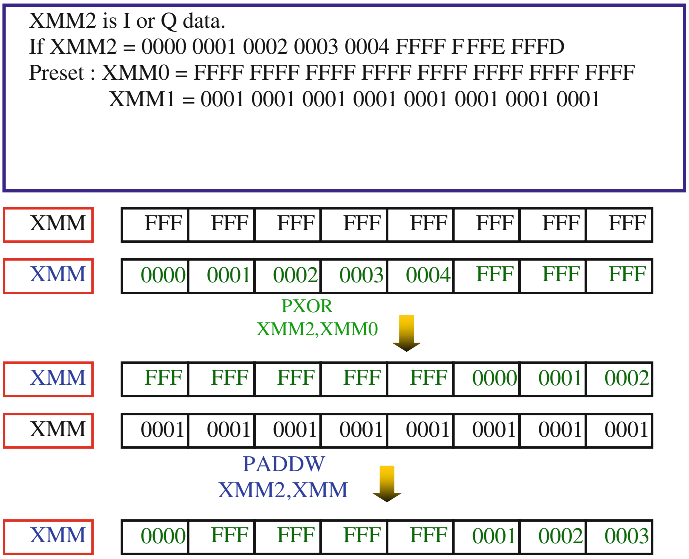


Fig. 2 De-mapper for i2 or q2 (**a**) the old scheme and (**b**) the new scheme

absolute words (PABSW) to replace the part of using the branch and also achieve the advantage of SIMD. Due to multiplication on the computer is more time-consuming, we use logical exclusive OR (PXOR) and packed add words (PADDW) to replace negate (NEG) operations, as shown in Fig. 3. The performance of de-mapper (C version), de-mapper_asm (old), and de-mapper_asm (new) are 277, 204, and 26.7 ms, respectively. The performance improvement is 6.64x


XMM2 is I or Q data.
If XMM2 = 0000 0001 0002 0003 0004 FFFF FFFE FFFD
Preset : XMM0 = FFFF FFFF FFFF FFFF FFFF FFFF FFFF FFFF
XMM1 = 0001 0001 0001 0001 0001 0001 0001 0001

XMM	FFF	FFF	FFF	FFF	FFF	FFF	FFF	FFF
XMM	0000	0001	0002	0003	0004	FFF	FFF	FFF
XMM	FFF	FFF	FFF	FFF	FFF	0000	0001	0002
XMM	0001	0001	0001	0001	0001	0001	0001	0001
XMM	0000	FFF	FFF	FFF	FFF	0001	0002	0003

Fig. 3 De-mapper for i1 or q1

Optimization of Expanding_channel Coefficient Value

In Viterbi decoding of DVB-T, we use the maximum likelihood (ML) algorithm. The log likelihood function is given by:

$$p(r|x) = \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-|r-hx|^2}{2\sigma^2}} \tag{2}$$

We assume h is the channel coefficient, x is the transmitted data, and r is the received data.

Where $|r - hx|^2$ is the metric and can be expressed as follows:

$$|r - hx|^2 = |h|^2 \left| \frac{r}{h} - x \right|^2 = Q \left| \frac{r}{h} - x \right|^2 = metric \tag{3}$$

where $Q = |h|^2$.

The Q value must be processed in the de-mapper, In 16-QAM, non-hierarchical modulation requires de-mapper to 4 bits for Viterbi decoding. So we copy Q value fore times, as shown in Fig. 4. The new scheme uses unpack low packed data (PUNPCKLWD) SIMD instruction twice to speed up this operation, as shown in Fig. 5. The performance of expanding_Q_Value (C version), expanding_Q_Value_asm (old),and expanding_Q_Value_asm (new) are 119, 35.25, and 22.42 ms, resoectively. The performance is up to 0.57x.

Fig. 4 The method of xpanding Q value (the old scheme)

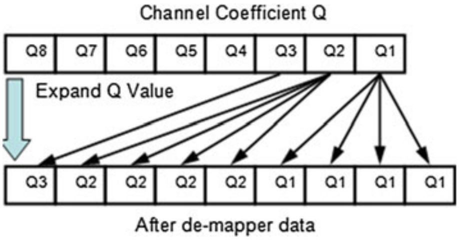
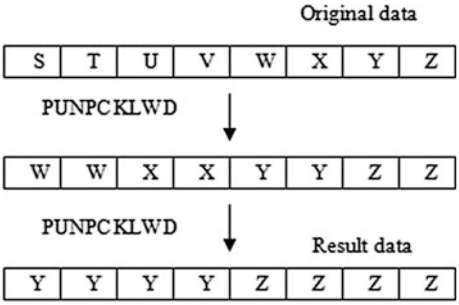


Fig. 5 The method of expanding Q value (the new scheme)



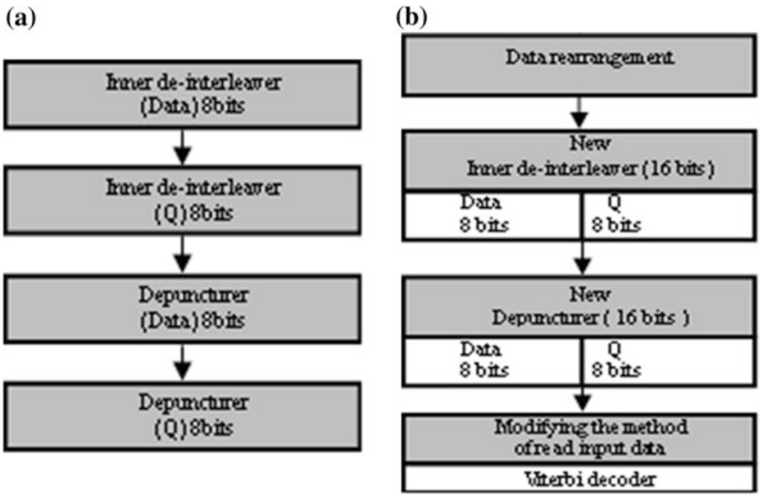

(a)
Inner de-interleaver (Data) 8bits
Inner de-interleaver (Q) 8bits
Depuncturer (Data) 8bits
Depuncturer (Q) 8bits
(b)
Data rearrangement
New Inner de-interleaver (16 bits)
Data 8 bits
Q 8 bits
New Depuncturer (16 bits)
Data 8 bits
Q 8 bits
Modifying the method of read input data
Viterbi decoder

Fig. 6 **a** The old scheme of inner de-interleaver and depuncturer on software DVB-T receiver. **b** The new scheme of inner de-interleaver and depuncturer on software DVB-T receiver

Optimization of Inner De-interleaver and Depuncturer

In the old scheme, we execute inner de-interleaver and depuncturer twice, as shown in Fig. 6a. One for *Q* value, and the other for data *(r/h)*. In Fig. 6, we show the new scheme. In Fig. 7, it shows that the arrangement of the old data format. In Fig. 8, it shows that we rearrange the data format from 8 to 16 bit. In the end, we modify the method of Viterbi to read input-data. Instead of processing date (8 bits) and *Q* value (8 bits) individually, we process them together (16 bits) in the inner

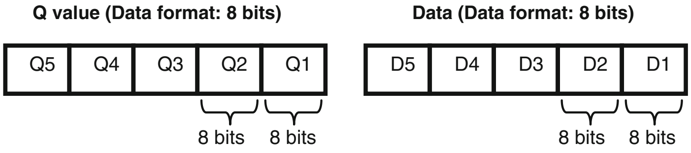


Fig. 7 Data arrange (the old scheme)

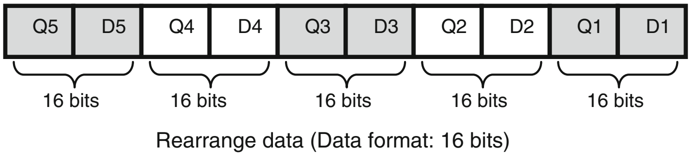
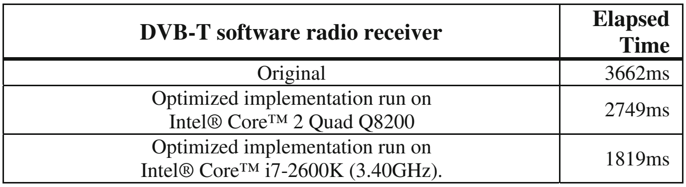


Fig. 8 The performance comparison between original method and reduction in operating time



DVB-T software radio receiver	Elapsed Time
Original	3662ms
Optimized implementation run on Intel® Core™ 2 Quad Q8200	2749ms
Optimized implementation run on Intel® Core™ i7-2600K (3.40GHz).	1819ms

Fig. 9 The elapsed time of the whole DVB-T software radio receiver (including all blocks including channel decoding) for 2,844 ms input data

de-interleaver and depuncturer operation. In Fig. 14, The performance of inner de-interleaver and depuncturer (C Version), inner de-interleaver and depuncturer asm (old), inner de-interleaver and depuncturer asm (new) are 767, 370, 197 ms, resoectively. The performance is up to 0.88x.

Conclusion

We use SIMD instruction set, data rearrangement, etc. to speed up the three functional blocks of software DVB-T receiver. The 16QAM de-mapper, expanding channel coefficient value, and inner de-interleaver and depuncturer is 6.64, 0.57 and 0.88x faster. The quantity of input data is 28 frames which are equal to 2,844 ms. Our software DVB-T receiver takes 2,749 ms to process the input data. Thus, the overall optimization of baseband signal processing of PC-based software DVB-T receiver is now complete and achieve real-time reception. In addition, if we use the latest CPU and the performance can reach to 1,819 ms. The overall performance improvement including channel decoding is shown in Fig. 9.

Acknowledgments This work was supported in part by National Science Council, Taiwan, under grant NSC 96-2622-E-027-015-CC3.

References

1. Kubo N, Kondo S, Yasuda A (2005) Evaluation of code multipath mitigation using a software GPS receiver. IEICE Trans Commun E88-B(11): 4204–4211 (2005)
2. Li S, Zhai C, Zhan X, Wang B (2009) Implement of real-time GPS L1 software receiver. In: Proceedings of the 4th international conference on computer science & education (ICCSE ‘09), pp 1132–1137
3. Tseng SM, Hsu YT, Chang MC, Chan HL (206) A notebook PC based real-time software radio DAB receiver. IEICE Trans Commun E89B(12):3208–3214
4. Di N, Gao P, Wan G, Li J, Li J (2010) A common SDR platform for digital audio broadcasting system. In: Proceedings of the 3rd international congress on image and signal processing (CISP 2010), vol. 8, pp. 3708–3711

5. Tseng S-M, Kuo Y-C, Ku Y-C Hsu Y-T (2009) Software viterbi decoder with SSE4 parallel processing instructions for software DVB-T receiver. In: Proceedings of the 7th IEEE international symposium on parallel and distributed processing with applications (ISPA-09), Chengdu and Jiuzhai Valley, China, pp 102–105
6. Tseng S-M, Hsu Y-T, Shih J-Z (2012) Reed-solomon decoder optimization for PC-based DVB-T software radio receiver. Inf Int Inter J 15(8):3485–3498
7. Tseng S-M, Yu J-C, Hsu Y-T (2011) A real-time PC based software radio DVB-T receiver. In: Proceedings of the 3rd international conference on future computational technologies and applications (FUTURE COMPUTING 2011), Rome, Italy, pp 86–91
8. Tseng S-M, Hsu Y-T, Lin H-K (2013) Iterative channel decoding for PC-based software radio DVB-T receiver. Wirel Personal Commun 69(1):403–411
9. Asif M, Farooq M, Taj IA (2010) Optimized implementation of motion compensation for H.264 decoder. In: Proceedings of 5th international conference on computer sciences and convergence information technology (ICCIT), pp 216–221
10. Shahbahrami A, Juurlink B, Vassiliadis S (208) Implementing the 2-D wavelet transform on SIMD-enhanced general-purpose processors. IEEE Trans Multimedia 10(1):43–51
11. Digital Vide Broadcasting (DVB) (2009) Framing structure, channel coding and modulation for digital terrestrial television, ETSI EN 300 744 V1.6.1, pp 22

In-Time Transaction Accelerator Architecture for RDBMS

Su Jin Kim, Seong Mo Lee, Ji Hoon Jang, Yeong Seob Jeong, Sang Don Kim and Seung Eun Lee

Abstract In this paper, we propose a hardware architecture for in-time transaction accelerator that reduces the bottlenecks between the DB server and the DB storage in margin FX trading system's RDBMS (Relational Database Management System). In-time transaction accelerator located between the DB server and the DB storage analyzes and processes the queries used for margin FX trading system by co-processing of the CPU and the FPGA. The accelerator analyzes the patterns and the consistency of the queries to reduce the total database access in order to increase the RDBMS's throughput.

Keywords In-time transaction accelerator · Margin FX trading · RDBMS · FPGA

Introduction

The field of stock markets has heavy traffic volume. Especially, a foreign exchange margin trading (a.k.a. margin FX trading) market that is the buying and selling of currencies generates much heavier traffic volume than the general stock market does. Furthermore, the traffic volume is growing up. However, there are limitations to improve the performance of the existing database system. The reason is that the obstacle named power wall blocks to improve the performance of the CPU and the memory access time is still markedly slow. When financial traffic volume is increasing, bottlenecks can occur between the server and the storage. To resolve

S. J. Kim · S. M. Lee · J. H. Jang · Y. S. Jeong · S. D. Kim · S. E. Lee (✉)
Department of Electronic and Information Engineering, Seoul National University of Science and Technology, 172 Gongreung-2-dong, Nowon-gu, Seoul-si, Korea
e-mail: seung.lee@seoultech.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_38, © Springer Science+Business Media Dordrecht 2014

the bottlenecks of DBMS, two groups of studies have been conducted. The first study is a new software architecture like NoSQL-based database such as the MongoDB [1] and database in memory such as the SAP HANA [2]. The second study is a new hardware architecture [3].

In this paper, we propose a hardware architecture for in-time transaction accelerator to reduce the bottlenecks that occur between the server and the storage. Our basic idea is using the Hardware parallelism by using the Field-Programmable Gate Array (FPGA). The FPGA is a possible parallel access and this feature facilitates the processing of streaming data [4].

In-time transaction accelerator architecture is located between the server and the storage. The accelerator is composed of CPU, FPGA, SSD and volatile memory. In order to have high-speed parallel access, we use the peripheral component interconnect (PCI) express interface between the CPU and the FPGA. Financial data (SQL query) is sequentially stored in the transaction accelerator. Then, the in-time transaction accelerator analyzes the *n*-query and optimizes the query to minimize the number of access to DB storage. Through this method, we can improve performance by extending the throughput between the DB server and DB storage.

We are working with the financial data from a financial solution company managing the margin fx trading system.

The rest of this paper is organized as follows. In section "Related Work", we review work related to improvement of the system by using FPGA. Section "RDBMS with In-Time Transaction Accelerator" describes the in-time transaction accelerator architecture. We conclude in section "Summary" by proposing the future works our architecture.

Related Work

The clock rate that is increasing at a high-speed is disturbed by the power wall. Hence, recently the researchers have been focusing on the architectural solutions to improve the CPU's performance. In [5], they proposed the co-processing of the CPU and the FPGA to improve the CPU's data processing performance. The performance is improved compared that of the original by using the FPGA's hardware parallelism. Furthermore, the system using a co-processing through the process where the system uses compressed the query along with the FPGA's decompression and process of the query increases the throughput by reducing the CPU usage about analyzing the query operation [6]. Our in-time transaction accelerator uses a co-processing of the CPU and the FPGA to process the data faster. Another approach uses a characteristic of the FPGA that is reconfigurable. When the queries are executed, the FPGA are reconfigured to the optimized design to process the data faster [7]. But the overhead of the configuration reduces the performance of the system. In the Netezza's FAST-engine, the FPGA fabrics are fixed and the internal registers of the FPGA are changed for the partial reconfiguration [8]. The Netezza's FAST-engine reduces the traffic volume between the memory and the CPU by using an engine's pipeline stage if the data pattern has a consistency. In our research, we hold a FPGA's design and use the [7]'s method to exclude a overhead when reconfiguration is needed.

RDBMS with In-Time Transaction Accelerator

Overview

The basic principle in order to improve the throughput of the RDBMS in in-time transaction accelerator architecture we proposed is to the commonalities between the requested queries and the pattern of the certain financial data. Therefore, we transfer the minimized queries to DB storage. As a result, we can reduce the DB access required to read and write. In-time transaction accelerator located between the DB server and the DB storage receives the queries and minimizes the query traffic volume. Minimizing the query traffic volume realizes CPU and co-processing by using the trait of parallelism of the FPGA. We explain detailed information in section "Query Analyzer". By reducing the number of the query to DB storage, we reduce the number of access to DB storage. As a result, we can improve the throughput of the RDBMS.

In-time transaction accelerator architecture is composed of the Query Analyzer (minimizing the query traffic volume), PCIe module (in order to parallel access of the FPGA) and Log buffer(saving the received queries). Figure 1 shows the RDBMS adding the in-time transaction accelerator.

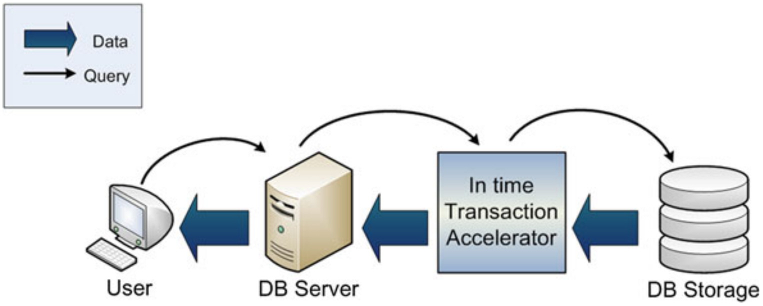


Fig. 1 RDBMS with in-time transaction accelerator

Query Analyzer

The Query Analyzer analyzes the pattern and consistency of the queries related the margin FX trading that is received sequentially from PCI Express and reduces the amount of query requesting the memory access to DB storage (Table 1).

For example, when above the query comes into Query Analyzer through the PCIe communication, the Query Analyzer analyzes the (1) and (2). The query (1) that requires the data can be realized through the query (2) that requires more data. Thus the query (2) is transferred to DB storage. In the case of the above example, the number of the query is reduced. In this way, the Query Analyzer reduces the number of access to DB storage. We reduce the number of the time-consuming memory access. We make the system to conduct to same work in lesser amount of time. In this way, we improve the entire throughput of the RDBMS.

PCIe Module

The PCI Express can realize the bandwidth from 200 MB/s to 6.4 GB/s as by serial bus system. The possible reason of the high bandwidth is that two PCI Express communication device can use a multi-lane link. A Lane an independent link, where can happen the single transmission at PCI Express, has an individual operating frequency. The multi-lane link is a large link comprised of the several independent Lane. Although the lane is serial bus, if multi-lane link consists the several Lane, it will transmit the data in the parallel form and be able to communicate with high bandwidth by operating frequency of an independent lane. The bidirectional data bandwidth is about 200 MB/s in a PCI Express 1× and the bidirectional data bandwidth is about 6.4 GB/s in a PCI Express 32x. A PCI module is the module to use the PCI Express which is based on high bandwidth. The data transfer system of the parallel form of PCI Express is made using the FPGA's hardware parallelism and supports the high speed transcipient from the CPU to the Query Analyzer.

Log Buffer

Log Buffer is responsible for storing the received queries to Solid State Disk (SSD). The Log buffer is designed to improve the safety of the financial system.

Table 1 Defined sample queries for explanation

select * from emp where ename like 'A %'; //Which starting 'A' -(1)
select * from emp where ename like ' %A %'; //Which includes 'A' -(2)

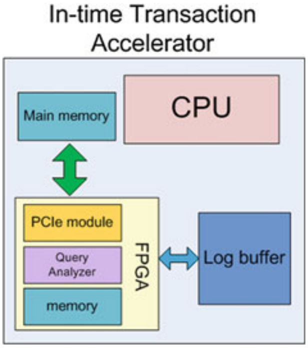


Fig. 2 In-time transaction accelerator

Uncompleted queries can disappear in unexpected situations (for example, system failure due to a disaster or outage and the system's internal computational error, etc.) during financial transaction. To prevent this, the queries coming into the FPGA are saved in SSD and are reloaded when we need it. Figure 2 shows architecture of the in-time transaction accelerator.

Summary

In this paper, we proposed the hardware architecture for in-time transaction accelerator that reduces the bottlenecks between the DB server and the DB storage in the RDBMS of margin FX trading system. PCIe interface is used for FPGA's hardware parallelism. The DB server transmits the queries to in-time transaction accelerator and then the CPU of the accelerator transmits the queries to the Query Analyzer of the FPGA through PCIe interface. The Query Analyzer analyzes and processes the queries received in a serial order. In consequence of the previous process, the primary queries are selected and are transmitted to the CPU with PCIe interface. And then the CPU transmits the primary queries received from the Query Analyzer to the DB storage. As a result, transmitting the primary queries can reduce the number of a memory access which takes a long time. This leads to improvement of RDBMS' throughput. Thus, performance of the RDBMS is enhanced. DB storage transmits the data demanded by the queries to the accelerator, and then process running on the CPU of accelerator transmits the data to all DB server that requests the data in the first step. Our In-time transaction accelerator architecture is appropriate for margin FX trading system's RDBMS as well as the big data systems that include a surge data stream. In the future, if the queries or the data are transmitted with InfiniBand which has wide Bandwidth from DB server to accelerator or from accelerator to DB storage on proposed our architecture, performance of in-time transaction accelerator architecture will be improved further.

Acknowledgments This study was supported in part by the IT R&D program of MKE/KEIT [10043896, Development of virtual memory system on multi-server and application software to provide real-time processing of exponential transaction and high availability service] and the Seoul National University of Science and Technology.

References

1. MongoDB, http://www.mongodb.org
2. SAP HANA, http://sap.com/HANA/
3. Nie C (2012) An FPGA-based smart database storage engine. Master's thesis, ETH zurich
4. Guha R, Al-Dabass D (2010) Performance prediction of parallel computation of streaming applications on FPGA platform. In: 12th international conference on computer modeling and simulation, UKSim, Cambridge pp 579–585
5. Mueller R, Teubner JM, Alonso G (2009) Data processing on FPGAs. J Proc VLDB Endowment, pp 910–921
6. Sukhwani B, Min H, Thoennes M, Dube P, Iyer B, Brezzo B, Dillenberger D, Asaad S (2012) Database analytics acceleration using FPGAs. In: Proceedings of the 21st international conference on parallel architectures and compilation techniques, ACM New York pp 411–420
7. Mueller R, Teubner J, Alonso G (2010) Glacier: a query-to-hardware compiler. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data, ACM New York pp 1159–1162
8. Francisco P (2011) The Netezza data appliance architecture: a platform for high performance data warehousing and analytics. Technical Report, IBM

Intra-Body Communication for Personal Area Network

Sang Don Kim, Ju Seong Lee, Yeong Seob Jeong, Ji Hoon Jang and Seung Eun Lee

Abstract The intra-body communication uses the human body as a conducting wire, providing the simplicity and the security. Although the communication distance is limited within a body-area, it is useful on the construction of personal area network. In this paper, we introduce our prototype intra-body communication module using the FPGA. The proposed system has the FSK modulator and the demodulator. These modulation methods are chosen after body-channel analysis. The experimental results demonstrate the feasibility of our intra-body communication module for establishing the PAN (Personal Area Network).

Keywords Intra-body communication · PAN (personal area network) · FPGA

Introduction

The intra-body communication uses the human body as a conducting wire in order to conduct the signals. The intra-body communication has a higher security than conventional wireless communication like as a wire communication. Furthermore, the intra-body communication provides convenience thanks to its wireless connectivity. However, the communication distance is limited within body area.

Modern handheld devices have high performance and contain many features such as telephone conversation, multimedia contents player, and data communications. The handheld devices have the PAN composed of various wireless communications for efficiently usage of these features. The PAN using the radio frequency has a risk of wiretapping and has a possibility to collide with other PAN in crowded region. Therefore, the PAN with higher security is required in recent

S. D. Kim · J. S. Lee · Y. S. Jeong · J. H. Jang · S. E. Lee (✉)
Department of Information Engineering, Seoul National University of Science and Technology, 172 Gongreung-2-dong, Nowon-gu, Seoul-si, Korea
e-mail: seung.lee@seoultech.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_39, © Springer Science+Business Media Dordrecht 2014

applications. Intra-body communication uses the human body instead of the radio frequency to transmit the electric signals thus it is free from the wiretapping and the radio interference. In addition, it has the simplicity because it does not require a conducting wire.

The intra-body communication was proposed in 1995 by Zimmerman [1]. He conducted various studies to analyze appropriate signals that can pass the human body. Recently, Schenk et al. have produced a prototyping that uses a human body as communication channel [2]. Joonsung Bae et al. have proposed about body channel communication from the surface of the human body by analyzing flow of signals in each of frequency band [3]. In this paper, we present the prototype of intra-body communication module to provide the PAN for handled devices.

The rest of this paper is organized as follows. We first briefly introduce the intra-body communication in section "Intra-Body Communication" and present the hardware features in section "Implementation". Section "Experimental Results" shows the experimental results on the data transmission. In section "Conclusion", we conclude this paper and introduce expected applications in the area of the intra-body communication module.

Intra-Body Communication

The intra-body communication is emerging as effective method which possesses the security and simplicity of the wireless communication. The human body has the conductivity because it includes water and some of electrolyte. The ECG is an example of the intra-body has the conductivity and the electric signals are conducted along the intra-body. Therefore, the external electric signal can be transmitted through the human body establishing communication channel. The weak biological signals are ignored or filtered. It is well known for the intra-body communication is not dangerous because the electric current passing through the human body is very low (about a milliampere).

Implementation

Human Body Channel Characterization

The signals to be transmitted should be modulated properly at a transmitter side for effective communication. A received signal from the human body is demodulated and data is reconstructed at the receiver module. In this case, the human body can be considered by one wire and provides the electric signal path from transmitter to receiver.

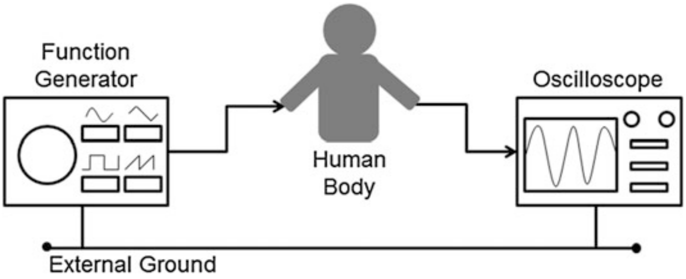


Fig. 1 Experimental environment of the signal transmission

Fig. 2 The output signal of a sine wave through the human body

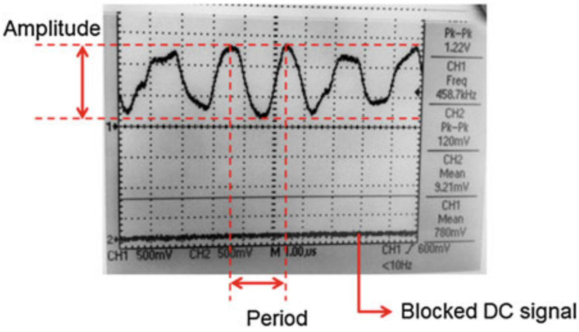


Figure 1 shows the experimental setup for human body channel characterization in order to decide the proper modulator scheme for the intra-body communication. Figure 2 illustrates the received waveform on oscilloscope. The DC electric signal is blocked and the AC electric signal is distorted when they pass through a human body. Therefore, the data have to be modulated into the AC signal. The modulator such as the ASK (Amplitude Shift Keying), the PSK (Phase Shift Keying) or the FSK (Frequency Shift Keying) can be applied to the data. The ASK is unsuitable for the intra-body communication because the distorted AC signal is not consistent at amplitude. The PSK has a more complex hardware than FSK and it may make an error when the transmitted signal is critically distorted. On the other hand, frequency at FSK is almost consistent although the signal was distorted thus we adopt FSK modulation in our intra-body connector.

The System Architecture

Our system consists of a transmitter and a receiver (See Fig. 3). The signal serializer in the transmitter block receives a parallel 8-bit data and serializes the data for serial communication. The modulator completes FSK modulation for intra-body communication. The modulated data is transmitted through the human body.

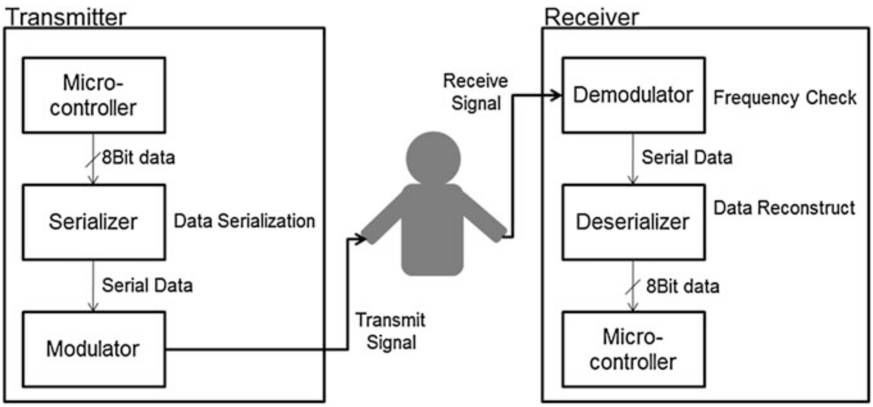


Fig. 3 The block diagram of a basic intra-body communication module

The demodulator in the receiver module checks the frequency of the received signal and reconstructs the data. A microcontroller at the transmitter side feeds the data stream should be transmitted through the intra-body communication and a microcontroller at the receiver side completes error rate.

Experimental Results

In order to verify the whole system, the intra-body communication module requires a communication protocol. We selected RS-232C protocol for serial communication. The data contains start bit and stop bit additionally. A byte data is transmitted through human body at a speed of 9,600 bps. We adopted Altera FPGA for realization of our proposed system. The electric signals which are modulated at transmitter are forwarded to the receiver through the human body. The restored data can be checked whether it is identical with transmitted data or not. An audio data was successfully transmitted to the receiver through the human body, demonstrating the feasibility of our proposal that the intra-body communication provides secure and simple wireless connection for the PAN (Fig. 4).

Conclusion

The intra-body communication has a simplicity and higher security compared with the conventional wire and wireless communication. In this paper, we presented the system of the intra-body communication by using the FSK modulation. The proper frequency band of conducting along the human body is selected by human body channel characterization. The audio data was successfully transmitted to the

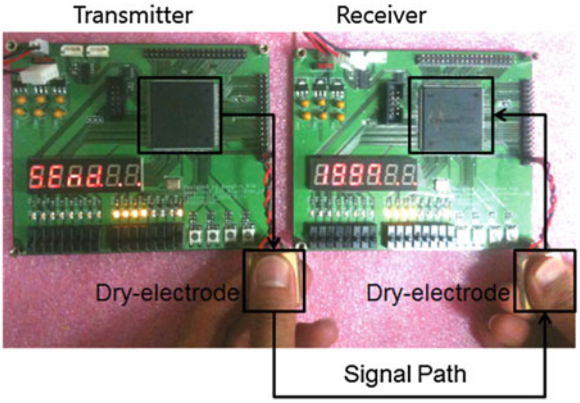


Fig. 4 The experimental environment of the data transfer

receiver through the human body. Experimental results demonstrated the feasibility of our proposal that the intra-body communication provides secure and simple wireless connection for the PAN.

In the applications such as financial services, where the higher security is required, the conventional wireless communications have the risk of the wiretapping. However, the intra-body communication can reduce the risk of the wiretapping in addition to can prevent other possible attacks. The safety can be improved by applying an encryption method. We expect the intra-body communication can be applied in various fields due to the simplicity and the safety.

Acknowledgments This work was supported by Seoul National University of Science and Technology, Korea.

References

1. Zimmerman TG (1995) Personal Area Networks (PAN): Near-field intra-body communication. Master's thesis, MIT, Cambridge, MA
2. Schenk TW, Mazloum NS, Tan L, Rutten P (2008) Experimental characterization of the body-coupled communications channel. Proc IEEE Int Symp Wearable Comput 234–239
3. Bae J, Cho H, Song K, Lee H, Yoo H-J (2012) The signal transmission mechanism on the surface of human body for body channel communication. IEEE Trans Microwave Theor Techn 60(3)

mrGlove: FPGA-Based Data Glove for Heterogeneous Devices

Seong Mo Lee, Ji Hoon Jang, Dae Young Park, Sang Don Kim, Ju Seong Lee, Seon Kyeong Kim and Seung Eun Lee

Abstract In this paper, we propose a glove based equipment (mrGlove) for a user interface that controls a device through hand motion recognition. The mrGlove provides more user experience compared to conventional devices such as a keyboard and a touchscreen. Our mrGlove is able to control a heterogeneous device (Windows or Android) that offers more convenience and user experience. Experimental results prove the feasibility of our proposal for enhancing convenience and user experience by obtaining control of an object through motion recognition in the PC and the smartphone.

Keywords Data glove · Heterogeneous device · Motion recognition · FPGA

Introduction

Over the past twenty years, many researches about a data glove have been done in entertainment field such as Mattel Power Glove or VPL Data Glove [1]. The glove-based controller controls a device through hand motion recognition. Control through motion recognition is intuitive and provides more user experience than a keyboard, mouse or touchscreen. Recently, controllers by using motion recognition such as Nintendo Wii [2] or Microsoft Kinect [3] have been widely used in entertainment field. Typically, motion recognition is classified into two groups. One is vision processing method and the other is the way to use mechanical equipment such as data glove. Data glove based motion recognition has the precise classification capability and fast response time compared to the vision processing method. Hence, it has been chosen for many systems [4].

S. M. Lee · J. H. Jang · D. Y. Park · S. D. Kim · J. S. Lee · S. K. Kim · S. E. Lee (✉)
Department of Electronic and Information Engineering, Seoul National University of Science and Technology, 172 Gongreung-2-dong, Nowon-gu, Seoul-si, Korea
e-mail: seung.lee@seoultech.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_40, © Springer Science+Business Media Dordrecht 2014

In this paper, we propose FPGA-based data glove, motion recognition Glove (mrGlove), to enhance user experience and convenience compared to a keyboard and a touchscreen. The mrGlove is able to control a heterogeneous device. Experimental results prove the feasibility of our proposal for enhancing convenience and user experience by obtaining control of an object through motion recognition.

The rest of this paper is organized as following: Section mrGlove describes the mrGlove in detail. Section Experimental Results explains the experimental results. We conclude in section Conclusion by proposing the future works on the mrGlove.

mrGlove

The mrGlove captures the user's motion and transmits data to the PC or the smartphone through Bluetooth communication channel. Software in the PC or the smartphone realizes the motion and controls an application.

Our mrGlove uses flex sensors and acceleration sensors to recognize user's hand motion. Flex sensors and switches are located in each finger, and signals from flex sensors are transmitted to FPGA for signal processing. The switches located on each fingertip enable a flex sensor to transmit a data to FPGA when they are turned on. When flex sensor is bended, resistance is changed. Acceleration sensor is used to detect user's hand motion. Signals from the acceleration sensor are transmitted to FPGA. Signals from the flex sensors and acceleration sensor are aggregated and processed in FPGA. Terasic DE0-nano board [5] including cyclone IV FPGA, ADC and digital accelerometer is used for implementation of the mrGlove. Figure 1 shows a prototype of the mrGlove.

A module for motion recognition in FPGA is composed of a data capture, data processing and a data transmit units. The data capture unit captures the signal from sensors. The data processing unit is used for data processing and the transmit unit

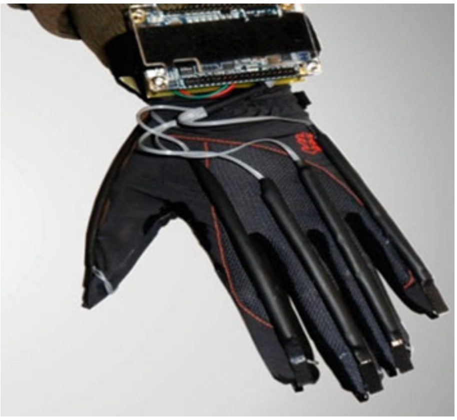


Fig. 1 Prototype of the mrGlove

forwards the data to Bluetooth module to communicate with a device. Figure 2 shows a system flow of the mrGlove.

Data Capture Unit. The data capture unit converts the analog flex signals into digital signals and brings the data from digital accelerometer by SPI (Serial to Peripheral Interface) communication. Calibration of resolution is possible in the analog-to-digital operation. The data capture unit includes the SPI controller to communicate with the digital accelerometer.

Data Processing Unit. The data processing unit sets a threshold of the data and assembles the data into a particular format for motion recognition. Figure 3 shows the data format for motion recognition. Basically, the data length is 8-bit for the RS232 communication standard. We determine five actions such as index finger, middle finger, balls of the feet, heel and tilt of the hand. We use only upper 5bits of the data according to five actions. When user generates each event, each bit corresponding to it is set high. For example, when an event corresponding to index finger is occur, the most significant bit of the data is set high.

Data Transmit Unit. The data transmit unit forwards the data from the data processing unit to the Bluetooth module. RS232 communication is used between the data transmit unit and the Bluetooth module. The data transmit unit includes the RS232 controller to communicate with the Bluetooth module. Calibration of a bps (bit per second) is possible.

Experimental Results

We demonstrate the motion recognition using the mrGlove on heterogeneous devices (PC and smartphone). The PC uses a Windows operating system and the smartphone uses an Android.

mrGlove with a PC

In order to demonstrate the mrGlove on the PC using Windows, we designed a middleware by using a LabVIEW [6]. The mrGlove communicates with the PC

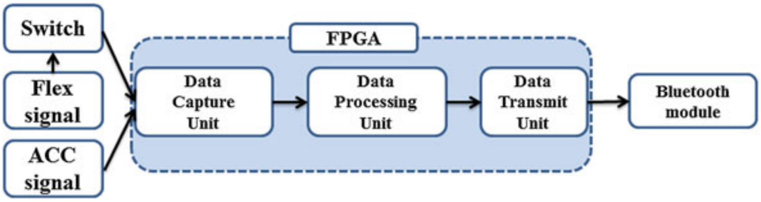


Fig. 2 System flow of the mrGlove

Bit	7	6	5	4	3	2	1	0
	Index finger	Middle finger	Balls of the feet	Heel	Tilt of the hand	-	-	-

Fig. 3 The data format for motion recognition

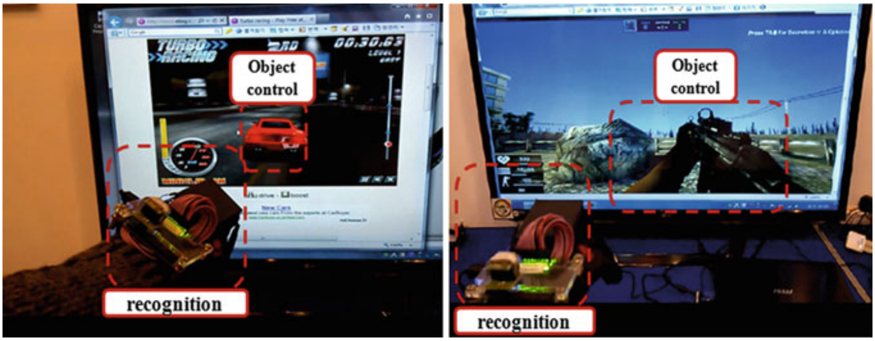
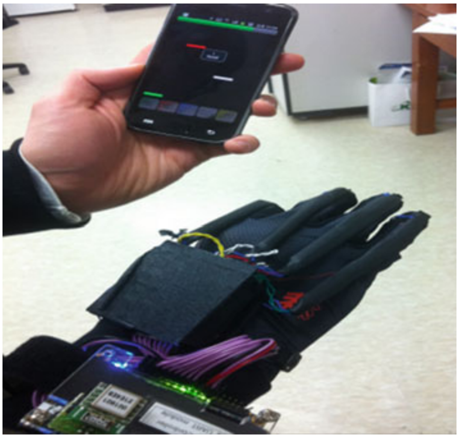


Fig. 4 A demonstration of motion recognition on the PC

Fig. 5 A demonstration of motion recognition on the smartphone



through a Bluetooth communication channel. The middleware processes the data and controls the PC as a keyboard does by using a Win32 library. Figure 4 shows the demonstration on the PC. The PC recognizes the hand motion through the mrGlove and controls the object in the racing game and the FPS (First-person shooter) game. Our mrGlove provides the way to control the most games with the same way.

mrGlove with a Smartphone

We designed an application (rhythm game) where five bar-shaped rectangles move down from top to bottom when music is selected. We tap the rectangles to the rhythm with the mrGlove. The application includes a class about a Bluetooth and a touch event. Figure 5 shows the touch event by using a switch of the mrGlove.

Conclusion

In this paper, we proposed a glove-based device for user interface that controls a heterogeneous device through hand motion recognition. We have succeeded in controlling the PC and a smartphone application. The mrGlove is able to control a smart device such as a tablet that has Bluetooth communication channel. We plan to add a feedback component in the mrGlove such as a vibration motor for a physical interaction game. The mrGlove is designed using a FPGA. Therefore, the units in the FPGA are easily converted to a single chip, minimizing the size of the mrGlove. We expect that the mrGlove brings a convenience and enhances a user experience in the entertainment field.

Acknowledgments This study was supported by Seoul National University of Science and Technology.

References

1. Dipietro L, Sabatini AM, Dario P (2008) A survey of glove-based systems and their applications. IEEE Trans Syst, Man, Cybernetics, Part C: Appli Rev 38(4)
2. Nintendo Wii http://www.nintendo.com/wii
3. Microsoft Kinect http://www.xbox.com/en-US/kinect
4. Jeong YM, Lim KT, Lee SE (2012) mGlove: Enhancing user experience through hand gesture recognition. Jin D, Lin S (eds) Advances in EECM. LNEE, vol 1 Springer, Heidelberg, pp 383–386
5. Terasic DE0-nano board, http://www.terasic.com.tw/en
6. LabVIEW, http://www.ni.com/labview

A Novel Wireless Context-Aware Network of Service Robot

Jianqi Liu, Qinruo Wang, Hehua Yan, Bi Zeng and Caifeng Zou

Abstract With the improvement of sensors, intelligence and wireless communication technologies, the service robot is faced with a new rapid development opportunity, which can utilize pervasive computing of wireless network to sensing everything happened in whole context. This paper presents a novel service robot global context-aware network, and puts emphasis on four key issues, such as ultra wideband radio, wireless positioning technology, wireless body area network and dynamic Bayesian network.

Keywords Service robot · Wireless sensor network · Wireless body area network · Context-aware · Dynamic bayesian network

Introduction

With the rapid development of context-aware technologies such as M2M communication technology [1], wireless positioning [2, 3] in recent years, many institutes and researchers have a strong interest in the service robot supported by

J. Liu · H. Yan (✉) · C. Zou
Guangdong Jidian Polytechnic, Guangzhou 510515, China
e-mail: hehua_yan@126.com

J. Liu
e-mail: liujianqi@ieee.org

C. Zou
e-mail: caifengzou@gmail.com

Q. Wang · B. Zeng
Guangdong University of Technology, Guangzhou 510515, China
e-mail: wangqr2006@gdut.edu.cn

B. Zeng
e-mail: z9215@163.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_41, © Springer Science+Business Media Dordrecht 2014

context-aware technique. The independent context-aware service robot, that integrated calculation, inference, wireless communication and adaptive control, becomes one of leading-edge interdisciplinary research areas today [4]. People above the age of 60 years in China have now exceeded 10 % of the total population, and elderly population will be total 200 million by 2015, which is a new challenge. It is expected that in the near future, nursing care for the elderly will become an important burden, which requires a large number of service robots to complete nursing affairs such as delivery items, cleaning, rehabilitation services, to make up for the lack of nursing staff. The service robot can help to take care of the elderly daily life and improve their quality of life. Therefore, service Robotics market has great potential demand, and will produce good economic results.

In the past, the service robot primarily made use of sensors on the robot body to obtain cognition of the ambient environment, such as using laser range finder to achieve positioning, using acceleration meter to sense collision, and using gas sensor to detect dangerous gas leaks. The defect of these methods is that the environment information is partial around the robot. Fortunately, with the development of context-aware technology based on wireless sensor network (WSN), deploying multiple types of sensors to realize global contextual perception, the robot can understand the each of contextual changes and take appropriate action quickly. For example, once the fire disaster happened, the robot should give an alarm and open the fire hydrants. Similarly, if someone is ill, and is made a diagnosis by a doctor in hospital traditionally, but some disease such as stroke, should be dealed with quickly, which is a disaster for the elderly living alone; so the constant monitor is necessary. The service robot can be competent for this challenge supported by wireless body area network (WBAN). This paper presents a new architecture of context-aware service robot, which integrates sensing, positioning, and body monitor.

The remainder of this paper is organized as follows: We discuss the architecture of context-aware network in section Architecture of Context-Aware Network of Service Robot. In section Key Issues, we review key issues of building network such as UWB, wireless positioning technology, WBAN, and dynamic Bayesian network body status forecasting model. Section Conclusion concludes this paper and outlines some problems that need to be resolved in the future.

Architecture of Context-Aware Network of Service Robot

The scheme of service robot puts WSN, wireless positioning technology, WBAN, dynamic Bayesian network forecasting technology together to acquire the global context information. System deploys sensor to acquire real-time contextual data, and transmits the data to service robot via a wireless communication network. Then the service robot analyzes data and reasons the tasks, which need to cope with imminently. Finally, appropriate control action is taken to complete the task. For example, in kitchen collecting gas and smoke data to judge whether there is gas leak and fire disaster or not; in balcony deploying sensor to detect sun light strength and outdoor weather; in bedroom deploying sensor to collect temperatures humidity; in human body deploying WBAN, using medical sensor to collect people's physiological vital signs, monitor and judge people's body status by dynamic Bayesian network prediction model.

Wireless sensor nodes complete data collection and communication, at the same time, they should complete high accuracy positioning, which is primarily used for the service robot navigation and people's positioning and tracking. The architecture is shown as Fig. 1.

The whole network can be divided into WBAN and WSN according the radio transmit range. In WBAN, all nodes and hubs are organized into logical sets, and coordinated by their respective hubs for medium access and power management as illustrated in left of Fig. 1. There is one and only one hub in a WBAN, whereas the number of nodes in a WBAN is to range from zero to 2 m. In a one-hop star WBAN, frame exchanges are to occur directly between nodes and the hub of the WBAN. In a two-hop extended star WBAN, the hub and a node exchanges frames optionally via a relay-capable node.

A WSN consists of spatially distributed autonomous sensors to monitor physical or environmental conditions, such as temperature, sound, pressure, etc., and to cooperatively pass their data through the network to a main location. The modern networks are bi-directional, also enabling control of sensor activity. The development of wireless sensor networks was motivated by military applications such as battlefield surveillance; today such networks are used in many industrial and consumer applications, such as industrial process monitoring and control, machine health monitoring, and so on [5].

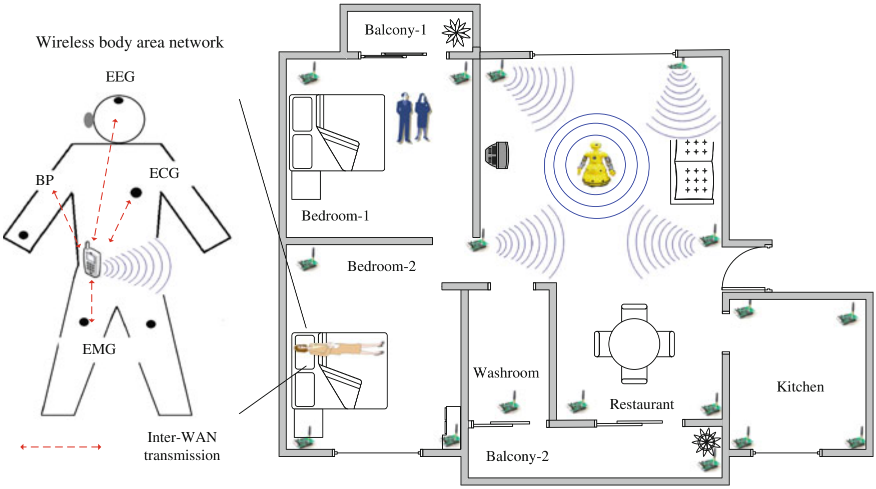


Fig. 1 Architecture of wireless context-aware network of service robot

Key Issues

The service robot wireless context-aware network mainly includes radio, wireless positioning technology, WBAN, and the physiological condition monitoring and prediction model. This paper will focus on four aspects of context-aware network.

UWB Radio

UWB was formerly known as "pulse radio". The FCC and the International Telecommunication Union Radio communication Sector currently define UWB in terms of a transmission from an antenna for which the emitted signal bandwidth exceeds the lesser of 500 MHz or 20 % of the center frequency. A significant difference between conventional radio transmissions and UWB is that conventional systems transmit information by varying the power level, frequency, and/or phase of a sinusoidal wave. UWB transmissions transmit information by generating radio energy at specific time intervals and occupying a large bandwidth, thus enabling pulse-position or time modulation. The information can also be modulated on UWB signals (pulses) by encoding the polarity of the pulse, its amplitude and/or by using orthogonal pulses. UWB pulses can be sent sporadically at relatively low pulse rates to support time or position modulation, but can also be sent at rates up to the inverse of the UWB pulse bandwidth. Pulse-UWB systems have been demonstrated at channel pulse rates in excess of 1.3 gig pulses per second using a continuous stream of UWB pulses (Continuous Pulse UWB or C-UWB), supporting forward error correction encoded data rates in excess of 675 Mbps [6].

A valuable aspect of UWB technology is the ability for a UWB radio system to determine the "time of flight" of the transmission at various frequencies. This helps overcome multipath propagation, as at least some of the frequencies have a line-of-sight trajectory. With a cooperative symmetric two-way metering technique, distances can be measured to high resolution and accuracy by compensating for local clock drift and stochastic inaccuracy.

Another feature of pulse-based UWB is that the pulses are very short (less than 60 cm for a 500 MHz-wide pulse, less than 23 cm for a 1.3 GHz-bandwidth pulse), so most signal reflections do not overlap the original pulse and the multipath fading of narrowband signals does not exist. However, there is still

Table 1 Comparison of UWB, CSS, and ZigBee

Radio	Accuracy of positioning	Bandwidth	Power dissipation	Anti-interference
UWB	10–30 cm	500 Mbps	Low	High
CSS	1–3 m	1 Mbps	Low	Medium
ZigBee	3–5 m	800 M: 20 kbps 2.4 G: 250 kbps	Medium	Low

multipath propagation and inter-pulse interference to fast-pulse systems which must be mitigated by coding techniques. Table 1 gives a comparison of UWB, CSS, and ZigBee.

Wireless Positioning Technology

The positioning plays an important role in system. We know that the positioning is difficult in non-line-of-sight (NLOS) environment. There are various obstacles, such as walls, which lead to multi-path effects [3]. Some interference and noise from other wireless networks such as WIFI, or electrical radiating equipment such as microwave ovens, degrade the accuracy of positioning. Irregular building geometry and the density of water vapor in the air leads to reflection, and extreme path loss. So indoor positioning is more complex.

The positioning algorithms can be classified into two categories: Rang-based and Rang-free. Range-Based algorithm positioning by measuring the distance between the anchor nodes and mobile node or angle information, use the trilateration, triangulation or Maximum Likelihood estimator (ML) positioning method for calculation of the node location; Range-Free location is not required to distance or angle information, thus only according to network connectivity information and so on can be realized. Range-based algorithms usually need some special hardware to obtain accurate absolute range measurements and can achieve higher positioning accuracy than range-free algorithms. Range-free algorithms, on the other hand, do not need special hardware and are at low cost [7]. Range-Based positioning algorithm includes Angle of Arrival (AOA) [8], Time of Arrival (TOA) [9], Time Difference of Arrival (TDOA) [10], Received Signal Strength Indication (RSSI) [11], Time of Flight (TOF) [12], Symmetrical Double Sided Two Way Ranging (SDS-TWR) [13], etc.

Ubisense 7,000 serial is an in-building UWB radio based tracking system developed by Ubisense company [14], which can determine the positions of people and objects to an accuracy of a few tens of centimeters, using small tags which are attached to objects and carried by personnel, and a network of receivers which are placed around buildings. UWB is well-suited to in-building of emergency field [15], because of its non-line-of-sight nature, 3D-position, modest infrastructure requirements and high tracking accuracy [16]. A properly-architected UWB tracking system is low-power, and the fundamental technology is simple and low-cost.

Wireless Body Area Network

A WBAN is a wireless network of wearable computing devices [17, 18]. WBAN is a basic technology, which can long-term monitor and record the signal of the human health. In particular, the network consists of several miniaturized body sensor units together with a single body central unit [19]. A typical WBAN requires vital sign monitoring sensors, motion detectors to help identify the location of the monitored individual and some form of communication, to transmit vital sign and motion readings to medical practitioners or care givers. A typical WBAN will consist of sensors, a processor, a transceiver and a battery. Physiological sensors, such as ECG and SpO2 sensors, have been developed. Other sensors such as a blood pressure sensor, EEG sensor and a PDA for BSN interface are under development [20].

Early application is used to continuously monitor and record of chronic diseases, such as diabetes, asthma and heart disease, health parameters in patients, to provide some way automatically therapy control. For example, once the insulin levels of the patients with diabetes drop down, his WBAN can immediately activate a pump automatically injecting insulin for patients, so that patients without a doctor can control insulin in the normal level. WBAN is the smallest coverage network, but it is benefiting a wide network. WBAN is applied into service robot, and robot use this advanced technology to monitor the physical condition of the elderly. Meanwhile, this technology may boost the real implementation of the telemedicine.

Dynamic Bayesian Network Monitor and Prediction Model

This scheme aims to discover the nonlinear association between the ill risk and body vital signs (such as temperature, blood sugar, blood pressure, heart rate and pulse, etc.). The dynamic Bayesian network model can be expressed for a probability dependent relationship between the physical status and body vital signs, which varies with time. Building health status monitoring and prediction model for the elderly consists of four main steps:

- Traditional static Bayesian network should be built in first;
- Collect the various physiological signs and analyze signs to find probability dependency relationship among them;
- Establish dynamic Bayesian network model, Probe the nonlinear association patterns between the ill risk and each of signs;
- Using system dynamics methods, computer simulation and clinical experience to validate and assess the dynamic nonlinear association model;
- Evaluate and verify the dynamic Bayesian network prediction model, attempting to make a reasonable medical explanation for the change of the human condition.

Conclusion

This paper discusses the service robot global context-aware network, gives an analysis of the limitations of the traditional robot, and proposes a framework for service robots perceive global environment. Four key issues, including UWB radio, wireless positioning technology, WBAN and dynamic Bayesian network model, are discussed. According to the progress of existing projects, if the four key technologies are grasped effectively, the establishment of the wireless context-aware network will become easy.

Acknowledgments The authors would like to thank the Natural Science Foundation of Guangdong Province, China (No.9151009001000021, S2011010001155), the Ministry of Education of Guangdong Province Special Fund Funded Projects through the Cooperative of China (No. 2009B090300 341), the National Natural Science Foundation of China (No. 61262013), the High-level Talent Project for Universities, Guangdong Province, China (No. 431, YueCaiJiao 2011), and the Chinese Society of Vocational and Technical Education 2012–2013 scientific research and planning projects (NO. 204921) for their support in this research.

References

1. Chen M et al (2012) Machine-to-machine communications: architectures, standards, and applications. KSII Trans Int Inf Syst 6:480–497
2. Liu J et al (2012) Towards real-time indoor localization in Wireless sensor networks, in Computer and Information Technology (CIT), 2012 IEEE 12th international conference on 2012, pp. 877–884
3. Suo H et al. (2012) Issues and challenges of wireless sensor networks localization in emerging applications, In: Proceedings of 2012 International Conference on Computer Science and Electronics Engineering, Hongzhou, pp 447–451
4. Wan J et al (2011) Advances in cyber-physical systems research. KSII Trans Int Inf Syst 5:1891–1908
5. Dargie W, Poellabauer C (2010) Fundamentals of wireless sensor networks: theory and practice: Wiley
6. Fernandes JR, Wentzloff D (2010) Recent advances in IR-UWB transceivers: an overview, in Circuits and Systems (ISCAS). In: Proceedings of IEEE International Symposium on 2010, pp 3284–3287
7. Zhang S et al (2010) Accurate and energy-efficient range-free localization for mobile sensor networks. IEEE Trans Mobile Comput 9:897–910
8. Priyantha NB et al. (2001) The cricket compass for context-aware mobile applications. In: Proceedings of the 7th annual international conference on mobile computing and networking, pp 1–14
9. Harter A et al (2002) The anatomy of a context-aware application. Wireless Netw 8:187–197
10. Girod L, Estrin D (2001) Robust range estimation using acoustic and multimodal sensing, in Intelligent robots and systems. In: Proceedings IEEE/RSJ International Conference on 2001, pp 1312–1320
11. Girod L et al (2002) Locating tiny sensors in time and space: A case study, In: Computer design: VLSI in computers and processors, 2002. Proceedings of IEEE International Conference on 2002, pp 214–219

12. Lanzisera S et al (2006) RF time of flight ranging for wireless sensor network localization in intelligent solutions in embedded systems, 2006 International Workshop on 2006, pp 1–12
13. *nanotron find*. Available: http://www.nanotron.com/EN/PR_find.php
14. *ubisense*. Available: http://www.ubisense.net/
15. Hill R et al (2004) A middleware architecture for securing ubiquitous computing cyber infrastructures. IEEE Distrib Syst Online 5:1
16. Mahfouz MR et al (2008) Investigation of high-accuracy indoor 3-D positioning using UWB technology. IEEE Trans Microw Theory Tech 56:1316–1330
17. Chen M et al (2011) Body area networks: a survey. Mobile Netw Appli 16:171–193
18. Chen M et al (2012) Machine-to-machine communications: architectures, Standards and applications
19. Ullah S et al (2012) A comprehensive survey of wireless body area networks, J Med Syst, vol 36, pp 1065–1094, 2012/06/01
20. O'Donovan T et al (2009) A context aware wireless body area network (BAN), In: Pervasive computing technologies for healthcare, 3rd International Conference on PervasiveHealth 2009, pp 1–8

Architecture of Desktop as a Service Supported by Cloud Computing

Jianqi Liu, Hehua Yan, Caifeng Zou and Hui Suo

Abstract Traditional desktop virtualization assumed the communication between client and cloud server happened on high-speed network, but as popularization of multimedia application, a large amount of data result in network congestion. With the development of cloud computing, desktop as a service is emerged as a new service for desktop delivery. This paper presents a new architecture of DaaS supported by client and protocol co-designed method, which deploys hardware microprocessor to encode and decode the channel data in client, and decreases the transmission data over the network by compression algorithm and protocol optimization.

Keywords Cloud computing · Virtual desktop infrastructure · Desktop as a service · Remote desktop protocol

Introduction

A new word, Desktop as a Service (DaaS) [1] is emerged, which is a natural evolution of desktop virtualization paradigm whereby desktops would be delivered as a service from a Desktop Cloud. It is similar to Software as a service (SaaS)

J. Liu · H. Yan (✉) · C. Zou · H. Suo
Guangdong Jidian Polytechnic, Guangzhou, China
e-mail: hehua_yan@126.com

J. Liu
e-mail: liujianqi@ieee.org

C. Zou
e-mail: caifengzou@gmail.com

H. Suo
e-mail: suohui79@163.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_42, © Springer Science+Business Media Dordrecht 2014

[2, 3] model. DaaS provides benefits of virtual desktop infrastructure (VDI) [4, 5] without the extra costs or risks of owning and managing physical resources. Customers who need large number of desktops for their employees are no longer required to provision all of the required resources, such as servers and storage, but can outsource these tasks to desktop cloud and focus on business critical task instead.

The essence of DaaS is desktop virtualization [6, 7], which is the combination optimization of computing and Communication, and shows the feature of centralization computing mode. In brief, DaaS places display and control on the front-end client and the computation or storage on back-end "Cloud" (data or computing center), the communication between front-end client and backend cloud supported by high speed network. The application execution is migrated from local (to the user) client to a remote data center. Client becomes a lightweight computer that handles only keyboard, mouse and monitor, as well as locally attached devices such as scanners and printers. Communication between client and desktop cloud is handled by remote desktop protocols (RDP). The virtual desktop paradigm has several advantages over the typical "fat-desktop" approach. Administrative costs of the DaaS are significantly lower because operating system (OS) images, applications, and data are no longer installed on a large number of distributed systems but in a desktop could, which improves manageability of the system as well as data and application security. Moreover, since the local client device is stateless, and it is very easy to troubleshoot and replace, thus on-site labor is significantly reduced. In a word, DaaS is emerging as an alternative to traditional desktop delivery.

Currently, most of the proposed desktop virtualization systems [4, 8–11] are based on the technologies to provide the control desktop. The premise of traditional client is that commodity networks are fast enough to use a low-level protocol to remotely serve graphical displays of common, GUI-based applications without any noticeable performance degradation. This leads to take the notion of thin clients to the limit by removing all state and computation from the desktop and designing a low-level hardware- and software-independent protocol to connect all user-accessible devices to the system's computational resources over a low-cost commodity network, for example, Stateless, Low-level Interface Machine (SLIM) [12]. They pay more attention to supplying high quality display effects on clients regardless network bandwidth consumption. These protocols do not lay emphasis on optimizing multimedia data transmission, do not improve the interaction experiences between users and machines, do not reduce the delay and bandwidth consumption of network, and ignore co-design between protocol and client.

To address the above issues, we present a new architecture of DaaS supported by protocol and client collaborative design, which optimizes communication protocol by data compression, and utilizes the hardware unit to realize image and video decoding. The new architecture can improve the network performance drastically through increasing very few time delay generated by decompression and compression.

Overview

DaaS represents a kind of new computing service; users can enjoy high computing performance and huge storage capacity supported by cloud server only through a lightweight client. Desktop cloud is a set of physical resources (such as storage, servers, networking gear, etc.) together with virtualization, connection brokering, and management software allowing for remote access to large numbers of desktops (potentially tens or hundred of thousands). The overview of DaaS is presented in Fig. 1. Desktops execute in data centers and users access them via the Internet by RDP. The response speed of system depends on three portions: thin client, RDP, desktop cloud server. As the desktop cloud server has huge storage and super computing ability to process desktop delivery, data process ability of thin client and network capability are two key factors to the DaaS. One of the direct methods is increasing the bandwidth of network, but in WAN, bandwidth is limited by telecommunication operator. Moreover, using more bandwidth will cost more communication fee. The other alternative is decreasing communication data amount. Before data transfer, we deploy some encoding technology to compress data in cloud server. But this method will bring a new problem: software decoding process will result in time delay, as the process ability of embedded system on chip (SOC) is low in thin client. So we should take these issues into account, and design a new architecture.

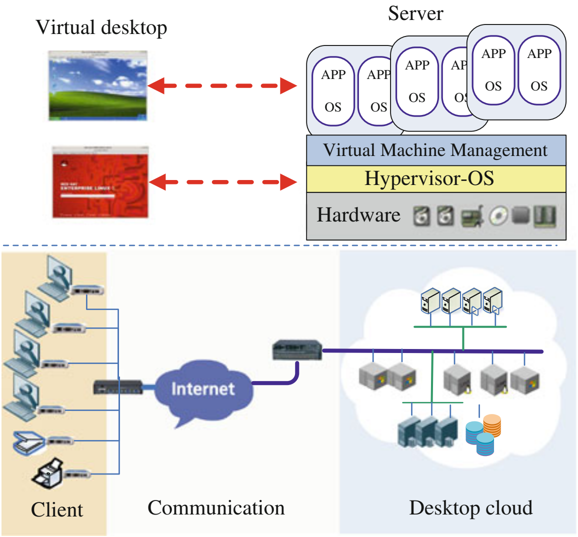


Fig. 1 Overview of DaaS. Users access their desktops and applications using thin client which forwards keyboard and mouse events and screen state updates over the local or wide area network. OS and applications run in virtual machine of the remote desktop cloud

Architecture

The desktop is delivered to the client from desktop cloud; the element of desktop (picture, text, mouse event, keyboard event etc.) from cloud should be processed by thin client quickly and transmit events and other data to desktop cloud promptly. The data is classified into image, video, sound, USB data, and keyboard event, mouse event. We deploy JPEG, MPEG4, MP3 encoding technology and lz4 to compress the data for transmission. In thin client, we deploy hardware decoding microchip to decode corresponding data. The details will be discussed in following three sections.

Thin Client

Formerly, we have become accustomed to the PC model, which allows every user to have their own CPU, hard disk, and memory to run their desktop application. However, in "Cloud" age, desktop virtualization is a modern way to delivery desktop, where multiple users share the processing power of cloud server by a simple client. The client usually deploys the system-on-chip (SoC) solution, which has several advantages over the traditional PC model, including lower costs, better energy efficiency, and simplified administration. Traditional independent communication protocol needs high-bandwidth network to transmit multimedia data. However, commodity wide-area networks are very slow, especially in some developing country, the independent protocol isolated by hardware result in bad user experience as the image and video time-consuming decoding. This paper deploys encoding and decoding chip to process channel data, such as image, sound and video.

In this architecture, S5PV210 chip of ARM CortexTM-A8 core is used as main control chip and also ARM V7 instruction set is adopted, relying on which, it can achieve 1 GHz of basic frequency. The hardware framework is shown as Fig. 2.

This thin client design has several features.

- High definition desktop display. Client is inbuilt with high-performance PowerVR SGX540 3Dgraphics engine and 2D graphics engine, supporting 2D/3D graphics acceleration. Its polygon formation rate is 28 million polygons per second, pixel fill rate can be up to 0.25 billion per second, and it supports PC level display technologies such as DX9, SM3.0, OpenGL2.0, etc. These can provide high performance desktop display update.
- Rapid encoding and decoding. Client's Multi Format Codec (MFC) supports the encoding and decoding of videos with MPEG4, H.263, H.264 and other formats, and supports simulated/digital TV output. With JPEG hardware encoding and decoding, the supported resolution ratio can be up to 8000 × 8000, and can play and record smoothly video documents of 1920 × 1080 pixel (1080p) at 30 frames per second. Audio encoding also supports MP3, WMA, EAAC + and

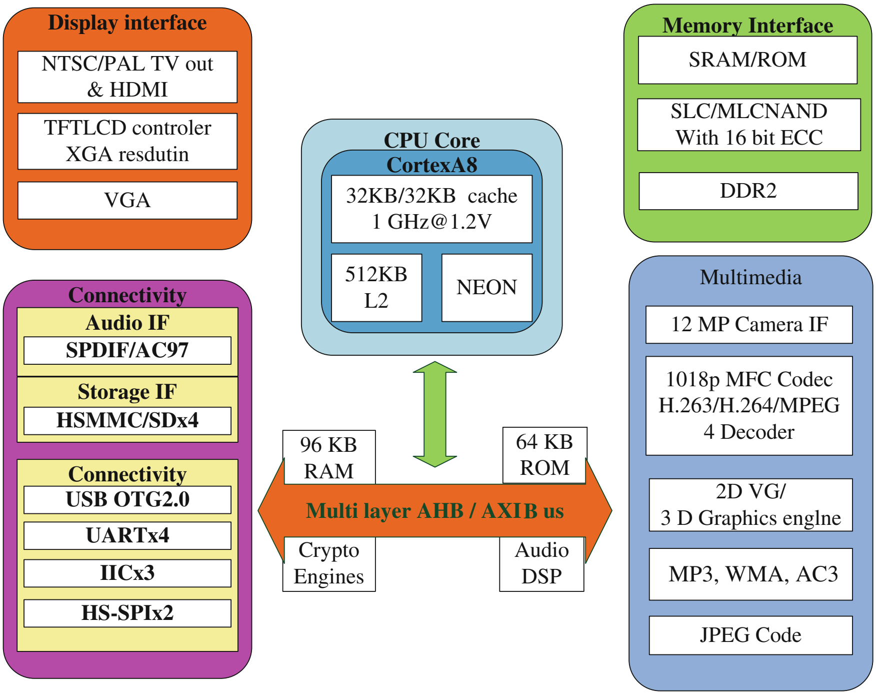


Fig. 2 Hardware framework

AC3. These hardware chips can complete encoding and decoding quickly, and people won't be affected by such a short time delay.

- Abundant interface. The client is inbuilt with HDMI interface, so that high-definition videos can be transmitted to external display. The client also provides three I2S, USB Host 2.0, and USB 2.0 OTG operating at high speed (480 Mbps), four SD Host and high-speed Multimedia Card interface.

Communication Protocol

Communication happened between front-end client and back-end cloud usually adopts the remote desktop protocols (RDP). The image, video, sound, keyboard event, mouse event and USB data are transmitted through the network. In order to allow some degree of flexibility in thin client and cloud server implementation, communication session is split into multiple communication channels (e.g., every channel is a remote device) to control communication and execution of messages according to the channel type, and to add and remove communication channels during run time. The following communication channels are defined: (a) the main channel services as the main session connection for sending control message; (b) events channel for sending mouse and keyboard events; (c) video channel for

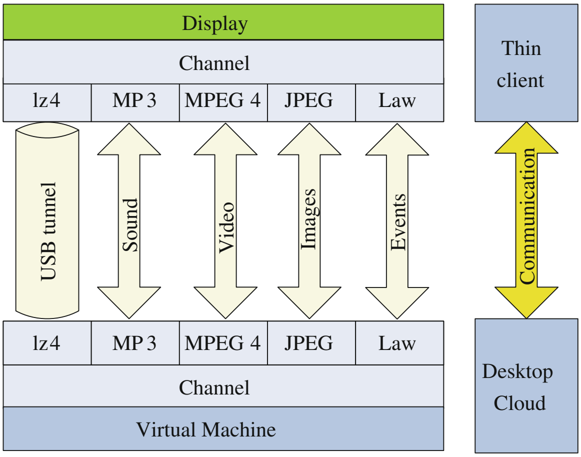


Fig. 3 Communication protocol of DaaS

receiving remote video updates; (d) sound channel for sending and receiving audio stream; (e) image channel for receiving display updates and (f) USB channel for sending and receiving USB data deploys as a tunnel. More channel types will be added as the protocol evolves expediently. In order to improve compatibility, communication protocol supports bidirectional exchange of channels compatibilities. In order to achieve higher performance, different data type deploy different compression algorithm. The Fig. 3 illustrates the communication protocol.

Desktop Cloud Server

The kernel of desktop cloud server is sharing hardware and software resource by virtual machine management module. Virtual machine can give better server utilization, better server management, and better power efficiency in desktop cloud. Now, there are some paradigms, such as VMware, Virtualbox, KVM, Xen and Hyper-V. In this paper, we utilize KVM as test platform. First, KVM is free software released under the general public license (GPL), and we can modify the remote desktop protocol (add compression algorithm in channel data process) to ease the pressure of communication network. Secondly, KVM enhances the performance by utilizing hardware virtual machine (HVM) technology, Intel VT and AMD-V, supported by CPU vendor. These instruction set extensions provide hardware assistance to virtual machine monitors. They enable running fully isolated virtual machines at native hardware speeds, for some workloads. Thirdly, KVM is part of Linux and uses the regular Linux scheduler and memory management. This means that KVM is much smaller and simpler to use.

Conclusion

Our main focus is to provide high-quality remote access from virtual machine of desktop cloud. We try to break down the barriers of traditional desktop virtualization. We present a new architecture of DaaS, which aims to decrease the transmission data over the network by compression algorithm and protocol optimization. In order to improve response speed, we deploy hardware encoding and decoding microchip in thin client to process data from or to communication channels. This architecture of DaaS supported by co-designed method overcomes the defects of traditional desktop virtualization, which has the following distinctive characteristics: (a) Low cost and energy-saving; (b) Multimedia supported; (c) Low bandwidth and low latency; (d) Good user experience. But some issues should be addressed in the near future, such as desktop migration, and dynamic resource allocation control.

Acknowledgements The authors would like to thank the Natural Science Foundation of Guangdong Province, China (No.9151009001000021, S2011010001155), the Ministry of Education of Guangdong Province Special Fund Funded Projects through the Cooperative of China (No. 2009B090300 341), the National Natural Science Foundation of China (No. 61262013), the High-level Talent Project for Universities, Guangdong Province, China (No. 431, YueCaiJiao 2011), and the Chinese Society of Vocational and Technical Education 2012–2013 scientific research and planning projects (NO. 204921) for their support in this research.

Reference

1. Cristofaro S et al. (2010) Virtual distro dispatcher: a light-weight desktop-as-a-service solution in cloud computing, Springer pp 247–260
2. Turner M et al (2003) Turning software into a service. Computer 36:38–44
3. Buxmann P et al (2008) Software as a service. Wirtschaftsinformatik 50:500–503
4. Velte A, Velte T (2009) Microsoft virtualization with Hyper-V: McGraw-Hill, Inc
5. Baratto RA et al. (2004) Mobidesk: mobile virtual desktop computing. In: Proceedings of the 10th annual international conference on mobile computing and networking, pp 1–15
6. Hazari S, Schnorr D (1999) Leveraging student feedback to improve teaching in web-based courses. The Journal 26:30–38
7. Watson J (2008) Virtualbox: bits and bytes masquerading as machines. Linux J 2008:1
8. Citrix Corporation,Citrix Application Delivery Infrastructure. Available: http://www.citrix.com/
9. Kivity et al. (2007) kvm: the Linux virtual machine monitor. In: Proceedings of the Linux Symposium, pp 225–230
10. Sugerman J et al. (2001) Virtualizing I/O devices on VMware workstation's hosted virtual machine monitor. In: Proceedings of the general track: 2002 USENIX annual technical conference, pp 1–14
11. Rosenblum M (1999) VMware's virtual platform^{TM}. In: Proceedings of hot chips, pp 185–196
12. Schmidt BK et al (1999) The interactive performance of SLIM: a stateless, thin-client architecture. ACM SIGOPS Oper Syst Rev 33:32–47

Lattice Boltzmann Method for the Velocity Analysis of Polymer Melt in Vane Extruder

Jianbo Li, Jinping Qu, Xiaoqiang Zhao and Guizhen Zhang

Abstract After some mixed methods in polymer processing are summarized, the new equipment is introduced which is called vane plasticizing extruder with elongational flow field. Then the unit of vane geometry is analyzed, and the analysis method of vane extruder conveying state is put forward by using the lattice Boltzmann method. On this base, physical model and mathematical model are established. Calculation of velocity distribution and quantitative description of the elongational deformation rate shows there has elongational rheological in the vane plasticizing extruder. Periodic extensional flow field has very good mixing effect in polymer processing.

Keywords Vane extruder · Lattice Boltzmann method · Positive displacement · Elongational rheology · Normal stress field

J. Li (✉) · J. Qu · X. Zhao · G. Zhang
The National Engineering Research Center of Novel Equipment for Polymer Processing,
Guangzhou, China
e-mail: gdjdljb@126.com

J. Qu
e-mail: jpqu@scut.edu.cn

X. Zhao
e-mail: xqzhaocn@gmail.com

G. Zhang
e-mail: gaoyuanxuefen@126.com

J. Li · J. Qu · X. Zhao · G. Zhang
The Key Laboratory of Polymer Processing Engineering of Ministry of Education South
China University of Technology, Guangzhou 510640, Guangdong, China

J. Li
Guangdong Jidian Polytechnic, Guangzhou 510515, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_43, © Springer Science+Business Media Dordrecht 2014

Introduction

Mixing, which is an indispensable stage in polymer processing technology, to a great extent, determines the final performance of product. Components are scattered through physical movement in the process of polymer melt mixing [1]. In order to get better mixing effect, a lot of work has been done by predecessors.

Tadmor analyzes the solid-phase breakage mechanism, and compares the shear and tensile field dispersion ability. Dumbbell ball is studied under the action of external field separation. The adhesion strength of the two small balls is equivalent to the linkage between them. Particle separation occurs when the field force is greater than the binding force between spheres. From the point of view of the largest force of solid phase particles, elongational flow effect is double shear flow effect at the same deformation rate (tensile rate and shear rate). So it can be spread to get higher efficiency and more effective dispersion effect using stretch flow field [2].

Spherical agglomerates which have a certain intensity distribution is studied by Manas [3], under the low Reynolds number,in the shear flow field, elongation flow field, uniaxial tension, biaxial stretching field. Dispersing ability is compared in two-dimensional flow field and three-dimensional flow field with or without the curl. Manas thinks a plane of spherical agglomerates determines its strength, and aggregates disconnection along the plane when the flow force is greater than its strength [4]. His research shows that elongational dispersion efficiency is better than that of the simple shear field.

Chris Rauwendaal opens many tapered slot on the wedge screw arris. When through the screw edge and screw arris of the tapered slot, the melt is doubly stretching flow field effect, which can produce strong stretching flow [5].

Suzaka opens multiple convergent channels with stretch on the runner plate, and discusses the draw ratio of big diameter and path of convergent channel [6].

In order to get continuous stretch flow field, a novel polymer processing equipment is designed by Qu [7], a positive displacement conveying facility which totally alters the shear conveying mechanism. In this article, the working principle of this equipment will be described, with the melt flow speed in flow field as the key of the research.

Model of the Vane Extruder

Structure of the Vane Extruder

The structure schematic diagram of the vane extruder is shown in Fig. 1. The components of the vane extruder are a stator, a rotor, and four vanes. The four vanes are installed in pairs, each other contacting at the bottom surface. When the rotor is rotating, the volume is periodically changing which is constituted by the stator, the rotor, and the vanes.

The materials in the previous unit are fed into the current vane unit when the volume becomes large. The materials in the current unit are discharged to the next vane unit when the volume becomes small.

Geometric Structure of the Vane Unit

The geometric structure of the vane extruder unit is shown in Fig. 2.

The effective radius vector of the vane unit is defined as the clearance between the stator and the rotor radius. Assume that the radius of the stator is R, the radius of the rotor is r, the eccentricity between the stator and rotor is e, the angle between the effective radius vector and x axis is θ. So, the effective radius vector can be expressed in the following expression:

$$\rho(\theta) = \sqrt{R^2 - e^2 \sin^2 \theta} + e\cos\theta - r \tag{1}$$

The Lattice Boltzmann Method

It is more difficult to use the analytic solution to describe melt flow in the vane unit, while the Lattice Boltzmann Equation can effectively analyze the melt flow in the vane extruder.

Lattice Boltzmann Equation (LBM) originates from Lattice Gas Automata (LGA). LGA model, taking the fluid as a hypothetical particle on a regular lattice, asserts that particle collisions and migration occur according to certain rules in the grid, and macroscopic quantity is obtained by statistical, such as density velocity etc. [8].

The earliest LBM model was first proposed by McNamam and Zanetti in 1988, whose evolution process uses the Boolean variable statistical distribution function of particles in the LGA [9].

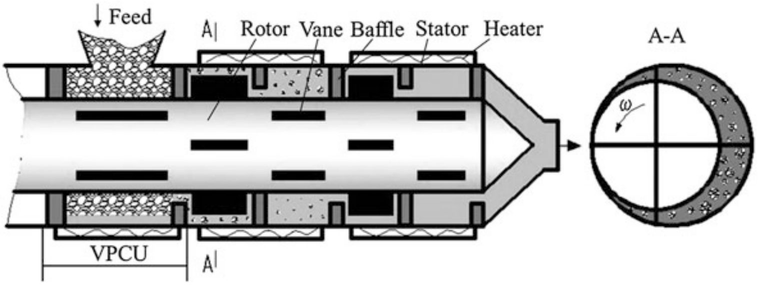
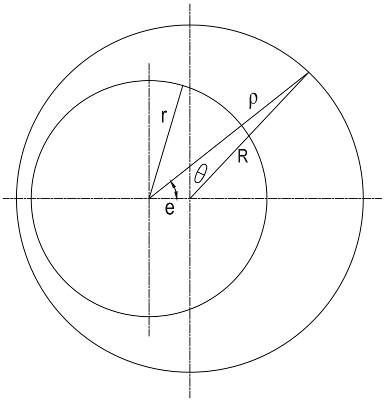


Fig. 1 The structure schematic diagram of the vane extruder

Fig. 2 Geometric structure of the vane unit



In 1991, single relaxation Bhatnagar-Gross-Krook model (BGK), a more convenient calculation, was proposed [10]. BGK model has a single relaxation time, and collision term convenient calculation, which make the LBGK model widely used.

DnQm model is the most representative model proposed by Qian [11], in which n is the dimension of the problem, m is the discrete velocity. In this paper, we use the D2Q9 model, two dimensions and nine discrete velocity, as shown in Fig. 3

The lattice Boltzmann equation of single relaxation can be approximated as [12]

$$f_k(x + \Delta x, t + \Delta t) = f_k(x, t) \cdot [1 - \omega] + \omega \cdot f_k^{eq}(x, t) \tag{2}$$

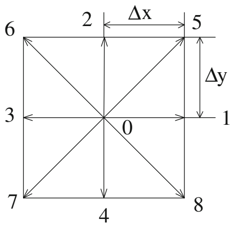
The equilibrium distribution function can be written as [12]

$$f_k^{eq}(x, t) = w_k \cdot \rho(x, t) \cdot \left[1 + \frac{\vec{c}_k \cdot \vec{u}}{c_s^2} + \frac{1}{2} \cdot \frac{(\vec{c}_k \cdot \vec{u})^2}{c_s^4} - \frac{1}{2} \cdot \frac{\vec{u}^2}{c_s^2}\right] \tag{3}$$

The macroscopic fluid density and the momentum can be represented as [12]

$$\rho = \sum_0^8 f_k \quad \vec{u} = \left(\sum_0^8 f_k \vec{c}_k\right) / \rho \tag{4}$$

Fig. 3 The D2Q9 model



k is the flow direction of the lattice, k = 0 ~ 8

w_k is the weight factor

$\vec{c}_k$ is grid unit vector in the flow direction

c_s is the speed of sound

ω is a diffusion coefficient about LBM relaxation time

ρ is the macro density in the grid at point x, at moment t

$f_k^{eq}(x, t)$ is the equilibrium distribution function in the grid at point x at moment t

$f_k(x, t)$ is the distribution function in the grid at point x, at moment t

$f_k(x + \Delta x, t + \Delta t)$. is the distribution function in the grid at point x + Δx, at the moment t + Δt

Results and Discussion

The Velocity Distribution

The rotor radius of the vane extruders r is 15 mm, The stator radius R is 20 mm, the eccentricity e is 2 mm. The circumference of the rotor is 2*PI*r = 94 mm. For simple calculations, we divide 200 at the x direction, 100 at the y direction. Assume that the inlet velocity uMax is 0.02; the Reynolds number Re is 100; the kinematic viscosity nu is uMax*2*r/Re = 0.006; the relaxation parameter omega is 1.0/(3.0*nu + 0.5) = 1.9305.

After numerical computation, dimensionless inlet velocity is 1. The velocity distribution in vane unit is shown in Fig. 4 after numerical calculation.

The Numerical Analysis of Velocity

In order to get more detailed information of velocity in the flow field, the velocity is analyzed in the middle part of the effective radius vector. The graphics of x

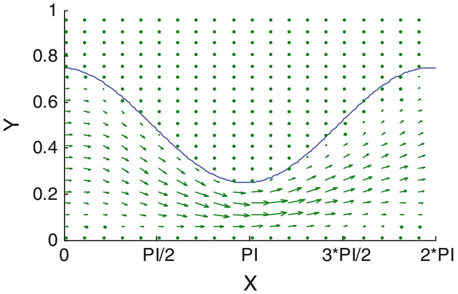


Fig. 4 The velocity distribution in vane unit

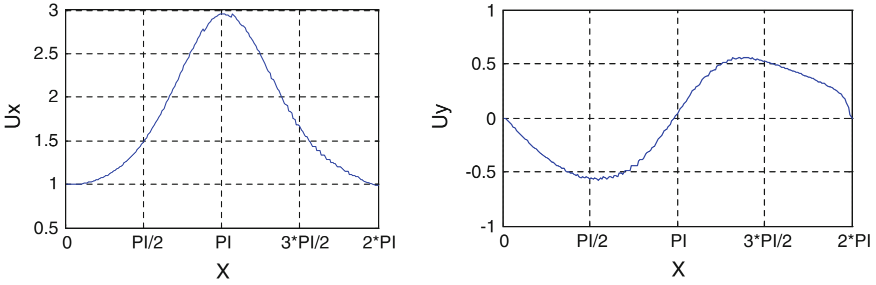


Fig. 5 The Ux and Uy in the middle of the effective radius vector

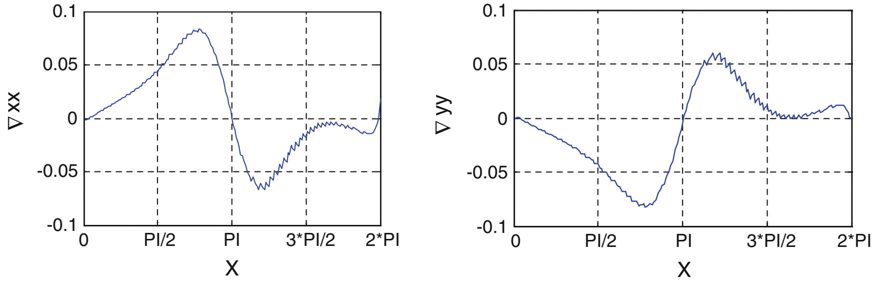


Fig. 6 The ∇xx and ∇yy in the middle of the effective radius vector

direction velocity and y in the middle part of the effective radius vector is shown in Fig. 5. The graphics of ∇xx and ∇yy is shown in Fig. 6.

Discussion

The Strain rate tensor component in the x direction ∇xx changes again in the vane unit. If $\nabla xx > 0$ in the lattice number $0 \sim$ PI of x direction, it has a maximum value 0.8 at about 130°; if $\nabla xx < 0$ in the number PI $\sim$ 2*PI, it has a minimum -0.8 at about $-130°$. Also in the y direction, if $Uy < 0$ in the lattice number $0 \sim$ PI of x direction, it has a minimum value -0.6 at about 130° lattice Point; if $Uy > 0$ in the number PI $\sim$ 2*PI, it has a maximum value 0.6 at about $-130°$.

In the lattice number $1-100$ of x direction, the melt is stretched in the x axis, compressed in the y axis; or, the melt is compressed in the x axis, stretched in the y axis. This kind of phenomenon is repeated in the next unit, that is, the elongation deformation rate of the melt changes again and again in the vane extruder.

Conclusions

As a novel equipment based on elongation rheology, method and equipment of the vane extruder is a completely new theory and practice for polymer processing. The vane extruder totally alters the shear conveying mechanism of the screw extruder in polymer processing, presenting good effects of mixing, distribution and dispersion. Through some experimental investigations, high stability, wide adaptability, shorter polymer thermo-mechanical history, better plasticating and mixing effects are found in the vane extruder compared with the screw extruder.

Acknowledgments The authors wish to acknowledge the National Nature Science Foundation of China(Grant 10872071, 50973035, 50903033, 51073061, and 21174044), National Key Technology R&D Program of China (Grant 2009BAI84B05 and 2009BAI84B06), the Fundamental Research Funds for the Central Universities (NO. 2012ZM0047), Program for New Century Excellent Talents in University (No.NCET-11-0152), Pearl River Talent Fund for Young Sci-Tech Researchers of Guangzhou City (No.2011J2200058), 973 Program (No.2012CB025902) and National Natural Science Foundation of China-Guangdong Joint Foundation Project (U1201242) for their financial supports.

References

1. Tadmor Z, Gogos CG (2006) Principles of polymer processing, 2nd edn. Wiley, New Jersey, pp 473–475
2. Tadmor Z (1976) Forces in dispersive mixing. Ind Eng Chem Res Fundam 15(4):346–348
3. Zloczower IM, Tadmor Z (1994) Dispersive mixing of solid additives. Mixing Compd Polymer-Theory Pract 23:55–84
4. Manas-Zloczower Ica (1994) Studies of mixing efficiency in batch and continuous mixers. Rubber Chem Technol 67(3):504–528
5. Rauwendaal C, Osswald T, Gramann P (1999) Design of dispersive mixing devices. Int Polym Proc 14(1):28–34
6. Suzaka Y (1982) Mixing device. USP. 4334783
7. Jinping Q (2009) A method and a device for plasticating and transporting polymer material based on elongational flow. EP. 2113355
8. Wolfram S (1986) Cellular automaton fluids 1: basic theory. J Stat Mech 45:471–529
9. Mcnamara GR, Zanetti G (1988) Use of the Boltzmann equation to simulate lattice gas automata. Phys Rev Lett 61(20):2332–2335
10. Bhatnagar P, Gross EP, Krook MK (1954) a model for collision processes in gases. I. Small amplitude processes in charged and neutral one—component systems. Phys Rev 94(3):511–525
11. Qian Y, d'Humieres D, Lallemand P (1992) Lattice BGK models for Navier-Stokes equation. Europhysits Lett 17:479–484
12. Mohamad AA (2011) Lattice Boltzmann method fundamentals and engineering applications with computer codes. Springer, London, pp 69–72

Integrated Approach for Modeling Cyber Physical Systems

Shuguang Feng and Lichen Zhang

Abstract Cyber physical systems contain three parts: control part, communication part and physical part. In this paper, we propose an integrated approach for modeling cyber physical systems. Differential Dynamic Logic is used for modeling control part, Communicating Sequential Process (CSP) is applied to specify communication part, and Modelica is used for modeling physical part of cyber physical systems. The proposed approach is illustrated by a case study of the one-street style of vehicles on the street.

Introduction

The Differential Dynamic Logic, whose abbreviation is dL, is developed by Platzer [1]. Differential Dynamic Logic dL is a Dynamic logic for Hybrid Systems and it has some advantages over Dynamic Logic [2]. Differential Dynamic Logic dL has differential equations for Hybrid Systems which contains both discrete and continuous dynamics. It is suitable for Differential Dynamic Logic dL to model complex systems.

Communicating Sequential Processes (CSP) is a formal language and firstly described by Hoare [3]. CSP is a tool for specifying and verifying the concurrent processes of systems. CSP is famous for its mathematical theories of concurrency which is known as process algebras. In this paper, we use CSP to describe a signal system. The simplified syntax of CSP is as below, where P is a process and p is an event.

Modelica is an object-oriented language for system modeling and simulation [4]. Modelica is also non-causal modeling language, and there are tools that

S. Feng · L. Zhang (✉)
Shanghai Key Laboratory of Trustworthy Computing East China Normal University,
Shanghai 200062, China
e-mail: zhanglichen1962@163.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_44, © Springer Science+Business Media Dordrecht 2014

support the model generation of Modelica. Below is the simple example of Modelica.

In this paper, we propose an integrated approach for modeling cyber physical systems. We build a model about one-street driving using dL, CSP, and Modelica. With dL, we build the control component of the model, which supervises the condition of the street and send signals to the cars on the street. With CSP, we build the signal transmission component of the model, which send data between control component and the cars on the street. The cars' data can be sent to control component via signal transmission component.

```
model simpleModel
equation
end simpleModel;
```

Model Description

The model is in Fig. 1. The street is unidirectional. Cars run on the street. No cars can run with another side by side but the moment of overtaking another car. Cars can accelerate and brake. And, one car can be driven to the street from outside or driven off the street. Control system calls the information of cars and sends the command to cars via signal system described by CSP. Signal system has four parts: CarReceive, CarSend, CtrlReceive and CtrlSend. THe CarSend part collects car's data. CtrlReceive part receives data from CarSend part. CtrlSend part sends

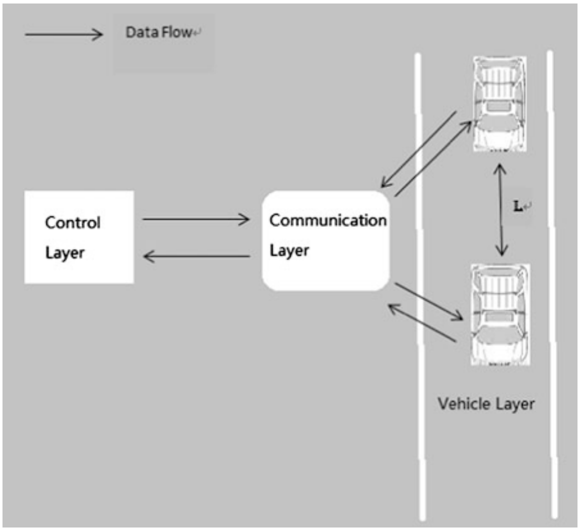


Fig. 1 Model of whole system

command from control system. CarReceive part receives command from CtrlSend part.

We introduce the variables needed: x is the position of the car, x is a scalar; v is the velocity of a car and a vector; a is the acceleration of a car and a vector; ID is the identification number of a car, which is assigned by the car's manufacturers and different from each other; state is the current state of a car.

Control System

Control system has five scenes to deal with. There is a standard time interval t and a length l. There is a length variable l. If two cars length is longer than l and the car can outstrip the one just in front of it within t, change the state of this car to ACCELERATION. If the distance between two cars is longer than I, keep the cars speed v to its current value and s to 0. If the distance between two cars is shorter than I and the car can't outstrip the one just in front of it within t, change the state of this car to DECELERATION.

The distance of a car runs in time t is:

$$\mathrm{s} = \frac{1}{2}at^2 + vt$$

$$\begin{aligned} ControlSystem \equiv & \begin{pmatrix} ?x_1 - x_2 < l \& s_1 - s_2 > x_1 - x_2; state = ACCELERATION; \\ \hfill a = b; \upsilon' = a; x' = \upsilon \end{pmatrix} \\ & \cup \\ & \begin{pmatrix} ?x_1 - x_2 < l \& s_1 - s_2 < x_1 - x_2; state = DECELERATION; \\ \hfill a = b_2; \upsilon' = a; x' = \upsilon; ?\upsilon > 0 \end{pmatrix} \\ & \\ & (?x_1 - x_2 \geq l; state = CONSTANTSPEED; a = 0;) \\ & \\ & (?state = SETOUT; a = b; \upsilon' = a; x' = \upsilon) \end{aligned} \tag{1}$$

Signal System

The SignalSystem works in this way: process Car sends its data to process of CarSend; process CarSend sends this data to CtrlReceive; CtrlReceive sends this data to ControlSystem. Above is the way of transmitting car's data to the ControlSystem. After computing on this data, ControlSystem sends its command for the car, which is also data like: a, v, state and x. The sequence is: ControlSystem sends its command to process CtrlSend; CtrlSend sends its command to process CarReceive; CarReceive sends this data to Car.

Formula 2 models the processing of a car, which accepts a call and sends the information out. Formula 3 delivers the command that the length is 5 to the car. Formula 4 receives the command that the length is 5 from control system. Formula 5 models the control system. Formula 6 receives commands of control system. Formula 7 receives information of car.

The formulas are below:

$$\begin{aligned}&Car = left?call \rightarrow right!x \rightarrow right!a \rightarrow right!v \rightarrow right!ID \rightarrow \\ &\quad right!state \rightarrow Car\end{aligned} \tag{2}$$

$$\begin{aligned}CarSend &= P\langle\rangle \\ &where \\ P_s &= right!s \rightarrow P\langle\rangle if \# s = 5 \\ P_s &= left?x \rightarrow P_{s\cap\langle x\rangle}\end{aligned} \tag{3}$$

And, x means the variable from Car, that is, x, a, v, ID and state.

$$\begin{aligned}Car\ Receive =& P\langle\rangle where \\ & P\langle\rangle = left!s \rightarrow P\{s\} \\ & P\langle x\rangle = right!x \rightarrow P\langle\rangle \\ & P\langle x\rangle^{smallfrown} s = right!x \rightarrow P_s \\ & left = \{s | s \in right * \# s = 5\}\end{aligned} \tag{4}$$

$$\begin{aligned}ControlSystem =& left?x \rightarrow left?a \rightarrow left?v \rightarrow left?ID \rightarrow left?state \rightarrow \\ & right!call \rightarrow right!s \rightarrow ControlSystem \\ & where \# s = 5\end{aligned} \tag{5}$$

$$CtrlSend = left?call \rightarrow left?s \rightarrow right!call \rightarrow right!sCtrlSend \tag{6}$$

$$Ctrl\mathrm{Receive} = left?call \rightarrow left?s \rightarrow right!call \rightarrow CtrlReceive \tag{7}$$

Car

Car has its attributes: x for assigning it's current position, v for assigning its current velocity, a for assigning its acceleration value, state for assigning its state value and ID for assigning its identification value. Car send its information which is kept into the attributes x, v, a, state and ID, to the ControlSystem via Signal-System. ControlSystem analyze these data using its algorithm and send the new data, which we can call it command, to the Car. Car set its attributes with the new data.

```
model Car
  input Real a_new;
  //Acceleration value from the Control System
  parameter Real x;
  //Current position of a car
  parameter Real v;
  //Current velocity of a car
  parameter Real a;
  //Current acceleration of a car
  parameter String state;
  //Current state of a car
  String ID;
  //The identification of a car
equation
          if a_new > a then
          state = Accelerating(a_new);
   elseif a_new < a then
          state = Deccelerating(a_new);
   end if;
   if a_new < 0 then
                  state = Parking(a_new, state);
   else
              state = SETOUT(a_new, state);
   end if;
end Car;
```

Conclusion

In this paper, we use three modeling methods: Differential Dynamic Logic dL, CSP and Modelica to model a traffic system. The Differential Dynamic Logic dL reasons about possible behaviour of a complex system with temporal logic for reasoning about the temporal behaviour during their operation and first-order dynamic logic. The dL logic supports both discrete and continuous evolution. CSP is used to verify the concurrent processes of a system. And, Modelica can model the physical world properly. Modelica with its object-oriented, equation-based properties can be useful for the analyzing a system. We use dL to model the ControlSystem, CSP to model the Signal System and Modelica for the Car, which is a vehicle in our system.

Acknowledgments This work is supported by national high technology research and development program of China (No. 2011AA010101), national basic research program of China (No. 2011CB302904), the national science foundation of China under grant No. 61173046, No. 61021004, No. 61061130541), doctoral program foundation of institutions of higher education of China (No. 200802690018), national; science foundation of Guangdong province under grant No. S2011010004905.

References

1. Platzer A (2008) A differential dynamic logic for hybrid systems. J Autom Reas 41(2):143–189
2. Harel D, Kozen D, Tiuryn J (2000) Dynamic logic. MIT, Cambridge
3. Brookes SD, Hoare CAR, Roscoe AW (1984) A theory of communicating sequential processes. J ACM 31(3):560–599
4. Fritzson P, Engelson V (1998) Modelica—a unified object-oriented language for system modeling and simulation. ECOOP'98—Object-Oriented Programming, pp 67–90

Specification of Railway Cyber Physical Systems Using AADL

Lichen Zhang

Abstract Railway cyber physical systems involve interactions between software controllers, communication networks, and physical devices. These systems are among the most complex cyber physical systems being designed by humans, but the complexities of railway cyber physical systems make their development a significant technical challenge. Various development technologies are now indispensable for quickly developing safe and reliable transportation systems. In this paper, we apply AADL to specify railway cyber physical systems and give a detailed analysis and design of the CBTC system. The CBTC system is split into four subsystems and makes friendly communication between the other three subsystems connecting to the data communication subsystem. We apply AADL to model each subsystem and give a detailed analysis and modeling, and make an effective integration of all subsystems together to form a complete CBTC system finally.

Keywords Railway cyber physical systems · AADL · Specification · CBTC

Introduction

The problems that must be addressed in operating a railway are numerous in quantity, complex in nature, and highly inter-related [1–3]. For example, collision and derailment, rear-end, head-on and side-on collisions are very dangerous and may occur between trains. Trains may collide at level crossings. Derailment is caused by excess speed, a wrong switch position and so on. The purpose of train control is to carry the passengers and goods to their destination, while preventing

L. Zhang (✉)
Shanghai Key Laboratory of Trustworthy Computing, East China Normal University,
Shanghai 200062, China
e-mail: zhanglichen1962@163.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_45, © Springer Science+Business Media Dordrecht 2014

them from encountering these dangers. Because of the timeliness constraints, safety and availability of train systems, the design principles and implementation techniques adopted must ensure to a reasonable extent avoidance of design errors both in hardware and software. The train to train collision accident that happened on July 23, 2011 in one of the high speed lines gave a big hit to the high speed railway development in China. Besides a great surprise, everybody is eager to know what has happened, what went wrong, whose responsibility it is. The accident investigation report published in December 2011 described the events and the software and hardware failures of the train control system equipments. Thus, a specific methodology relevant, to design should be applied for train control systems development. The dependability of the railway cyber physical system should arouse more attention [4, 5].

The development of railway cyber physical systems is a challenging process. On the one hand, the railway domain experts have to make the requirement analysis for the railway cyber physical systems in such a way that they are implementable. On the other hand, the software engineer has to understand these domain-specific requirements to be able to implement them correctly. The high need for product quality is beyond dispute as human life may be endangered if a railway controller is malfunctioning. The struggle for high-quality software development methods is of highest importance in railway cyber physical area.

The SAE Architecture Analysis and Design Language (AADL) [6–8] defines a language for describing both the software architecture and the execution platform architectures of performance-critical, embedded, real-time systems. An AADL model describes a system as a hierarchy of components with their interfaces and their interconnections. Properties are associated to these constructions. AADL components fall into two major categories: those that represent the physical hardware and those representing the application software. The former is typified by processors, buses, memory, and devices, the latter by application software functions, data, threads, and processes. The model describes how these components interact and are integrated to form complete systems. It describes both functional interfaces and aspects critical for performance of individual components and assemblies of components. The changes to the runtime architecture are modeled as operational modes and mode transitions.

In this paper, we propose an approach to specify railway cyber physical systems based on AADL.

The Proposed Method for Specifying Railway Cyber Physical Systems Based on AADL

AADL [7–12] is an architecture description language developed to describe embedded systems is shown in Fig. 1. Architecture Analysis and Design Language (AADL), which is a modeling language that supports text and graphics, was

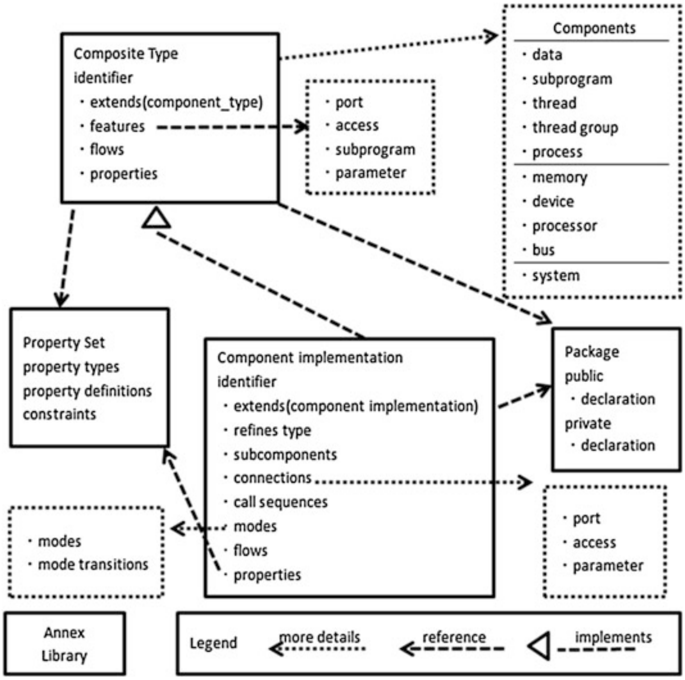


Fig. 1 AADL elements

approved as the industrial standard AS5506 in November 2004. Component is the most important concept in AADL. The main components in AADL are divided into three parts: software components, hardware components and composite components. Software components include data, thread, thread group, process and subprogram. Hardware components include processor, memory, bus and device. Composite components include system (Fig. 2).

In its conformity to the ADL definition, AADL provides support for various kinds of non-functional analyses along with conventional modeling [14].

Flow Latency Analysis: Understand the amount of time consumed for information flows within a system, particularly the end-to-end time consumed from a starting point to a destination.

Resource Consumption Analysis: Allows system architects to perform resource allocation for processors, memory, and network bandwidth and analyze the requirements against the available resources.

Real-Time Schedulability Analysis: AADL models bind software elements such as threads to hardware elements like processors. Schedulability analysis helps in examining such bindings and scheduling policies.

Safety Analysis: Checks the safety criticality level of system components and highlights potential safety hazards that may occur because of communication among components with different safety levels.

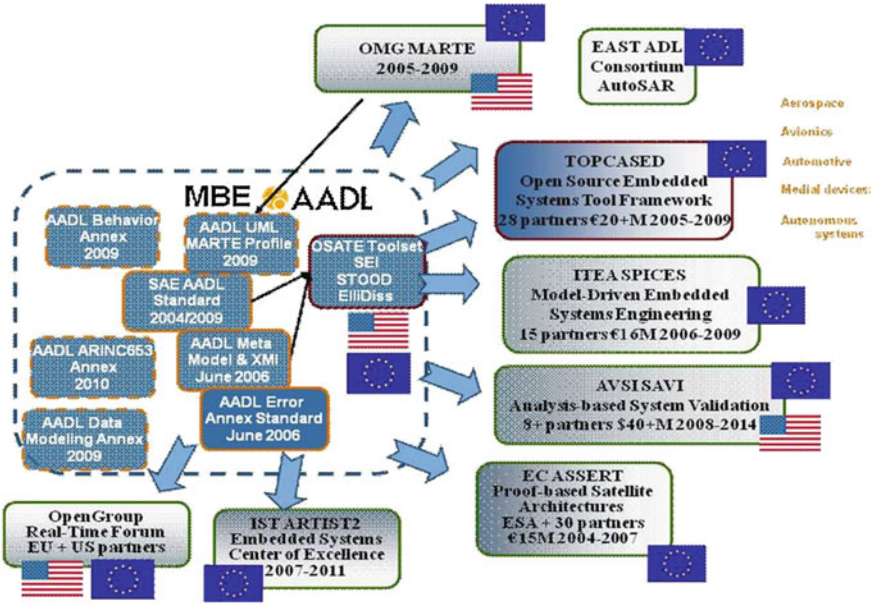

OMG MARTE
2005-2009
EAST ADL
Consortium
AutoSAR
Aerospace
Avionics
Automotive
Medial devices:
Autonomous systems
TOPCASED
Open Source Embedded Systems Tool Framework
28 partners €20+M 2005-2009
MBE AADL
AADL Behavior Annex 2009
AADL UML MARTE Profile 2009
OSATE Toolset SEI
STOOD
ElliDiss
SAE AADL Standard 2004/2009
ITEA SPICES
Model-Driven Embedded Systems Engineering
15 partners €16M 2006-2009
AADL ARINC653 Annex 2010
AADL Meta Model & XMI June 2006
AVSI SAVI
Analysis-based System Validation
8+ partners $40+M 2008-2014
AADL Error Annex Standard June 2006
AADL Data Modeling Annex 2009
EC ASSERT
Proof-based Satellite Architectures
ESA + 30 partners
€15M 2004-2007
OpenGroup
Real-Time Forum
EU + US partners
IST ARTIST2
Embedded Systems Center of Excellence
2007-2011

Fig. 2 Industry initiatives utilizing SAE AADL [13]

Security Analysis: Like safety levels, AADL components can be assigned various security levels. The analysis helps in identifying the security loopholes that may happen because of mismatches in security levels between a component and its subcomponents, and communication among components with different security levels.

AADL defines two main extension mechanisms: property sets and sublanguages (known as annexes). It is possible to extend the AADL concepts either by introducing new properties to the modeling elements, by addition of new modeling notations, or by developing a sublanguage as annex to the AADL standard Properties are label-value pairs used to annotate components. These properties can be grouped into named sets. These sets are then used in analysis tools that process AADL models to be able to verify characteristics of the modeled system. One example of a property set is the standard property set for RMA that contains period, execution time, and deadline among other properties. These properties are associated with threads to be able to derive a real-time task set amenable of timing analysis. Sublanguages, on the other hand, enable the encoding of complex statements about components for which syntactic verification makes sense. The syntax of the language is defined inside the annex that implements the language. Annexes and properties allow the addition of complex annotations to AADL models that accommodate the needs of multiple concerns. These annotations, along with their corresponding analysis plug-ins, provide a powerful combination for the architect to evaluate his/her design choices from different perspectives.

The extension mechanisms in AADL enable these perspectives to evolve in number and complexity as the knowledge on them also evolves [15, 16].

In this paper, we propose an approach to specify railway cyber physical systems based on AADL:

(1) Specify the physical world of railway cyber physical systems with Modelica [17–19].
(2) Combine SysML and AADL for the design, validation and implementation of railway cyber physical systems [20].
(3) Extend AADL to express the dynamic continous features by proposing new AADL annex.
(4) Specify spatial–temporal features by Cellular automata (CA) [21, 22].

Case Study: Specifying Communication Based Train Control Systems

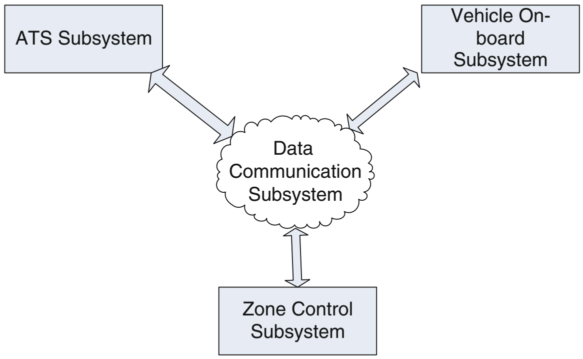
It is known that CBTC system can be effectively split into four subsystems: automatic train supervision subsystem (ATS), zone control subsystem, vehicle on-board subsystem and data communication subsystem. The ATS subsystem, also named center control subsystem, includes a series of servers for different purposes, printers, displays, many workstations and so on. Via the data communication subsystem, the ATS subsystem gets respectively the information of databases, train position, wayside devices and movement authority from the database storage unit in data communication subsystem, vehicle on-board subsystem and zone control subsystem. After handling them, they are transferred to the correlated devices and subsystems. The vehicle on-board subsystem includes the vehicle on-board controller (VOBC), Driver Machine Interface (DMI) and so on. The subsystems accepts many different types of data in zone control subsystem and ATS subsystem, then calculates the train movement curve, measures the train speed and movement distance associating with the guide way databases to protect the safety of train movement. The zone control subsystem contains zone controller, computer interlocking (CI) devices, axle counter, signals, platform doors, switches and other wayside devices. The subsystem gets and handles the useful data and statue information from other subsystems to generate the movement authority for the trains in control zones and update persistently as required. Then the subsystem transfers the movement authority to the vehicle on-board subsystem through the data communication subsystem in order to control the train movement. It also controls the switches, signals, and platform doors, and acknowledges the request of adjacent zone controller. The data communication subsystem mainly contains database storage unit, backbone fie-optical network, wayside access points, on-board wireless units and network switches. Data communication subsystem is the other subsystems communication bridge making normal communication between the subsystems and ensuring the safety of train operation. As shown in Fig. 3, the communications between the subsystems are via the data communication subsystem.

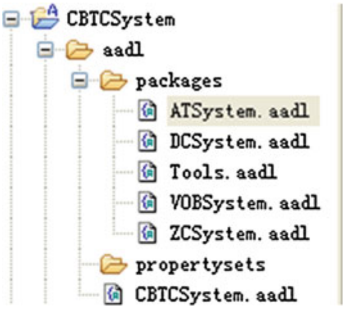
In OSATE environment, we will use the AADL to give the design and modeling of the CBTC system. The CBTC system's file structure is shown in Fig. 4.

Each subsystem corresponds to a file whose name ends with aadl. In view of the repeated use of some tools, the system provides a common tool file (Tools.aadl). Four subsystems are combined to form a complete CBTC system through a file (CBTCSystem.aadl). The four subsystems are achieved by using the system components in AADL, and the connections between the subsystems use the bus components. In AADL, components contain component type and component implementation. Component type specifies the external interfaces of component implementation, and component implementation describes the internal structure of the component. Each component implementation corresponds to a component type, a component type can have zero or more component implementations. The AADL code of CBTC system overall framework is as follows:

```
system CBTCSystem
end CBTCSystem;
system implementation CBTCSystem.Impl
  subcomponents
    ATSys: system ATSystem::ATSys.Impl;
    ZCSys: system ZCSystem::ZCSys.Impl;
    VOBSys: system VOBSystem::VOBSys.Impl;
    DCSys: system DCSystem::DCSys.Impl;
  connections
    conn1: bus access ATSys.toDCS -> DCSys.fromATS;
    conn2: bus access ZCSys.toDCS -> DCSys.fromZC;
    conn3: bus access VOBSys.ToDCS -> DCSys.fromVOBS;
    conn4: bus access DCSys.toATS -> ATSys.fromDCS;
    conn5: bus access DCSys.toZC -> ZCSys.fromDCS;
conn6: bus access DCSys.toVOBS -> VOBSys.FromDCS;
…
end CBTCSystem.Impl;
```

Fig. 3 The communication between the subsystems





```
CBTCSystem
    aadl
        packages
            ATSystem.aadl
            DCSystem.aadl
            Tools.aadl
            VOBSystem.aadl
            ZCSystem.aadl
        propertysets
        CBTCSystem.aadl
```

Fig. 4 CBTC system's file structure

As can be seen from the above code, the CBTC system contains four subsystems, data communication subsystem is a bridge of communication between the other three subsystems. OSATE will automatically generate an aaxl file for each aadl file, we can take advantage of aaxl file to new a corresponding graphics file (aaxldi file). The file lists the graphical representation of all components. And each implementation of the system components will generate a system instance diagram. The system instance diagram of the CBTC system implementation (CBTCSystem.Impl) is shown in Fig. 5.

The dynamic continuous features of the train is specified by AADL annexb as follows:

```
Device train
    Features
        In_data :in data port;
        Out_data: out data port;
    Properties
        Equation => {
            $F = 300 - 0.284 v_t$
            $W_0 = A + B v_t + C v_t^2$ ;
            $W_i = 1000 \tan\theta$ ;
            $W_r = \frac{600}{R}$ ;
            $W_s = 0.00013 L_s$ ;
            $W_f = W_i + W_r + W_s$;
            $W1 = W_f * L_t$;
          B=M ;
          $a = \frac{F - W_0 - W_1 - B}{(1+\gamma)M}$ ;
        };
       Const =>{M, γ};
        Const_value =>{M,γ};
        Var=>{vt,A,B,C,R,Ls,Lt, θ};
    End Train;
```

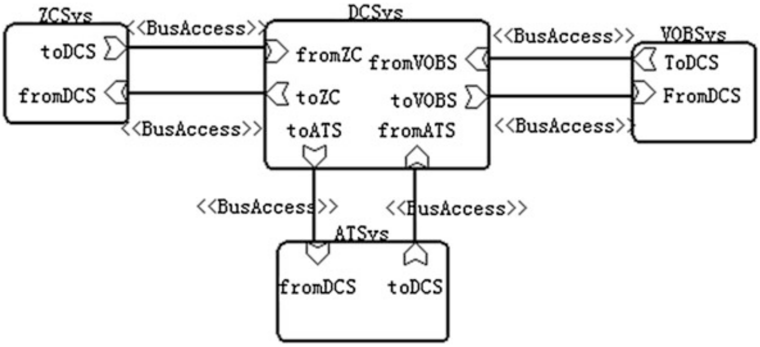
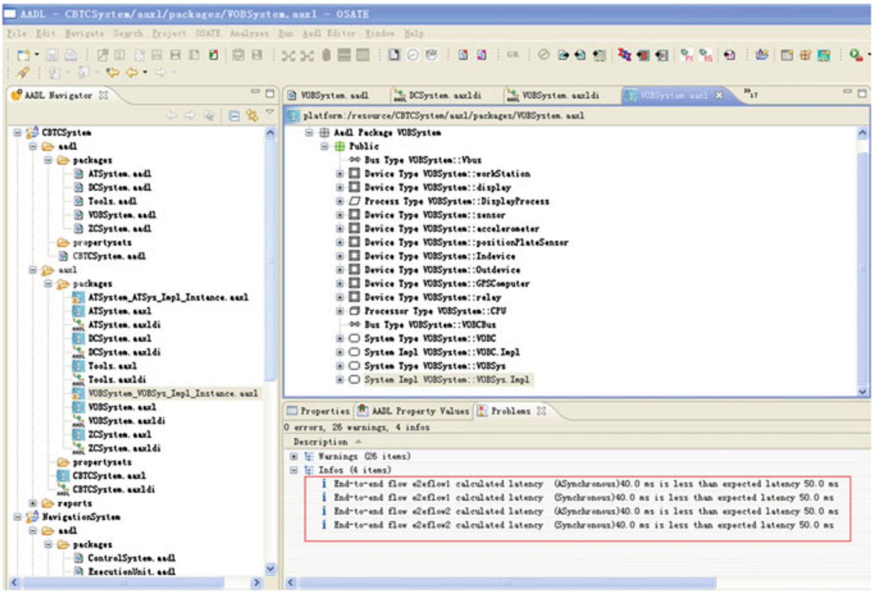


Fig. 5 System instance diagram of the of CBTC system implementation


AADL - CBTCSystem/aaxl/packages/VOBSystem.aaxl - OSATE
File Edit Navigate Search Project OSATE Analyses Run Aadl Editor Window Help
AADL Navigator
CBTCSystem
aadl
packages
ATSystem.aadl
DCSystem.aadl
Tools.aadl
VOBSystem.aadl
ZCSystem.aadl
propertysets
CBTCSystem.aadl
aaxl
packages
ATSystem_ATSys_Impl_Instance.aaxl
ATSystem.aaxl
ATSystem.aaxldi
DCSystem.aaxl
DCSystem.aaxldi
Tools.aaxl
Tools.aaxldi
VOBSystem_VOBSys_Impl_Instance.aaxl
VOBSystem.aaxl
VOBSystem.aaxldi
ZCSystem.aaxl
ZCSystem.aaxldi
propertysets
CBTCSystem.aaxl
CBTCSystem.aaxldi
reports
NavigationSystem
aadl
packages
ControlSystem.aadl
ExecutionUnit.aadl
VOBSystem.aadl
DCSystem.aaxldi
VOBSystem.aaxldi
VOBSystem.aaxl
platform:/resource/CBTCSystem/aaxl/packages/VOBSystem.aaxl
Aadl Package VOBSystem
Public
Bus Type VOBSystem::Vbus
Device Type VOBSystem::workStation
Device Type VOBSystem::display
Process Type VOBSystem::DisplayProcess
Device Type VOBSystem::sensor
Device Type VOBSystem::accelerometer
Device Type VOBSystem::positionPlateSensor
Device Type VOBSystem::Indevice
Device Type VOBSystem::Outdevice
Device Type VOBSystem::GPSComputer
Device Type VOBSystem::relay
Processor Type VOBSystem::CPU
Bus Type VOBSystem::VOBCBus
System Type VOBSystem::VOBC
System Impl VOBSystem::VOBC.Impl
System Type VOBSystem::VOBSys
System Impl VOBSystem::VOBSys.Impl
Properties
AADL Property Values
Problems

```
0 errors, 26 warnings, 4 infos
Description
Warnings (26 items)
Infos (4 items)
    End-to-end flow e2eflow1 calculated latency  (ASynchronous)40.0 ms is less than expected latency 50.0 ms
    End-to-end flow e2eflow1 calculated latency  (Synchronous)40.0 ms is less than expected latency 50.0 ms
    End-to-end flow e2eflow2 calculated latency  (ASynchronous)40.0 ms is less than expected latency 50.0 ms
    End-to-end flow e2eflow2 calculated latency  (Synchronous)40.0 ms is less than expected latency 50.0 ms
```

Fig. 6 End to end flow analysis of VOBS

The spatial requirements of railway cyber physical systems is specified as follows:

```
    dn: in parameter Types::float;
    dmin: in parameter Types::float;
    ltrain: in parameter Types::float;
    at: in parameter Types::float;
    bt: in parameter Types::float;
    vmax: in parameter Types::float;
    Xc: in parameter Types::float;
    Dgap: in parameter Types::float;
    vn: in parameter Types::float;
    xn: in parameter Types::float;
    vn: out parameter Types::float;
    xn: out parameter Types::float;
end space;
subprogram implementation space.default
    annex space_specification {**
states
    s1: initial state;
    s2:return state;
transitions
    normal: s1 -[ on 加速]-> s2 {
        if(isTrain){
              if(dn-ltrain>dmin){
                     vn=min(vn+at,vmax);
            }else if(dn-ltrain<dmin){
              vn=max(vn-bt,0);
            }else{
              vn=vn;
            }
            xn=xn+vn;
        }else{
              if( Dgap>Xc){
                     vn=min(vn+at,vmax);
            }else if( Dgap>Xc){
              vn=max(vn-bt,0);
            } else{
              vn=vn;
            }
            xn=xn+vn;
        }
    };
    normal: s1 -[on 减速 ]-> s2 {
        if(isTrain){
                vn=min(vn,dn-ltrain-1);
              xn=xn+vn;
          }else{
                vn=min(vn,Dgap);
              xn=xn+vn;
          }
      } ;
  **};
  end space.default;
```

The end to end flow analysis of VOBS is shown as Fig. 6.

Conclusion

In this paper, we proposed an approach to specify railway cyber physical systems based on AADL. We illustrated the proposed method by the specification of the communication based train control systems. The specification process of the communication based train control systems demonstrated the extension of AADL approach that can be used for modeling complex railway cyber physical system, effectively reduce the complexity of software development.

The further work is devoted to developing tools to support the automatic generation of model and code.

Acknowledgments This work is supported by Shanghai Knowledge Service Platform Project (No. ZF1213), national high technology research and development program of China (No. 2011AA010101), national basic research program of China (No. 2011CB302904), the national science foundation of China under grant (No. 61173046, No. 61021004, No. 61061130541, No. 91118008), doctoral program foundation of institutions of higher education of China (No. 20120076130003),national science foundation of Guangdong province under grant (No.S2011010004905).

References

1. IEC62278:2002 Railway applications: specification and demonstration of reliability, availability, maintainability and safety (RAMS)
2. IEC62279:2002 Railway applications: communications, signaling and processing systems–Software for railway control and protection systems
3. IEC62280:2002 Railway applications: communication, signaling and processing systems – Safety related electronic systems for signaling
4. Laprie C (1992) Dependability: basic concepts and terminology. Springer, Berlin
5. Svizienis A, Laprie JC, Randell B (2000) Dependability of computer systems: fundamental concepts, terminology, and examples. Technical report, LAAS-CNRS
6. Feiler PH, Gluch DP, Hudak JJ (2006) The architecture analysis and design language (AADL): an introduction. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST
7. Feiler PH, Lewis B, Vestal S et al (2005) An overview of the SAE architecture analysis and design language (AADL) standard: a basis for model-based architecture-driven embedded systems engineering. Springer, US, pp 3–15 (Architecture Description Languages)
8. Hudak JJ, Feiler PH (2007) Developing aadl models for control systems: a practitioner's guide
9. Feiler PH, Gluch DP (2012) Model-based engineering with AADL: an introduction to the SAE architecture analysis and design language. Addison-Wesley Professional
10. SAE AS-2C (2012) Architecture analysis and design language. SAE international document AS5506B(2012) Revision 2.1 of the SAE AADL standard
11. Delange J (2012) Towards a model-driven engineering software development framework. In: The third analytic virtual integration of cyber-physical systems workshop, 04 Dec 2012, Porto Rico
12. Feiler P, Hugues J, Sokolsky P (Eds) (2012) Oleg Architecture-driven semantic analysis of embedded systems. Dagstuhl Seminar 12272. Dagstuhl Report, 2(7):30–55. ISSN 2192-5283

13. The story of AADL (2012) AADL Wiki. Software Engineering Institute, 2010. Web. 06 Jan 2012
14. Muhammad N, Vandewoude Y, Berbers Y, van Loo S Modelling embedded systems with AADL: a practical study. www.intechopen.com/download/pdf/10732
15. de Niz D, Feiler PH Aspects in the industry standard AADL. In: AOM '07 Proceedings of the 10th international workshop on aspect-oriented modeling. pp 15–20
16. Michotte L, Vergnaud T, Feiler P, France R (2008) Aspect oriented modeling of component architectures using AADL. In: Proceedings of the second international conference on new technologies, mobility and security, 5–7 Nov 2008
17. Modelica Association (2002) Modelica—a unified object-oriented language for physical systems modelling. Language specification. Technical report
18. Modelica Association (2007) Modelica: A unified object- oriented language for physical systems modeling: language specification version 3.0. www.modelica.org
19. OMG OMG unified modeling language TM (OMG UML). Superstructure Version 2.2, February 20
20. De Saqui-Sannes P, Hugues J (2012) Combining SysML and AADL for the design, validation and implementation of critical systems. In: ERTSS 2012 (Embedded Real Time Software and Systems), Toulouse, France, 01–03 Feb 2012
21. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. Phys I France 2(12):2221–2229
22. Culik K, Hurd LP (1990) Formal languages and global cellular automaton behavior. Phys D 45(13):396–403

Formal Specification of Railway Control Systems

Bingqing Xu and Lichen Zhang

Abstract Train control systems must provide a high level of safety as they are a very important component and responsible for the safe operation of a train. To meet safety and reliability requirements, formal techniques must be used to specify train control systems. In this paper, we uses CSP, Object-Z and Clock to specify the Railway Control System concerning both the linear track and crossing area, especially the time delay between any two aspects of the railway system.

Keywords Railway control systems · Object-Z · CSP · Clock theory · Formal specification

Introduction

Train control systems must provide a high level of safety as they are a very important component and responsible for the safe operation of a train. To meet safety and reliability requirements, the relative international standards recommend the application of formal methods in specifying development specifications and design for train control systems [1, 2]. Complicated system such as railway control system is a system with many complex behavioral aspects. And the mechanism of communication between different aspects is hard to define. With help of formal methods, we now find a way to construct a detailed specification of each aspect and the link mechanism among various aspects. While a communication mechanism is not enough to describe the state change and data change in the system. Above all, the author tends to use Communicating Sequential Processes (CSP) [3] to specify the communication part of the Railway Control System. Concerning the

B. Xu · L. Zhang (✉)
Shanghai Key Laboratory of Trustworthy Computing, East China Normal University,
Shanghai 200062, China
e-mail: zhanglichen1962@163.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_46, © Springer Science+Business Media Dordrecht 2014

time characteristics in the system, Clock [4–7] specifies the system time requirements better. For the state and data changes, Object-Z is ideal for analysis in data change in a schema box form.

In this paper, we uses CSP, Object-Z and Clock to specify the Railway Control System concerning both the linear track and crossing area, especially the time delay between any two aspects of the railway system.

Relative Works

Hoenicke [8–12] uses a combination of three techniques for the specification of processes, data and time: CSP, Object-Z and Duration Calculus. The basic building block in our combined formalism CSP-OZ-DC is a class. First, the communication channels of the class are declared. Every channel has a type which restricts the values that it can communicate. There are also local channels that are visible only inside the class and that are used by the CSP, Z, and DC parts for interaction. Second, the CSP part follows; it is given by a system of (recursive) process equations. Third, the Z part is given which itself consists of the state space, the Init schema and communication schemas. For each communication event a corresponding communication schema specifies in which way the state should be changed when the event occurs. Finally, below a horizontal line the DC part is stated. The combination is used to specify parts of a novel case study on radio controlled railway crossings. Johannes Faber formally specifies a part of the European Train Control System (ETCS) with the specification language CSPOZ-DC treating the handling of emergency messages.

Gnesi et al. [13] described an important experiment in formal specification and validation, both performed in the context of an industrial project jointly performed by Ansaldobreda Segnalamento Ferroviario and CNR Institutes IEI and CNUCE of Pisa. Within this project they developed two formal models of a control system which is part of a wider safety–critical system for the management of medium-large railway networks. Each model describes different aspects of the system at a different level of abstraction. On these models they performed verification of both safety properties—in the hypothesis of Byzantine errors or in presence of some defined hardware faults—and liveness properties of a dependable communication protocols. The properties has been specified by means of assertions and temporal logical formulae. As a specification language we used Promela language while the verification was performed using the model checker Spin.

Hyun-Jeong Jo, Yong-Ki Yoon and Jong-Gyu Hwang proposed an eclectic approach to incorporate Z (Zed) formal language and ‘Statemate MAGNUM’, formal method tools using Statechart. Also they applied the proposed method to train control systems for the formal requirement specification and analyzed the specification results [1].

In Zafar paper [14], formal methods which are advanced software engineering techniques, in term of Z notation, are applied for the specification of critical components of automated train control system. At first graph theory is used for modeling of static components of the system and then integrated with Z notation to describe its entire state space. At first real topology is transferred to model topology in graph theory and then switches, crossings, and level crossing are formalized. At the end, these components are composed to define the entire interlocking system. Formal specification of the system is described in Z notation and the model is analyzed using Z/EVES tool.

Peleska et al. [15] motivate and illustrate an approach for domain-specific specification languages in the field of small local railway or tramway control systems. Adopting terms and concepts from the application domain, such languages are the ideal means of communication between users, system engineers and control computer specialists. Semantic rigour is provided by a transformation of the domain-specific representation into wide-spectrum formal specification languages. These transformations offer the possibilities of formal verification and testing against formal specifications, as well as automatic generation of executable programs from specifications. The construction of the transformation can be substantially simplified if the class of control systems is formalised in a generic way using design patterns and frameworks.

Haxthausen and Peleska [16] introduce the concept for a distributed railway control system and present the specification and verification of the main algorithm used for safe distributed control. Their design and verification approach is based on the RAISE method, starting with highly abstract algebraic specifications which are transformed into directly implementable distributed control processes by applying a series of refinement and verification steps. Concrete safety requirements are derived from an abstract version that can be easily validated with respect to soundness and completeness. Complexity is further reduced by separating the system model into a domain model and a controller model. The domain model describes the physical system in absence of control and the controller model introduces the safety-related control mechanisms as a separate entity monitoring observables of the physical system to decide whether it is safe for a train to move or for a point to be switched.

Guo Xie, Akira Asano, Sei Takahashi, Hideo Nakamura presents a formal specification of an Automatic Train Protection and Block (ATPB) model for local line railway system [17] and validates the model by internal consistency proving and systematic testing. The system consists of two parts, the on-board subsystem and ground subsystem. The former is to detect the basic state of train, such as position, speed and integrity, monitor the speed, communicate with ground equipment and record the relative events. And the latter is responsible for communicating with train, controlling the route and interlocking, and decision-making for train operation adjustment. The main purpose of this project is to improve the efficiency and guarantee that there is no collision, no derailment and no over speeding at the same. The formal language used in this project is VDM ++. And the state and specification of operation are all checked and validated using VDMTools. The results confirm the correctness of this system and the model throws new light on practical system design.

Bernardeschi et al. [18] outline an experience on formal specification and verification carried out in a pilot project aiming at the validation of a railway computer based interlocking system. Both the specification and the verification phases were carried out in the Just Another Concurrency Kit (JACK) integrated environment. The formal specification of the system was done by means of process algebra terms. The formal verification of the safety requirements was done first by giving a logical specification of such safety requirements, and then by means of model checking algorithms. Abstraction techniques were defined to make the problem of safety requirements validation tractable by the JACK environment.

The European Train Control System (ETCS) is a control system for the interoperability of the railways across Europe. A. Chiappini et al. report on the activities of the EuRailCheck project, promoted by the European Railway Agency [19], for the development of a methodology and tools for the formalization and validation of the ETCS specifications. Within the project, we achieved three main results. First, they developed a methodology for the formalization and validation of the ETCS specifications. The methodology is based on a three-phases approach that goes from the informal analysis of the requirements, to their formalization and validation. Second, they developed a set of support tools, covering the various phases of the methodology. Third, they formalized a realistic subset of the specification in an industrial setting. The results of the project were positively evaluated by domain experts from different manufacturing and railway companies.

Constance Heitmeyer and Nancy Lynch give a new solution to the Generalized Railroad Crossing problem [20], based on timed automata, invariants and simulation mappings, is presented and evaluated. The solution shows formally the correspondence between four system descriptions: an axiomatic specification, an operational specification, a discrete system implementation, and a system implementation that works with a continuous gate model.

Formal Specification of Train Control Systems

The Problem that must be addressed in operating a railway are numerous in quantity, complex in nature, and highly inter-related. For example, collision and derailment, rear-end, head-on and side-on collisions are very dangers and may occur between trains. Trains collide at level crossing. Derailment is caused by excess speed, wrong switch position and so on. The purpose of train control is to carry the passengers and goods to their destination, while preventing them from encountering these dangers. Because of the timeliness constraints, safety and availability of train systems, the design principles and implementation techniques adopted must ensure to a reasonable extent avoidance of design errors both in hardware and software. Thus, a formal technique relevant to design should be applied for train systems development. The purpose of our exercise is to apply aspect -oriented formal method s to develop a controller for train systems that tasks as input: a description of track configuration and a sequence of description of the moves of each of these trains.

The controller should take care of trains running over the track. It should control the safety of the configuration, i.e. No two trains may enter the critical section. When one critical section is occupied, some others, which share some part of section with this one, should be locked. The controlled can control the status, speed, position of trains.

In order to keep the description focused, we concentrate on some particular points in train control systems rather than the detailed descriptions of all development process. The specification is made by integrating Object-Z, CSP and Clock Theory.

Clock theory [4–7] puts forward the possibility to describe the event in physical world by using a clock, and can analyze, records the event by clock. To use clock to specify Cyber Physical Systems, the time description is clearer to every event and can link continuous world with discrete world better, the definition and linking mechanism of clock theory is provided as below.

Definition 1 *A clock c is an increasing sequence of real numbers. We define its low and high rates by*

$$\begin{aligned}&\mathit{\Delta}(c) = df \quad \boldsymbol{inf}\{(c[i+1] - c[i])|i \in \\ &Nat\}\nabla(c) = df \ \ \boldsymbol{sup}\{(c[i+1] - c[i])|i \in \\ &Nat\}\end{aligned}$$

Here, c [1] and c′ stand for the first element of c, and the resultant sequence after removal of c[1] from c respectively.

If c is a *healthy clock*, it does not speed up infinitely. Then

$$\Delta(c) > 0$$

If c runs faster than d if for all $i\in$ *Nat*

$$c[i] \leq d[i]$$

then the relation of c and d can be denoted by c 主 d and this is a kind of partial order in clock.

Lemma

$$c \preccurlyeq c^I$$

Let c and d be clocks. We define the transition latency between the two clocks as

$$\rho(c,d) = df \quad \mathbf{sup}\{|c[i] - d[i]||i \in Nat\}$$

Lemma

$$\rho(c,d) \geq 0$$

Let e be an event. **clock**(e) denotes the clock that records the time instants when the event e occurs. And we use **clock**(**event**(c)) to denote the event that take place at every time instant c[i].

Definition 2 (Local Clock and Global Clock). *Let l be a label denoting a location, and c a clock. Then l:c denotes clock c that locates at l.*

$$l : c \preccurlyeq l : d = df \quad (c \preccurlyeq d)$$
$$\rho(l : c, l : d) = df\, \rho(c, d)$$

A global clock [l:c] is defined as an equivalent class of local clocks,

$$\rho([l1 : c1], [l2 : c2]) = df \quad \rho(l1 : c1, l2 : c2)$$

Definition 3 (Discrete Variable and Continuous Variable). *Some dynamic features of continuous variable can be described better.*

climb(u,r) is introduced to describe the time instants when the value of u rises up to r.
drop(u,r) is introduced to describe the times instants when the value of u falls below r.

Definition 4 (Linking Mechanism). *Here we assign continuous variables to discrete variables.*

$$x = u \; every \; c \; init \; x0$$

x is a discrete variable, u is a continuous variable, c is a clock with $c[1] \geq 0$*, and* $x0$ *is an initial value. In this equation, we assign the value of u to x at the every instant of clock c.*

A train controller limits the speed of the train, decides when it is time to switch points and secure crossings, and makes sure that the train does not enter them too early as shown in Fig. 1.

To extend state-based and behavioural techniques with real-time aspects, different approaches exist. One approach is unifying a state based language and Timed CSP, an extension of CSP with a time-out operator. In this paper, we take a different approach. We integrate CSP and Object-Z with clock theory to specify real-time aspects as shown in Figs. 2 and 9.

Rail crossing control is modeled as shown in Fig. 3.

Finally, woven model is shown as Fig. 4.

```
TrainController
    position :  [p ! : Position]  chan
    speed :  [s ? : Speed]  chan
    t_now :   [t  ? : Clock]  local_chan
    t_out :   [t ? : Clock]  chan
    t_wait :   [t ? : Clock]  chan
    t_idle :   [t ? : Clock]  chan
    record : [r ! : record ]chan
    state : [st ? : State] chan

    Init
     t_now = Clock_now
     speed = 0
     t_out = 0
     t_wait = 0
     t_idle = 0

    sendPosition                      sendSpeed
     pos ! : Position                  speed ! : Speed

    updateCloc                        crossControl
    k
     Δ(t_out, t_wait,t_idle)           Δ(speed, position)
     t_out ? :Clock                    speed ?: Speed
     t_wait? :Clock                    position  ?: Position
     t_idle? :Clock
                                       speed' = speed |
     t_out'= t_out                              0<speed <max_speed
     t_wait'= t_wait                   position'= Position
     t_idle'= t_idle

    log
     Δ(record)

    record' = record
```

Fig. 1 Model of train controler

ClockController[X]

$\upharpoonleft(Clock_{now}, Clock_{out}, Clock_{wait}, Clock_{idle}, Init, updateClock)$

train$_i$: ***Train***
train$_j$: ***Train***
t_pos_b : ***Track_Position***
t_pos_d : ***Track_Position***
t_now : ***Clock***
t_out : ***Clock***
t_wait : ***Clock***
t_idle : ***Clock***

Init

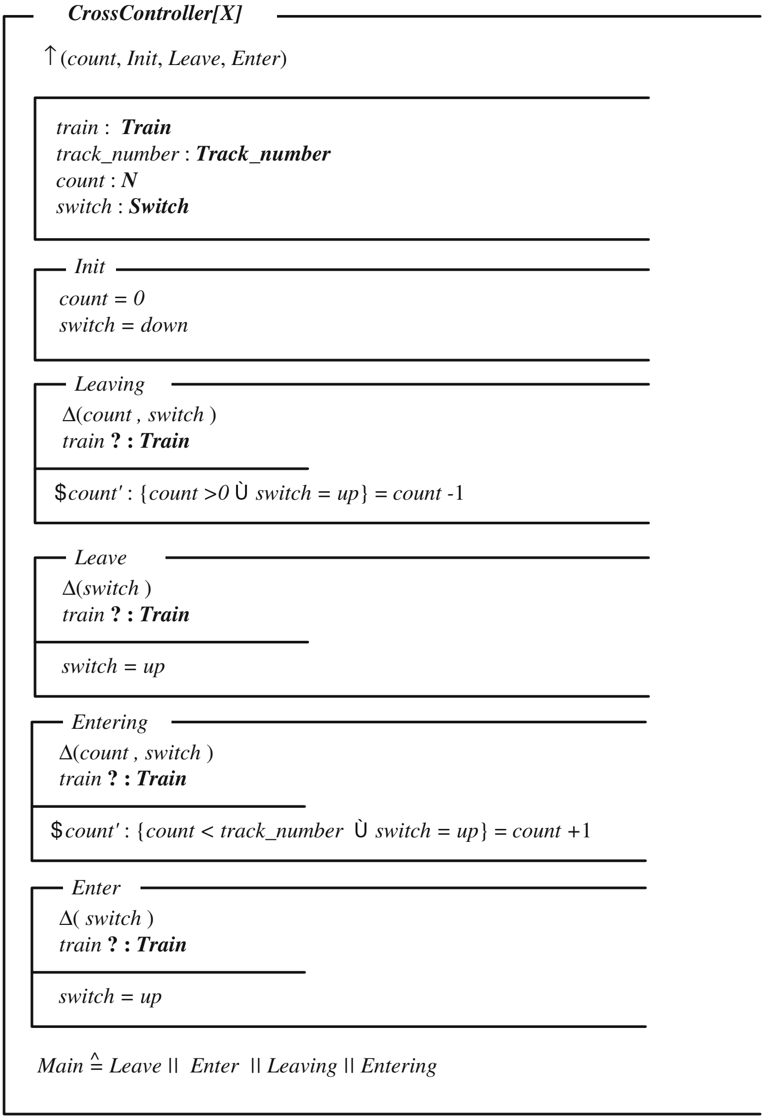
$t\_now = Clock_{now}$
$t\_out = Clock_{out}$
$t\_wait = Clock_{wait}$
$t\_idle = Clock_{idle}$

updateClock

$\Delta(t\_out, t\_wait, t\_idle)$
train$_i$ **? :** ***Train***
train$_j$ **? :** ***Train***
t_pos_b **? :** ***Track_Position***
t_pos_d **? :** ***Track_Position***

$t\_out' := Clock_{now} + Clock_{now} - Clock_{common}$
$| U\_C^{ij}_{fun\_pursued}$ Ú $U\_C^{ij}_{fun\_meet}$
$t\_idle' := Clock_{common} - Clock_{now} | U\_C^{ij}_{fun\_pursued}$ Ú $U\_C^{ij}_{fun\_meet}$
$t\_wait' := Fun\_Track(i, Track\_Position).Clock_{wait} - Clock_{out}$
$+ Clock_{idle}$
$| U\_C^{ij}_{fun\_pursued}$ Ú $U\_C^{ij}_{fun\_meet}$

Fig. 2 Model of clock



CrossController[X]

↾(*count*, *Init*, *Leave*, *Enter*)

train : ***Train***
track_number : ***Track_number***
count : ***N***
switch : ***Switch***

Init
count = 0
switch = down

Leaving
Δ(*count* , *switch*)
train **? :** ***Train***

$*count'* : {*count >0* Ù *switch* = *up*} = *count* -1

Leave
Δ(*switch*)
train **? :** ***Train***

switch = *up*

Entering
Δ(*count* , *switch*)
train **? :** ***Train***

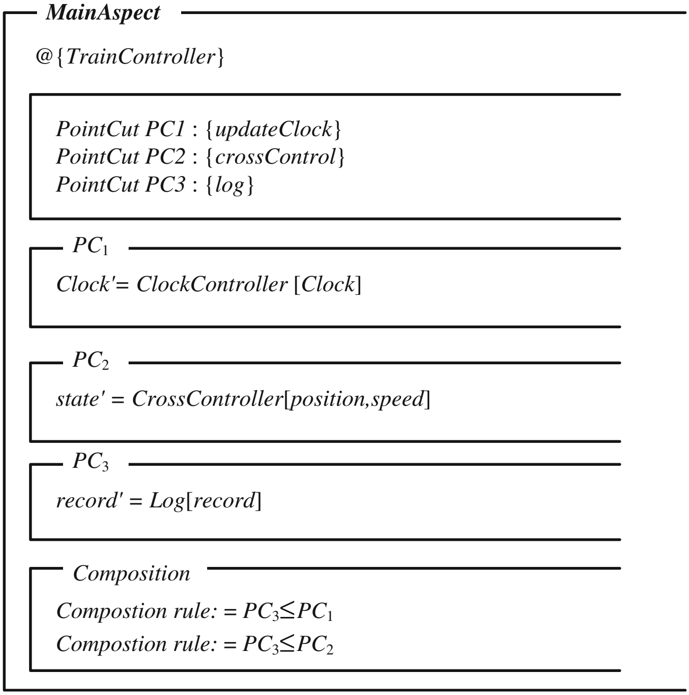
$*count'* : {*count* < *track_number* Ù *switch* = *up*} = *count* +1

Enter
Δ(*switch*)
train **? :** ***Train***

switch = *up*

Main ≙ *Leave* || *Enter* || *Leaving* || *Entering*

Fig. 3 Model of railway crossing control



MainAspect

@{*TrainController*}

PointCut PC1 : {*updateClock*}
PointCut PC2 : {*crossControl*}
PointCut PC3 : {*log*}

PC_1

Clock'= ClockController [*Clock*]

PC_2

state' = CrossController[*position,speed*]

PC_3

record' = Log[*record*]

Composition

Compostion rule: $= PC_3 \leq PC_1$
Compostion rule: $= PC_3 \leq PC_2$

Fig. 4 Woven aspects of diagram

Conclusion

Train control systems must provide a high level of safety as they are a very important component and responsible for the safe operation of a train. To meet safety and reliability requirements, formal techniques must be used to specify train control systems. In this paper, we uses CSP, Object-Z and Clock to specify the Railway Control System concerning both the linear track and crossing area, especially the time delay between any two aspects of the railway system.

Future work focuses on the verification tool development of our proposed method.

Acknowledgments This work is supported by Shanghai Knowledge Service Platform Project (No.ZF1213), national high technology research and development program of China (No.2011AA010101), national basic research program of China (No.2011CB302904), the national science foundation of China under grant (No.61173046, No.61021004, No.61061130541, No.91118008), doctoral program foundation of institutions of higher education of China (No.20120076130003),national science foundation of Guangdong province under grant (No.S2011010004905).

References

1. Jo H-J, Yoon Y-K, Hwang J-G (2009) Analysis of the formal specification application for train control systems. J Electr Eng Technol 4(1):87–92
2. IEC62278:2002 Railway applications: Specification and demonstration of reliability, availability, maintainability and safety (RAMS)
3. Reed GM, Roseoe AW (1986) A timed model for communicating sequential processes. Pro ICALP'86. Lecture notes in computer science. Springer, Berlin
4. He J (2013) A clock-based framework for constructions of hybrid systems. Key talk. In the Proceedings of ICTAC'2013
5. Xu B, He J, Zhang L (2013) Specification of cyber physical systems based on clock theory. Int J Hybrid Inf Technol 6(3):45–54
6. Xu B et al (2013) Specification of cyber physical systems by clock. AST2013. ASTL 20: 111–114, Yeosu, South Korea
7. He J (2012) Link continuous world with discrete world. Shanghai Key Laboratory of Trustworthy Computing East China Normal University, China
8. Hoenicke J Specification of Radio based railway crossings with the combination of CSP, OZ, and DC. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.4394
9. Hoenicke J (2006) Combination of processes, data, and time. PhD thesis, University of Oldenburg
10. Hoenicke J, Maier P (2005) Model-checking of specifications integrating processes, data and time. In: Fitzgerald JS, Hayes IJ, Tarlecki A (eds) FM 2005, volume 3582 of LNCS, Springer, pp 465–480
11. Hoenicke J, Olderog E-R (2002) CSP-OZ-DC: a combination of specification techniques for processes, data and time. Nordic J Comput 9(4):301–334
12. Hoenicke J, E-R Olderog (2002) Combining specification techniques for processes data and time. In: Butler M, Petre L, Sere K (eds) Integrated formal methods, volume 2335 of lecture notes in computer science, Springer, pp 245–266
13. Gnesi S, Latella D, Lenzini G, Amendola A, Abbaneo C, Marmo P (2000) A formal specification and validation of a safety critical railway control system. In: Fifth international workshop on formal methods for industrial critical systems, FMICS 2000, Berlin, Germany, April 3–4 2000
14. Zafar NA (2006) Modeling and formal specification of automated train control system using Z notation. Multitopic conference. INMIC '06. IEEE, pp 438–443, 23–24 December 2006
15. Peleska J Baer A, Haxthausen AE Towards domain-specific formal specification languages for railway control systems.http://www.informatik.uni-bremen.de/agbs/jp/papers/trans2000.html
16. Haxthausen AE, Peleska J (2000) Formal development and verification of a distributed railway control system. IEEE Trans Software Eng 26(8):687–70
17. Xie G, Asano A, Sei Takahashi, Hideo Nakamura, (2011) Study on formal specification of automatic train protection and block system for local line. ssiri-c, Fifth international conference on secure software integration and reliability improvement—companion, pp 35–40
18. Bernardeschi C, Fantechi A, Gnesi S, Larosa S, Mongardi G, Romano D (1998) A formal verification environment for railway signaling system design. Formal Methods Syst Design 12:139–161
19. Chiappini A et al (2010) Formalization and validation of a subset of the European train control system. ICSE'10, Cape Town, South Africa, 2–8 May 2010
20. Heitmeyer C, Lynch N (1994) The generalized railroad crossing: a case study in formal verification of real-time systems. In: Proceedings of real-time systems symposium, pp 120–131

A Clock Based Approach to the Formal Specification of Cyber Physical Systems

Bingqing Xu and Lichen Zhang

Abstract Many cyber-physical systems are under real-time constraints, and thus a critical research challenge is how to ensure time predictability of cyber physical systems. The paper describes the case studies in applying clock theory to the industrial problems. The clock theory described is very simple, in that it models clocks as potentially infinite lists of reals. Xeno's paradox and similar problems are avoided by specifying limits on clock rates, which effectively means that the model sits somewhere between a discrete synchronous model and a fully dense continuous time model as assumed by some other formalisms. Case studies show that using clock theory to specify cyber physical systems can give a more detailed description of the every subsystem and give a much more considerate observation of the time line and sequence of every event.

Keywords Cyber physical systems · Continuous and discrete · Clock · Time analysis

Introduction

Cyber physical systems related research is based on two, originally different world views: on the one hand the dynamics and control (DC) world view, and on the other hand the computer science (CS) world view [1]. The DC world view is that of a predominantly continuous-time system, which is modeled by means of differential (algebraic) equations, or by means of a set of trajectories. The evolution of a hybrid system in the continuous-time domain is considered as a set of piecewise continuous functions of time. The CS world view is that of a predominantly discrete-event system. A well-known model is a (hybrid) automaton, but modeling of discrete-event systems is also based on. As new CPS applications start to interact with the

B. Xu · L. Zhang (✉)
Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China
e-mail: zhanglichen1962@163.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_47, © Springer Science+Business Media Dordrecht 2014

physical world using sensors and actuators, there is a great need for ensuring that the actions initiated by the CPS is timely. This will require new time analysis functionality and mechanisms for CPS. Cyber physical systems are characterized by their stringent requirements for time constraints such as predictable end-to-end latencies, timeliness. Ensuring the time constraint needs of cyber physical systems entails the need to specify and analyze timing requirements correctly.

Since Cyber Physical Systems are dynamic systems that exhibit both continuous and discrete dynamic behavior, and the continuously bilateral interaction between discrete events and continuous time flow makes it hard to know the dynamic feature of the system. So specifying the timing issues is a really vital work in the early stage. This paper describes the case studies in applying clock theory [2, 3] to the industrial problems. The clock theory described is very simple, in that it models clocks as potentially infinite lists of reals. Xeno's paradox and similar problems are avoided by specifying limits on clock rates, which effectively means that the model sits somewhere between a discrete synchronous model and a fully dense continuous time model as assumed by some other formalisms. Case studies show that using clock theory to specify cyber physical systems can give a more detailed description of the every subsystem and give a much more considerate observation of the time line and sequence of every event.

Related Works

The UML profile for Modeling and Analysis of Real-Time and Embedded (Marte) [4] systems has been adopted by the OMG earlier this year. Marte supersedes the UML Profile for Schedulability, Performance and Time (SPT) [5] and extends the mainly untimed uml with several new constructs. Amongst several other things, Marte proposes a new resource model, an extensible way to express non functional properties and a time model. The time model adapts SPT timing mechanisms to the Unified Modeling Language (uml2) simple time and offers completely new features, like the support of logical and multiform time. The Clock Constraint Specification Language (CCSL) [6] defines a set of time patterns between clocks that apply to infinitely many instant relations. A CCSL specification consists of clock declarations and conjunctions of clock relations between clock expressions. A clock expression defines a set of new clocks from existing ones. Most expressions deterministically define one singleclock. In the paper of Lamport [7], concept of "happening before" defines an invariant partial ordering of the events in a distributed multiprocess system. He described an algorithm for extending that partial ordering to a somewhat arbitrary total ordering, and showed how this total ordering can be used to solve a simple synchronization problem. The representation of a closed finitary real-time system as a graph annotated with clock constraints is called a timed automaton [8], since clocks range over the nonnegative reals, every nontrivial timed automaton has infinitely many states. If the clocks of a finitary real-time system are permitted to drift with constant, rational drift bounds, one obtains a finitary drifting-clock system. The representation of a closed finitary drifting-clock system as a graph annotated with constraints on drifting clocks is called an initialized rectangular automaton [9]. Two popular specification languages for the algorithmic verification of untimed systems are finite automata and propositional temporal logics. In order to specify timing constraints, these languages can be extended by adding clock variables. If we judiciously add clocks to finite automata, One obtains the timed automata (TA); from propositional linear temporal logic, one obtains the real-time logic TPTL [10]; from the propositional, branching-time logic CTL, one obtains the real-time logic TCTL [11].

Silva and Krogh [12] introduced sampled data hybrid automata (SDHA) as a formal model of hybrid systems that result from clock-driven computer control of continuous dynamic systems. In contrast to standard hybrid automata, the discrete state transitions in the SDHA can occur only at valid sampling times when the guard conditions are evaluated. Sequences of valid sampling times are defined by a clock structure that specifies bounds on the possible initial phases, period variations and jitter. Approximate quotient transition systems are then defined for SDHA as a theoretical framework for performing formal verification.

Bujorianu et al. [13] presented a multiclock model for real time abstractions of hybrid systems. They call Hybrid Time systems the resulting model, which is constructed using category theory. Such systems are characterized by heterogeneous timing, some components having discrete time and others continuous time. They define a timed (or clock) system as a functor from a category of states to a category of time values. They further define concurrent composition operators and bisimulation.

Clock Theory

Clock theory [2, 14, 15] puts forward the possibility to describe the event in physical world by using a clock, and can analyze, records the event by clock. To use clock to specify Cyber Physical Systems, the time description is clearer to every event and can link continuous world with discrete world better, the definition and linking mechanism of clock theory is provided as below.

Definition 1 *A clock c is an increasing sequence of real numbers. We define its low and high rates by*

$$\Delta(c) =df\, \boldsymbol{inf}\{(c[i+1]-c[i])|i \in Nat\}\nabla(c)$$
$$=df\, \boldsymbol{sup}\{(c[i+1]-c[i])|i \in Nat\}$$

Here, c[1] and c' stand for the first element of c, and the resultant sequence after removal of c[1] from c respectively.

If c is a *healthy clock*, it does not speed up infinitely. Then

$$\Delta(c) > 0$$

If c runs faster than d if for all $i \in Nat$

$$c[i] \leq d[i]$$

then the relation of c and d can be denoted by c 主 d and this is a kind of partial order in clock.

Lemma

$$c \preccurlyeq c^{I}$$

Let c and d be clocks. We define the transition latency between the two clocks as

Lemma

$$\rho(c,d) = df\, \mathbf{sup}\{|c[i] - d[i]|| i \in Nat\}$$

$$\rho(c,d) \geq 0$$

Let e be an event **clock**(e) denotes the clock that records the time instants when the event e occurs. And we use **clock**(**event**(c)) to denote the event that take place at every time instant c[i].

Definition 2 (Local Clock and Global Clock). *Let l be a label denoting a location, and c a clock. Then l:c denotes clock c that locates at l.*

$$\begin{aligned} l : c \preccurlyeq l : d =& df \quad (c \preccurlyeq d)\rho(l : c, l : d) \\ =& df \quad \rho(c,d) \end{aligned}$$

A global clock [l:c] is defined as an equivalent class of local clocks,

$$\rho([l1 : c1],[l2 : c2]) = df \quad \rho(l1 : c1, l2 : c2)$$

Definition 3 (Discrete Variable and Continuous Variable). *Some dynamic features of continuous variable can be described better.*

climb(u,r) is introduced to describe the time instants when the value of u rises up to r. drop(u,r) is introduced to describe the times instants when the value of u falls below r.

Definition 4 (Linking Mechanism). *Here we assign continuous variables to discrete variables.*

$$x = u \text{ every } c \text{ init}$$

$$x0$$

x is a discrete variable, u is a continuous variable, c is a clock with $c[1] \geq 0$*, and* $x0$ *is an initial value. In this equation, we assign the value of u to x at the every instant of clock c.*

In differential equation, u is a continuous variable, and f is an expression.

$$\dot{u} = f \ init \ u0$$

sets up the relation between u and f

$$(\dot{u} = f) \wedge (u(0) = u0)$$

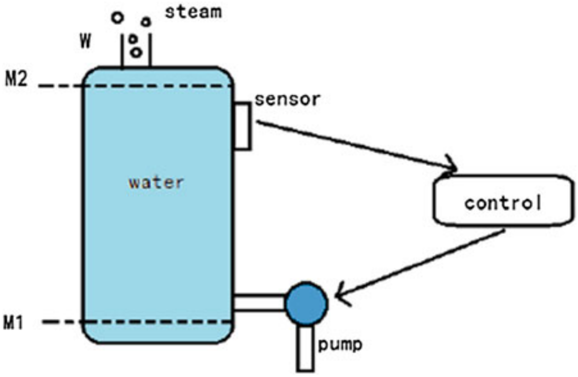
Implementation of Clock Function by Modelica

We use Modelica to implement the clock functions: climb(u,r), drop(u,r), cross(u, r). climb(u,r) and drop(u,r) is coded as follows:

```
( climb function )
function climb
  input Time stime;
  input Real r;
  input Time etime;
  input Time s;
  output Boolean b;
algorithm
  if u_time(stime) > r then
    b := false;
    return;
  end if;
  if _ut(true, s, etime) == r and u__t(true, s, stime) == r then
    b := true;
    return;
  end if;
end climb;
```

```
( drop function )
function drop
  input Time stime;
  input Real r;
  input Time etime;
  input Time s;
  output Boolean b;
algorithm
  if u_time(stime) < r then
    b := false;
    return;
  end if;
  if _ut(false, s, etime) == r and u__t(false, s, stime) == r then
    b := true;
    return;
  end if;
end drop;
```

Fig. 1 Steam boiler control system



Case Study: Specification of the Steam Boiler Control System Based on Clock Theory

As Fig. 1 shows, the steam boiler control system consists of two parts. One is the steam boiler which acts continuously, and the other is the discrete control part which gives the command when it receives messages from sensors. To steam boiler, it has got limits of water level such as minimal water level and maximal water level, and the water level influences the steam rate directly since little water cannot produce steam and too much water is likely to break the limit of steam rate. Sensor collects the information of water level and steam rate and transfers the message to the controller. To the controller, it gives discrete commands to change the quantity of pumps so as to control the water volume and steam rate, and it has got some commands to reply the sensor fault and instrument fault. In addition, this paper only focuses on the basic steam boiler control system, but cannot tolerate the container volume expansion which is caused by heat etc. All the parameters are listed in Table 1.

Table 1 Parameter list for steam boiler control system

Parameter	Value
υ_p	Current speed of water
υ_s	Current speed of steam
V_w	Current volume of water
e	Volume of incoming water
z	Volume change of water
M_1	Minimal volume of water
M_2	Maximal volume of water
C	Capacity of steam boiler
W	Limit of steam rate

To guarantee the safety of the control system, the specification should obey the next two rules. The current steam rate is always less than the highest rate W and the maximal volume of water cannot exceed the capacity of the steam boiler.

$$
\begin{gathered}
0 < \upsilon_s < W \\
0 < M_1 < M_2 < C
\end{gathered}
$$

First of all, we consider the water level; e}is the continuous variable which denotes the volume of incoming water; minus_pump} and add_pumpare events which control the quantity of pumps in order to adjust the steam rate. As the water volume is approaching the maximum, the controller is ready to make some pumps to stop, and the two events have time difference and sequence.

$$
\begin{gathered}
climb(e, M_2 - z) \preccurlyeq clock(minus\_pump) \\
\rho(climb(e, M_2 - z),\ clock(minus\_pump)) \le z/\upsilon_p \\
drop(e, M_1 + z) \preccurlyeq clock(add\_pump) \\
\rho(drop(e, M_1 + z),\ clock(add\_pump)) \le z/\upsilon_p \\
z \le min(M_2 - V_w,\ V_w - M_1)/2
\end{gathered}
$$

Since the water level can influence the steam rate, the following relationship is similar.

$$
\begin{gathered}
clock(minus\_pump) \preccurlyeq clock(low\_steam) \\
\rho(clock(minus\_pump),\ clock(low\_steam)) \le z/\upsilon_p \\
clock(add\_pump) \preccurlyeq clock(high\_steam) \\
\rho(clock(add\_pump),\ clock(high\_steam)) \le z/\upsilon_p
\end{gathered}
$$

And it is vital that some couples of events in the equations above have noninterference.

$$
\begin{gathered}
clock(add\_pump)[1] > 0 \\
clock(add\_pump),\ \preccurlyeq clock(minus\_pump) \preccurlyeq lock(add\_pump)' \\
clock(add\_pump) \wedge clock(minus\_pump) = \phi \\
clock(high\_steam) \wedge clock(low\_steam) = \phi
\end{gathered}
$$

There are also some timing issues in the system, since sensor and controller all need time to transfer the message and command. sensor_delay and control_delay denote the time consumed to transfer information.

$$
\begin{gathered}
(clock(high_{water}) \preccurlyeq clock(sensor\_delay)) \\
\le clock(minus\_pump) \preccurlyeq clock(control\_delay) \preccurlyeq clock(low\_steam) \\
clock(low_{water}) \preccurlyeq clock(sensor\_delay) \\
\le clock(add\_pump) \preccurlyeq clock(control\_delay) \preccurlyeq clock(high\_steam)
\end{gathered}
$$

And no matter which part meets problem, the controller is fault-tolerant.

$$
\begin{gathered}
(clock(sensor\_error) \vee clock(pump\_error)) \\
\preccurlyeq clock(control\_stop) \\
\preccurlyeq clock(SYSTEM\_STOP)
\end{gathered}
$$

In the equations, Ådenotes continuous speed change, and vp denotes discrete speed at each clock unit. The continuous variable and discrete variable can be linked as below:

$$
\begin{aligned}
\dot{e} &= \upsilon_p \; init \; \upsilon_{po} \\
\upsilon_p &= e \; every \; c \; init \; \upsilon_{po}
\end{aligned}
$$

Conclusion

The paper presented the case study in applying clock theory to the industrial problems. Case study shows that using clock theory to specify cyber physical systems can give a more detailed description of the every subsystem and give a much more considerate observation of the time line and sequence of every event.

It is brilliant to connect the event with clock, while it is so difficult to handle so many local clocks with the global clock. In my point of view, it is always hard work to make local clocks keeping consistent with the global clock, and the verification of the security and accuracy of the synchronization is very complicated, and we need more ideas to do this work.

Acknowledgments This work is supported by Shanghai Knowledge Service Platform Project (No. ZF1213), national high technology research and development program of China (No. 2011AA010101), national basic research program of China (No. 2011CB302904), the national science foundation of China under grant (No. 61173046, No. 61021004, No. 61061130541, No. 91118008), doctoral program foundation of institutions of higher education of China (No. 20120076130003), national science foundation of Guangdong province under grant (No. S2011010004905).

References

1. Man KL, Schiffelers RRH (2006) Formal specification and analysis of hybrid systems. Universiteitsdrukkerij Technische Universiteit Eindhoven, ISBN-10: 90-386-2997-4
2. He J (2013) A clock-based framework for constructions of hybrid systems. Key talk. In the proceedings of ICTAC'2013
3. He J (2012) Link continuous world with discrete world. Shanghai Key Laboratory of Trustworthy Computing East China Normal University, China
4. Object Management Group (2009) UML profile for MARTE, v1.0.formal/2009-11-02
5. Object Management Group (2005) UML profile for schedulability, performance, and time specification. OMG document:formal/05-01-02 (v1.1)

6. Mallet F, DeAntoni J, AndrÅ C, de Simone R (2010) The clock constraint specification language for building timed causality models. Innovations Syst Softw Eng 6:99–106
7. Lamport L (1978) Time, clocks, and the ordering of events in a distributed system. Commun ACM 21(7):558–565
8. Alur R, Dill DL (1994) A theory of timed automata. Theor Comput Sci 126:183–235
9. Henzinger TA, Kopke PW, Puri A, Varaiya P (1995) What's decidable about hybrid automata? In: Proceedings of the 27th annual symposium on theory of computing, ACM Press, pp 373–382
10. Alur R, Henzinger TA (1994) A really temporal logic. J ACM 41(1):181–204
11. Alur R, Courcoubetis C, Dill DL (1993) Model checking in dense real time. Inf Comput 104 (1):2–34
12. Silva BI, Krogh BH Modeling and verification of sampled-data hybrid systems. www.ece.cmu.edu/~krogh/checkmate/…/adpm_sampled_data.p
13. Bujorianu MC, Bujorianu LM, Langerak R (2008) An interpretation of concurrent hybrid time systems over multi-clock systems. In: Proceedings of the 17th IFAC world congress, Seoul, Korea, 6–11 July 2008
14. Xu B, He J, Zhang L (2013) Specification of cyber physical systems based on clock theory. Int J Hybrid Inf Technol 6(3):45–54
15. Xu B et al (2013) Specification of cyber physical systems by clock. AST2013.Yeosu, South Korea, ASTL, vol 20, pp 111–114
16. Leeb G, Lynch N (1996) Proving safety properties of the steam boiler controller, lecture notes in computer science. 1165:318–338
17. Abrial J, Bger E, Langmaack H (eds) (1996) Formal methods for industrial applications—specifying and programming the steam boiler control, lecture notes in computer science, vol 1165, Springer, Berlin
18. Duval G, Cattel T (1996) Specifying and verifying the steam boiler problem with SPIN. In: Abrial J-R, Bger E, Langmaack H (eds) Formal methods for industrial applications—specifying and programming the steam boiler control, lecture notes in computer science, vol 1165. Springer, Berlin, pp 203–217
19. Willig A, Schieferdecker I (1996) Specifying and verifying the steam boiler control system with time extended LOTOS. In: Abrial J-R, Bger E, Langmaack H (eds) Formal methods for industrial applications—specifying and programming the steam boiler control, lecture notes in computer science, vol 1165. Springer, Berlin, pp 473–492
20. Carreira PJF, Costa MEF (2003) Automatically verifying an object-oriented specification of the steam-boiler system. Sci Comput Program 46:197–217

A Genetic-Based Load Balancing Algorithm in OpenFlow Network

Li-Der Chou, Yao-Tsung Yang, Yuan-Mao Hong, Jhih-Kai Hu and Bill Jean

Abstract Load balancing service is essential for distributing workload across server farms or data centers and mainly provided by dedicated hardware. In recent years, the concept of Software-Defined Networking (SDN) has been applied successfully in the real network environment, especially by OpenFlow designs. This paper presents an OpenFlow-based load balancing system with the genetic algorithm. This system can distribute large data from clients to different servers more efficiently according to load balancing policies. Furthermore, with the pre-configured flow table entries, each flow can be directed in advance. Once the traffic burst or server loading increased suddenly, the proposed genetic algorithm can help balance workload of server farms. The experiments demonstrate the better performance of the proposed method compared to other approaches.

Keywords Load balancing · Genetic algorithm · OpenFlow · Software-defined networking

Introduction

As growth of the Internet, server overloads always happen due to network traffic of excessive requests or malicious attacks increasing unexpectedly. Load balancing service is a necessary option to ease such burden or sufferings. It can distribute the incoming workload over server farms of duplicate servers to work for more clients.

L.-D. Chou (✉) · Y.-T. Yang · Y.-M. Hong
Department of Computer Science and Information Engineering,
National Central University, Taoyuan, Taiwan, Republic of China
e-mail: cld@csie.ncu.edu.tw

J.-K. Hu · B. Jean
Xinguard Company, Limited, Taipei, Taiwan, Republic of China
e-mail: openflow@xinguard.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_48, © Springer Science+Business Media Dordrecht 2014

Such operations can be done through the process of rewriting the destination of header, and applied by several methods, such as random, round robin, or load-based, to keep the accesses with lower latency and higher throughput, even if more service requests submitted. However, load balancing function is always provided by dedicated software and hardware, which makes it less flexible.

In recent years, the idea of Software-Defined Networking (SDN) has prevailed and practiced successfully in several network environments [1]. OpenFlow is an open source software to implement the SDN architecture and a potential approach to provide abstract elements to rebuild or separate the network topology. The major purpose of using OpenFlow is to increase the flexibility of network management and traffic routing. Moreover, its programmable control interface can help quickly and easily deploy different policies around network devices on the fly instead of redundant manual configuration.

Therefore, we hope through using OpenFlow can manage and deploy the applications which satisfied with our demand to improve the performance of network transmission, raise the flexibility of packet transmission, simplify the network complexity and increase the network management capability even to define the self-network by ourselves.

In this paper, we use Pica8 Open Switch [2] to be the device which support OpenFlow and Open vSwitch (OVS) integration. And we deploy application control programs that using policy-based load balancing methods to achieve OpenFlow-based load balancing and raising the flexibility for network data transmission. We hope to achieve better load balancing performance through the OpenFlow capability and toward to the goal of Software Defined Networking finally.

This paper describes related work in Section Related Work, system design in Section System Design, and experimental results In Section Experimental Results. Finally in Section Conclusion and Future Work, the conclusion and future work are remarked.

Related Work

In this section, load balancing is first introduced; then the Open vSwitch and OpenFlow Controller are described in the final.

SDN is the innovative network architecture that provides a quick deployment instead of manually configuring policies. And OpenFlow is the best solution for achieving the objective of SDN now.

Load Balancing

Cloud computing becomes more and more popular in recent years, the reasons include not only the convenience for end users but also the virtualization technology and distributed computation. The number of computing nodes will also increase quickly because the need for support big data distributed processing. Therefore, it's much important for load balancing in the cloud network that distribute the traffic to separate servers. In cloud network architecture, the network traffic increase because of the big data from clients. So the single server can't afford the capability to process the large traffic. And in order to avoid wasting the resources with servers, load balancing technology was provided [3].

For network applications, it's necessary for load balancing and when the traffic are too large to handle, and we can also cut down the traffic load because of it. In [4], we can use the central load balancing decision module. It's not only increasing the decision speed but also allow single point of failure for the virtual machine. In [5], it present Plug-n-Server system which control the traffic load by OpenFlow-based routing methods in order to minimize the response time from client to server. But this system didn't consider the advanced load balancing algorithms to increase the network performance. On the other hand, someone also present the novel ideal for improving the bad performance for OpenFlow controller because the huge of flows caused by separating flows for each client connection [6]. The advantage for this paper is avoiding the cost for processing the huge of flows. On the contrary, there are scant physical OpenFlow switches support the module features presented by this paper as its drawbacks. And in OpenFlow-based load balancing technology, some researches present the comparison between simple load balancing algorithms such as random choice, time slice based choice and weighted balancing [7]. And from the research [8], it designs the load balancing architecture which apply in web services applications and it also presents three load-balancing policies which compared with our Genetic algorithm-based load balancing algorithm in this paper. The first is random-based load balancing policy that selects the registered servers randomly. Secondly, it uses the round robin load balancing policy to rotates the registered servers to serve the requests. Last, by using load-based load balancing algorithm which chooses the server with the current lowest load to serve the requests. In [9], it uses multiple OpenFlow controllers to distribute the different services that integrating the network and the load balancing functionality to reduce the maintenance effort. For example, one controller handles the load for e-mail servers and the other one processes the load for network traffic. It increases the efficiency by separating the load for multiple controllers which implemented by the open source project named FlowVisor [10].

Open vSwitch and OpenFlow Controller

OpenFlow is an open standard originated from the Clean slate project in Stanford University. And Open vSwitch (OVS) is a multilayer virtual switch licensed under the open source Apache 2.0 license which is also an Ethernet switch that has flow table inside to insert and delete flow entries [11]. It can operate both as a soft switch running within the hypervisor, and as the control stack for switching silicon.

Currently, it has many OpenFlow controllers which released with open source project, such as Beacon [12], NOX [13], POX [14], Helios, BigSwitch and so on. OpenFlow Controllers communicate to the OpenFlow switches by the secure channel which defined by OpenFlow protocol.

System Design

The proposed OpenFlow-based load balancing system is designed for redirecting the flows in order to balance the loads for the servers in load balancing pool and mirror the packets which have the abnormal traffic to backend detection server.

Figure 1 shows the OpenFlow-based load balancing system architecture, It separates two basic components: OpenFlow switch and OpenFlow controller. Two of components have multiple modules inside which perform its own functions and communicate each other.

In this paper, we propose an intelligent load balancing algorithm using Genetic Algorithm (GA). We assume each client sends multiple requests and the servers also have different workload on it. So the flow redirection problem become NP-Complete problem that we hope to find the best order for redirecting to servers to modify the flow rules. Suppose that we have *N* flows to do pre-configured, and

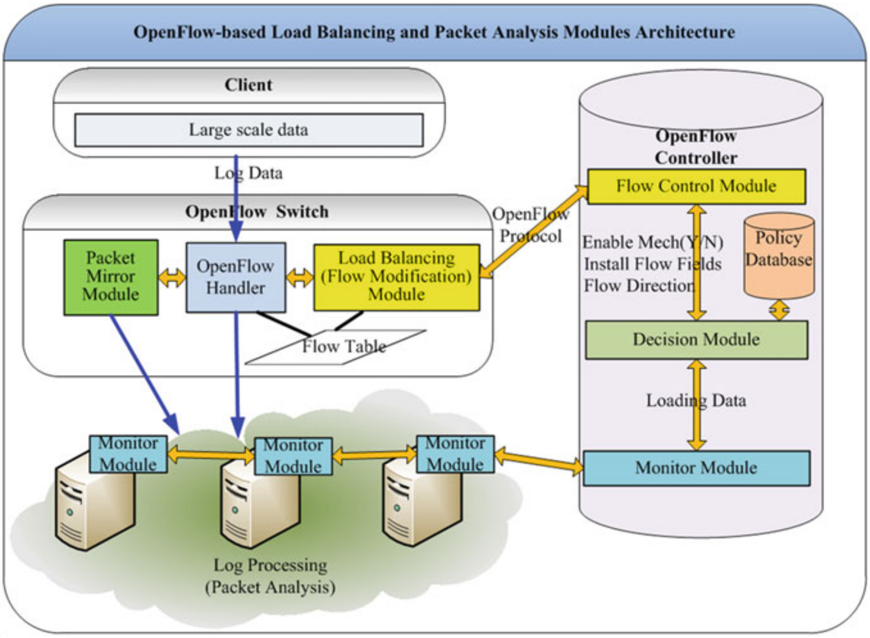


Fig. 1 System architecture

each flow have different load which noted the set of *Ω*. On the other hand, we have *K* servers to receive the data and each server has different workload which noted the set of *X*. Moreover, the flow redirection problem is considered with the objective of minimizing the server's coefficient of variation for traffic in OpenFlow environment. For proposed genetic-based load balancing algorithm in this paper, we define the fitness function by the coefficient of variation for server's traffic as the following formula:

$$\text{Min}\ \frac{\sqrt{\left(\left(\sum_{\mathrm{j}=1}^{K} \mathrm{X[j]}^2\right) - \left(\left(\sum_{\mathrm{j}=1}^{K} \mathrm{X[j]}\right)/K\right)^2\right)/K}}{\left(\sum_{\mathrm{j}=1}^{K} \mathrm{X[j]}\right)/K} \tag{1}$$

For the evolution processes in proposed algorithm, we choose the roulette wheel selection for reproduction, single point crossover and single point mutation.

Experimental Results

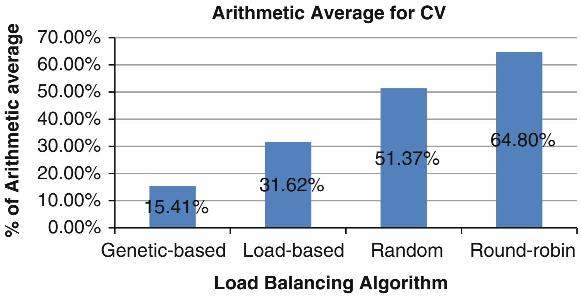
In order to compare the performance for proposed load balancing algorithm, we use the other three algorithms to be compared together. In order to achieve load balancing, we use the arithmetic average for coefficient of variation (CV) to be the indicator for the performance evaluation. The more lower of CV, it represents the extent of variability are lower. Therefore, if the CV calculated for load balancing algorithm is lower, it shows the algorithm get more better load balancing result. And we assume we can get the load for the flows at each clients, the simulation parameters in Table 1:

The following diagram, shown in Fig. 2, is the simulation results for the four load balancing algorithms. We can see the genetic-based load balancing algorithm has the best performance because it has the lowest percentage of arithmetic average for CV. In the other words, genetic-based load balancing algorithm can achieve the best load balancing result which compared with the other three algorithms.

Table 1 Simulation parameters

Experiment times Variables/parameters	20 times Input
K (number of servers)	4
N (number of flows)	4
Population size	5
Crossover rate	0.9
Mutation rate	0.2

Fig. 2 Simulation results



Conclusion and Future Work

In this paper, we present flexible load balancing algorithms by our OpenFlow-based load balancing system. Through the proposed genetic-based load balancing algorithm and self-definition flow entries by this paper, the traffic follow the pre-configured flows and redirect to the nodes which in the load balancing pool. It can save the cost and avoid the bottleneck because of the throughput for the controller channel. In the simulation results, we can see the significant performance on our proposed genetic-based load balancing algorithm which evaluated by arithmetic average for coefficient of variation. In order to enhance the scalability and stability for our system, we will focus on scaling the network topology to enterprise class and providing the high availability properties for our system in the future.

Acknowledgments This work was supported by the National Science Council of the Republic of China under grant NSC 102-2622-E-008-002-CC2. And the author also thanks Xinguard for their support to Pica8 Open Switch.

References

1. Haleplidis E, Denazis S, Koufopavlou O, Halpern J, Salim JH (2012) Software-defined networking: experimenting with the control to forwarding plane interface. In: Proceedings of european workshop software defined networking (EWSDN), pp 91–96, Oct 2012
2. Pica8 Open switch http://www.pica8.com/open-switching/1-gbe-10gbe-open-switches.php
3. Maguluri ST, Srikant R, Lei Y (2012) Heavy traffic optimal resource allocation algorithms for cloud computing clusters. In: Proceedings of the 24th international teletraffic congress, pp 1–8, Sep 2012
4. Radojevic B, Zagar M (2011) Analysis of issues with load balancing algorithms in hosted (cloud) environments. In: Proceedings of the 34th international convention MIPRO, pp 416–420, 23–27 May 2011
5. Handigol N, Seetharaman S, Flajslik M, McKeown N, Johari R (2009) Plug-n-serve: load-balancing web traffic using openflow. In: Proceedings of demo at ACM SIGCOMM, Aug 2009

6. Wang R, Butnariu D, Rexford J (2011) OpenFlow-based server load balancing gone wild. In: Proceedings of the 11th USENIX conference on hot topics in management of internet, cloud, and enterprise networks and services, pp 12–12, Mar 2011
7. Marcon DS, Bays LR (2011) Flow based load balancing: optimizing web servers resource utilization. J Appl Comput Res Dec 2011
8. Uppal H, Brandon D (2010) OpenFlow based load balancing. In: Proceedings of CSE561: networking. project report. University of Washington, Spring 2010
9. Koerner M, Kao O (2012) Multiple service load-balancing with OpenFlow. In: Proceedings of the 13th international conference on high performance switching and routing, pp 210–214, 24–27 Jun 2012
10. Eokhong M, Eungju K, Yong L, Ngchul K, Taek H, Guk K (2012) Implementation of an openflow network virtualization for multi-controller environment. In: Proceedings of the 14th international conference on advanced communication technology, pp 589–592, Feb 2012
11. Open vSwitch. http://openvswitch.org/
12. Beacon: a java-based openflow control platform, Oct 2011. http://www.beaconcontroller.net/
13. Gude N, Koponen T, Pettit J, Pfaff B, Casado M, McKeown N, Shenker S (2008) Nox: towards an operating system for networks. ACM SIGCOMM Comput Commun Rev Jul 2008
14. POX www.noxrepo.org/pox/about-pox/
15. Jarschel M, Oechsner S, Schlosser D, Pries R, Goll S, Tran-Gia P (2011) Modeling and performance evaluation of an OpenFlow architecture. In: Proceedings of the 23rd international teletraffic congress, pp 1–7, Sep 2011

QoS Modeling of Cyber Physical Systems by the Integration of AADL and Aspect-Oriented Methods

Lichen Zhang

Abstract This paper proposes an aspect-oriented QoS modeling method based on AADL. Aspect-Oriented development method can decrease the complexity of models by separating its different concerns. In model-based development of cyber physical systems this separation of concerns is more important given the QoS concerns addressed by Cyber physical Systems. These concerns can include timeliness, fault-tolerance, and security Architecture Analysis and Design Language (AADL) is a standard architecture description language to design and evaluate software architectures for embedded systems already in use by a number of organizations around the world. In this paper, we present our current effort to extend AADL to include new features for separation of concerns., we make a in-depth study of AADL extension for QoS. Finally, we illustrate QoS aspect-oriented modeling via an example of transportation cyber physical system.

Keywords QoS · Cyber physical systems · Aspect-Oriented · AADL

Introduction

The dependability of the software [1] has become an international issue of universal concern, the impact of the recent software fault and failure is growing, such as the paralysis of the Beijing Olympics ticketing system and the recent plane crash of the President of Poland. Therefore, the importance and urgency of the digital computing system's dependability began arousing more and more attention. A digital computing system's dependability refers to the integrative competence of the system that can provide the comprehensive capacity services, mainly related to the reliability, availability, testability, maintainability and safety. With the

L. Zhang (✉)
Shanghai Key Laboratory of Trustworthy Computing, East China Normal University,
Shanghai 200062, China
e-mail: zhanglichen1962@163.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_49, © Springer Science+Business Media Dordrecht 2014

increasing of the importance and urgency of the software in any domain, the dependability of the distributed real-time system should arouse more attention.

Aspect-oriented programming (AOP) [2] is a new software development technique, which is based on the separation of concerns. Systems could be separated into different crosscutting concerns and designed independently by using AOP techniques. Every concern is called an "aspect". Before AOP, as applications became more sophisticated, important program design decisions were difficult to capture in actual code. The implementation of the design decisions were scattered throughout, resulting in tangled code that was hard to develop and maintain. But AOP techniques can solve the problem above well, and increase comprehensibility, adaptability, and reusability of the system. AOSD model separates systems into tow parts: the core component and aspects.

With the deepening of the dependable computing research, the system's dependability has becoming a important direction of cyber physical systems, the modeling and design of cyber physical systems has become a new field. The dependable cyber physical system has a high requirement of reliability, safety and timing, these non-functional properties dispersed in the various functional components of system, so the Object-Oriented design has lost its superiority very obviously. The QoS of dependable real-time system [3] is very complex, currently the QoS research still hasn't a completely and technical system, and there isn't any solution meeting all the QoS requirements. We design the QoS of dependable real-time system as a separate Aspect using AOP, and proposed the classification of complex QoS, divided into the timing, reliability and safety and other sub-aspects. These sub-aspects inherit t members and operations from the abstract QoS aspect. We design each sub-aspects through aspect-Oriented modeling, to ensure the Quality of dependable real-time system meeting the requirements of the dependability.

This paper proposes an aspect-oriented QoS modeling method based on AADL [4]. Aspect-Oriented development method can decrease the complexity of models by separating its different concerns. In model-based development of cyber physical systems this separation of concerns is more important given the QoS concerns addressed by Cyber physical Systems. These concerns can include timeliness, fault-tolerance, and security Architecture Analysis and Design Language (AADL) is a standard architecture description language to design and evaluate software architectures for embedded systems already in use by a number of organizations around the world. In this paper we present our current effort to extend AADL to include new features for separation of concerns., we make a in-depth study of AADL extension for QoS. Finally, we illustrate QoS aspect-oriented modeling via an example of transportation cyber physical system.

Aspect-Oriented Specification of QoS Based on AADL

AOP is a new modularity technique that aims to cleanly separate the implementation of crosscutting concerns. It builds on Object-Orientation, and addresses some of the points that are not addressed by OO. AOP provides mechanisms for decomposing a problem into functional components and aspectual components called aspects [5]. An aspect is a modular unit of crosscutting the functional components, which is designed to encapsulate state and behavior that affect multiple classes into reusable modules. Distribution, logging, fault tolerance, real-time and synchronization are examples of aspects. The AOP approach proposes a solution to the crosscutting concerns problem by encapsulating these into an aspect, and uses the weaving mechanism to combine them with the main components of the software system and produces the final system. We think that the phenomenon of handling multiple orthogonal design requirements is in the category of crosscutting concerns, which are well addressed by aspect oriented techniques. Hence, we believe that system architecture is one of the ideal places where we can apply aspect oriented programming (AOP) methods to obtain a modularity level that is unattainable via traditional programming techniques. To follow that theoretical conjecture, it is necessary to identify and to analyze these crosscutting phenomena in existing system implementations. Furthermore, by using aspect oriented languages, we should be able to resolve the concern crosscutting and to yield a system architecture that is more logically coherent. It is then possible to quantify and to closely approximate the benefit of applying AOP to the system architecture [6].

AADL [4] is an architecture description language developed to describe embedded systems is shown in Fig. 1. AADL (Architecture Analysis and Design Language), which is a modeling language that supports text and graphics, was approved as the industrial standard AS5506 in November 2004. Component is the most important concept in AADL. The main components in AADL are divided into three parts: software components, hardware components and composite components. Software components include data, thread, thread group, process and subprogram. Hardware components include processor, memory, bus and device. Composite components include system [7–9].

In its conformity to the ADL definition, AADL provides support for various kinds of non-functional analyses along with conventional modeling as shown in Fig. 2 [7–9]:

Flow Latency Analysis: Understand the amount of time consumed for information flows within a system, particularly the end-to-end time consumed from a starting point to a destination.

Resource Consumption Analysis: Allows system architects to perform resource allocation for processors, memory, and network bandwidth and analyze the requirements against the available resources.

Real-Time Schedulability Analysis: AADL models bind software elements such as threads to hardware elements like processors. Schedulability analysis helps in examining such bindings and scheduling policies.

Safety Analysis: Checks the safety criticality level of system components and highlights potential safety hazards that may occur because of communication among components with different safety levels.

Security Analysis: Like safety levels, AADL components can be assigned various security levels. The analysis helps in identifying the security loopholes that

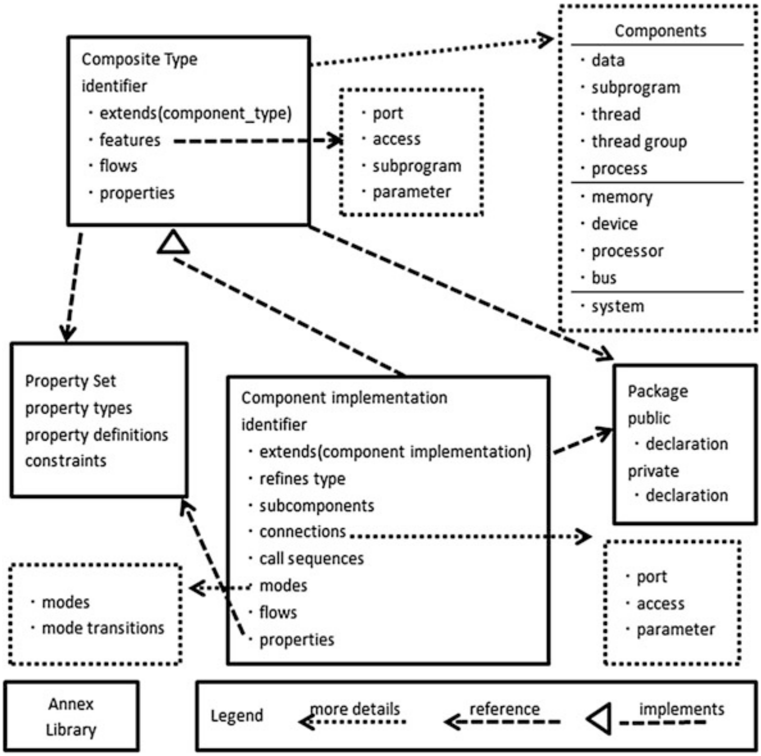
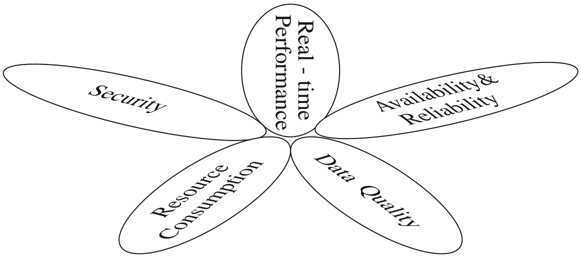


Fig. 1 AADL elements

Fig. 2 AADL and non-functional analysis



may happen because of mismatches in security levels between a component and its subcomponents, and communication among components with different security levels.

AADL defines two main extension mechanisms: property sets as shown in Fig. 3 and sublanguages (known as annexes). Annexes and properties allow the addition of complex annotations to AADL models that accommodate the needs of multiple concerns. These annotations, along with their corresponding analysis plug-ins, provide a powerful combination for the architect to evaluate his/her

```
property set Clemson is
MbitPerSec : type units (MPS, GPS => MPS*1000);
Band_width: type aadlinteger units Clemson::MbitPerSec;
Radio_band_width: Clemson::Band_width applies to (all);
Band_width_802_11g: constant Clemson::Band_width => 54 MPS;
Band_width_802_11n: constant Clemson::Band_width => 300 MPS;
Band_width_fast_ethernet: constant Clemson::Band_width => 100 MPS;
end Clemson;
```

Fig. 3 Property sets of AADL

design choices from different perspectives. The extension mechanisms in AADL enable these perspectives to evolve in number and complexity as the knowledge on them also evolves [10, 11].

AO4AADL is an aspect oriented extension for AADL as shown in Fig. 4 [11–13]. This language considers aspects as an extension concept of AADL components called aspect annex. Instead of defining a new aspect oriented ADL, we extend AADL, a well-known ADL, with an aspect annex. So we consider, in this work, that aspects can be specified in a language other than AADL, and then integrated in AADL models as annexes. Based on the annex extension mechanism,

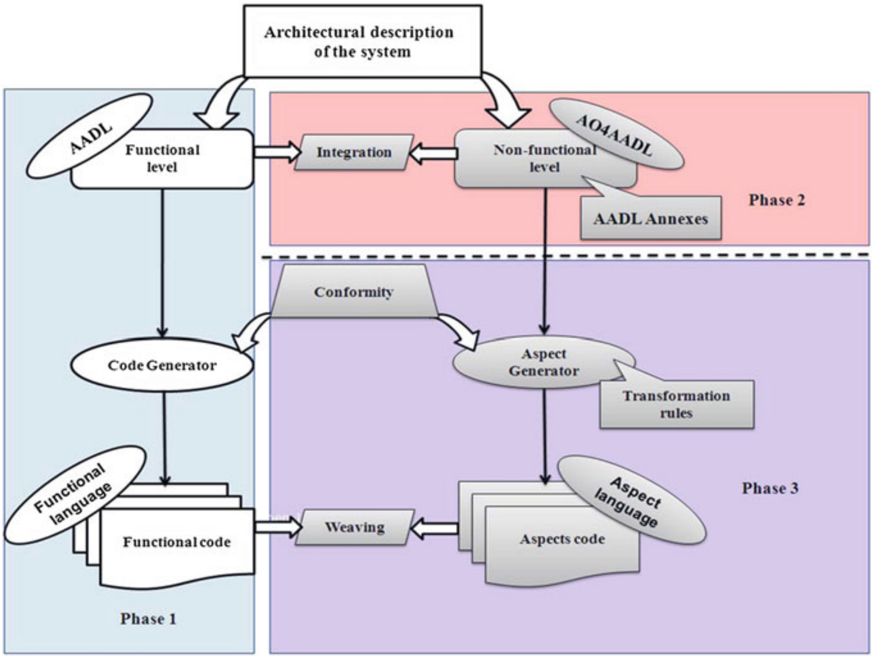


Fig. 4 AO4AADL structue

we propose to enrich AADL specifications with aspect concepts. AO4AADL Consists mainly in three phases:

Implementing Functional code: In this phase, the designer should focus on the main functionalities of his application without considering any non-functional properties.

Designing non-functional properties: At this phase, the designer defines the non functional safety properties and states the conditions under which his application operates correctly, such as security, availability, etc.

Generating non-functional properties code: AO4AADL aspects can be translated in different aspect languages such as AspectJ or JAC for Java language, AspectAda for Ada language, AspectC for C language, etc.

Ana-Elena Rugina, Karama Kanoun and Mohamed Kaâniche proposed a four step modeling dependability methods based on AADL [14]: The *first step* is devoted to the modeling of the application architecture in AADL. The *second step* concerns the specification of the application behavior in the presence of faults through AADL error models associated with components of the architecture model. The *third step* aims at building an analytical dependability evaluation model, from the AADL dependability model, based on model transformation rules. The *fourth step* is devoted to the dependability evaluation model processing that aims at evaluating quantitative measures characterizing dependability attributes.

In this paper, we extend AADL by aspect-oriented method in following aspect:

Physical world aspect: Cyber physical systems are often complex and span multiple physical domains, whereas mostly these systems are computer controlled.

Dynamic Continuous dynamics Aspect: Cyber physical systems are mixtures of continuous dynamic and discrete events. These continuous and discrete dynamics not only coexist, but interact and changes occur both in response to discrete, instantaneous, events and in response to dynamics as described differential or difference equations in time.

Formal Specification Aspect of Data: A formal specification aspect of data captures the static relation between the object and data. Formal data aspect emphasizes the static structure of the system using objects, attributes, operations and relationships based on formal techniques.

Formal Specification aspect of Information flow and control flow: Formal Specification aspect of Information flow and control flow aims at facilitating the description and evaluation of various flow properties measures. It provides an ideal semantic basis to characterize the behavior of the component communications in the system.

Spatial Aspect: The analysis and understanding of railway cyber physical systems spatial behavior—such as guiding, approaching, departing, or coordinating movements is very important.

Case Study: Aspect-Oriented Specification of QoS of VANET

Presently study in VANET is still at the preliminary stage [15–18]. We model the most promising system of VANET (Vehicular Ad-hoc NETwork) as shown in Fig. 5.

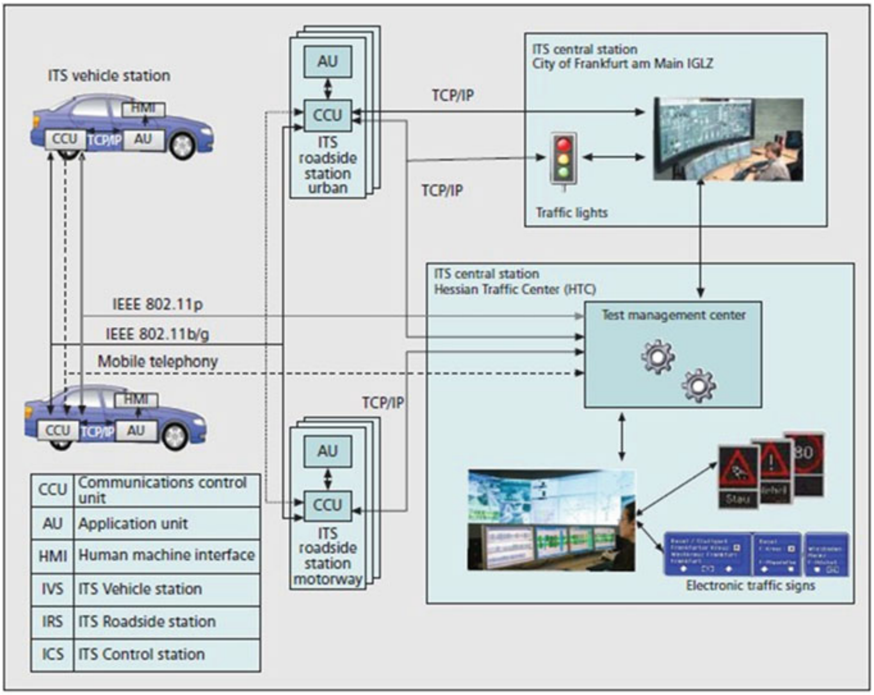
The AADL Model of Vehicle Station is shown in Fig. 6.

Delay Analysis of Data Flow is shown in Fig. 7.

```
flows

  ETE_F1: end to end flow BRAKE.Flow1
  -> C22 -> CC.brake_flow_1
  -> C29 -> TA.Flow1
  {
  Latency => 200 Ms;
  };
flows

  brake_flow_1: flow path brake_status
```



CCU	Communications control unit
AU	Application unit
HMI	Human machine interface
IVS	ITS Vehicle station
IRS	ITS Roadside station
ICS	ITS Control station

Fig. 5 The architecture of VANET [19]

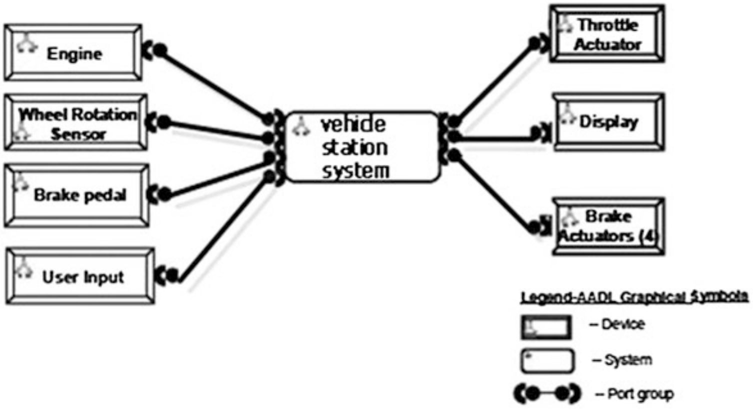


Fig. 6 AADL model of vehicle station

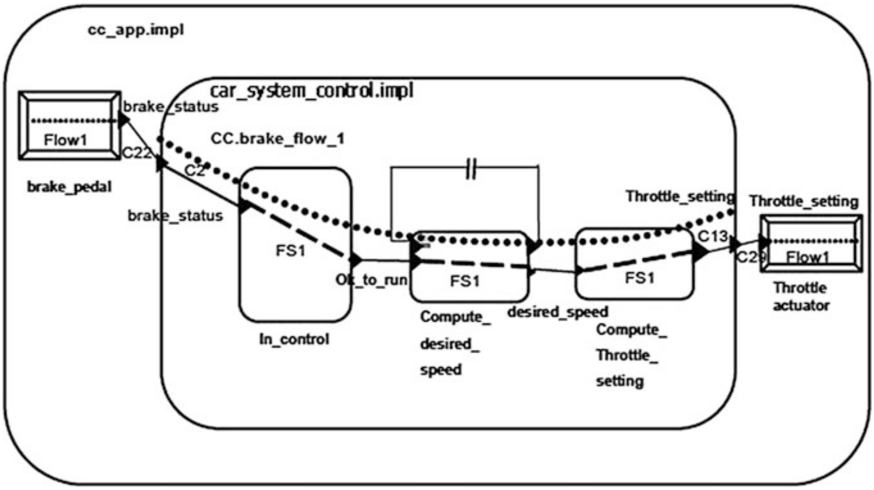


Fig. 7 Delay analysis of data flow

```
-> C2 -> I_C.FS1
-> C9 -> C_D_S.FS1
-> C12 -> C_T_S.FS1
-> C13 -> throttle_setting{
Latency => 130 Ms;
};
```

Conclusion

This paper proposed an aspect-oriented QoS modeling method based on AADL. Aspect-Oriented development method can decrease the complexity of models by separating its different concerns. In model-based development of cyber physical systems this separation of concerns is more important given the QoS concerns addressed by Cyber physical Systems. These concerns can include timeliness, fault-tolerance, and security. Architecture Analysis and Design Language (AADL) is a standard architecture description language to design and evaluate software architectures for embedded systems already in use by a number of organizations around the world. In this paper we present our current effort to extended AADL to include new features for separation of concerns., we made a in-depth study of AADL extension for QoS. Finally, we illustrated QoS aspect-oriented modeling via an example of the specification of VANET based on AADL.

The future work focuses on the integration AADL and formal techniques to specify and verification of QoS of cyber physical systems.

Acknowledgments This work is supported by Shanghai Knowledge Service Platform Project (No. ZF1213), national high technology research and development program of China (No. 2011AA010101), national basic research program of China (No. 2011CB302904), the national science foundation of China under grant (No. 61173046, No. 61021004, No. 61061130541, No. 91118008), doctoral program foundation of institutions of higher education of China (No. 20120076130003),national science foundation of Guangdong province under grant (No. S2011010004905).

References

1. Laprie JC (ed) (1992) Dependability: basic concepts and terminology. Springer, Berlin
2. Kiczales G et al (1997) Aspect-oriented programming. In: Proceedings of the 11th european conference on object-oriented programming, June 1997
3. Frolund Svend, Koistinen Jari (1998) Quality of service specification in distributed object systems. IEE/BCS Distrib Syst Eng J 5:179–202
4. AE Aerospace (2009) SAE AS5506A[S]: architecture analysis and design language V2.0
5. Aldawud O, Elrad T, Bader A (2001) A UML profile for aspect oriented modeling. In: Proceedings of workshop on AOP
6. Wehrmeister MA, Freitas EP, Pereira CE et al (2007) An aspect-oriented approach for dealing with non-functional requirements in a model-driven development of distributed embedded real-time systems. In: Proceedings of 10th IEEE international symposium on object and component-oriented real-time distributed computing, IEEE Computer Society. Santorini Island, Greece, 7–9 May 2007 pp 428–432
7. Feiler P, Hugues J, Sokolsky O (eds) (2012) Architecture-driven semantic analysis of embedded systems. Dagstuhl seminar 12272, Dagstuhl Report, vol 2 (7). pp 30–55. ISSN 2192-5283
8. The story of AADL (2010) AADL wiki. software engineering institute, Web. 06 Jan 2012
9. Muhammad N, Vandewoude Y, Berbers Y, van Loo S (2010) Modelling embedded systems with AADL: a practical study. www.intechopen.com/download/pdf/10732

10. de Niz D, Feiler PH (2007) Aspects in the industry standard AADL. In: Proceedings of AOM '07 Proceedings of the 10th international workshop on aspect-oriented modeling. pp 15–20
11. Michotte L, Vergnaud T, Feiler P, France R (2008) Aspect oriented modeling of component architectures using AADL. In: Proceedings of the 2nd international conference on new technologies, mobility and security, 5–7 Nov 2008
12. Loukil S, Kallel S, Zalila B, Jmaiel M (2010) Poster -AO4AADL: aspect oriented ADL for embedded systems. In: Proceedings of international conference on new technologies of distributed systems (NOTERE)
13. Loukil S, Kallel S, Zalila B, Jmaiel M (2010) AO4AADL: an aspect oriented ADL for embedded systems. In: Proceedings of the 4th european conference on software architecture (ECSA 2010), LNCS. Springer, Copenhagen
14. Rugina AE, Kanoun K, Kaaniche M (2006) An architecture-based dependability modeling framework using AADL. In: Proceedings of 10th IASTED international conference on software engineering and applications (SEA'2006), Dallas (USA), 13–15 Nov 2006 (13/11/2006), pp 222–227
15. Festag A, Ëussler H, Hartenstein H, Sarma A, Schmitz R (2004) Fleet net: bringing car-to-car communication into the real world. In: Proceedings of the 11th ITS world congress and exhibtion [C], Nagoya, Japan, pp 1–8
16. CVIS. Cooperative vehicle-infrastructure systems [EB/OL]. http://www.cvisproject.org. Accessed on 19 Feb 2013
17. SAFESPOT. Cooperative vehicles and road infrastructure for road safety[EB/OL]. http://www.safespot-eu.org. Accessed on 19 Mar 2013
18. Kargl F, Papadimitratos P, Buttyan L (2008) Secure vehicular communication systems: implementation, performance, and research challenges. IEEE Commun Mag 46(11):110–118
19. Stubing Hagen (2010) Adam opel gmbh marc bechler.simTD: a car-to-x system architecture for field operational tests. IEEE Commun Mag 48(5):148–154

Assessment of Performance in Data Center Network Based on Maximum Flow

Kai Peng, Rongheng Lin, Binbin Huang, Hua Zou and Fangchun Yang

Abstract Recently, data center networks (DCN) have received significant attention from the academic and industry. However, researches of DCN are mainly concentrated on the improvement of network architectures and the design of routing protocols, or the performance evaluation from the perspective of node importance. In contrast to existing solutions, in this paper, we propose using maximum-flow theory to assess the network performance. Firstly, we abstract two kinds of typical DCN architectures and then formulate and convert the performance analysis of those architectures into a maximum-flow problem including a supersource and a supersink. Secondly, we get the value of maximum-flow by using Edmonds and Goldberg algorithm. Last but not the least, based on the theory of maximum-flow and Minimal cut sets, we get the critical edges for each architecture. Extended experiments and analysis show that our method is effective and indeed introduce low overhead on computation. In addition, the method and issues observed in this paper is generic and can be widely used in newly proposed DCN architectures.

Keywords DCN · Topologies · Maximum-flow · Assessment

K. Peng (✉) · R. Lin · B. Huang · H. Zou · F. Yang
State Key Laboratory of Networking and Switching Technology, Beijing
University of Posts and Telecommunications, Beijing, China
e-mail: pkbupt@gmail.com

R. Lin
e-mail: rhlin@bupt.edu.cn

B. Huang
e-mail: binbinHUang@bupt.edu.cn

H. Zou
e-mail: zouhua@bupt.edu.cn

F. Yang
e-mail: fcyang@bupt.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_50, © Springer Science+Business Media Dordrecht 2014

Introduction

Driven by the recent proliferation of cloud services [1, 2] and trend of IT systems, data center networks have gained a wide attention from both industry and research community. Several network architectures [3–8] have been proposed for these extensive data centers.

In general, all existing DCN architectures can be divided into two classes, hierarchical architecture and flat architecture. The hierarchical architecture is represented by the traditional Multi-rooted Tree Architecture and employs a layered structure which is constituted recursively by lower-level components. The other one (for example, FatTree [6] and VL2 [7] architectures) which organizes all servers into the same level by using some sort topologies. Actually, the main research of these new architectures is limited to load balancing or improvement of architectures as the increasing new services [9]. Although some of researchers are engaged in investigating how these architectures are affected in visualized environments [10], while a few of existing research concerned about the security of the data center network, especially for the vulnerability evaluation of these new architectures. In our previous research, we mainly focus on the assessment of node importance for those newly proposed ones. In contrast to previous solution, in this paper, we fill this void by conducting an experimental evaluation of the-state-of-the-art architectures from the perspective of network flow. The main idea is as follows. When the network traffic reaches the maximum, the paths or the edges which are in the saturated status are the critical ones. In addition, critical paths or edges are treated as the focus of attack and defense, which need to be given more attention. For one thing, NSP (Network service provider) should increase the traffic capabilities of the critical links in order to increase the capabilities of entire network. For another, for the attackers, once they get the information of critical paths, try to destroy the critical paths may greatly reduce the cost of attack so as to maximize their benefits. Unfortunately, edges importance of these new DCN architectures has never been investigated before.

In this paper, we try to conduct experiments on Multi-rooted Tree architecture and VL2 architecture, each respectively as a representative of hierarchical and flat architectures. Taking any of them for example, we first use a directed and weight graph to describe its topology. As we know, there are multi-source and multi-sink in this architecture, therefore, we make this be converted into an ordinary flow network with only one single source and one single sink by adding a supersource and a supersink. Secondly, we get the maximum-flow and corresponding paths by using Edmonds and Goldberg algorithm. Finally, based on maximum-flow and Minimal cut sets theory, we get the critical edges.

Our contributions are summarized as two aspects.

1. For one thing, we formulate and transform the DCN architecture as a maximum-flow problem from the perspective of network flow, and then show it feasibility.

2. We formulate and conduct an experimental evaluation of two typical DCN architectures. The results show that our proposed method indeed finds the maximum-flow and critical edges for each architecture. Furthermore, the method can be generalized and applied as guidelines for the newly proposed DCN architectures.

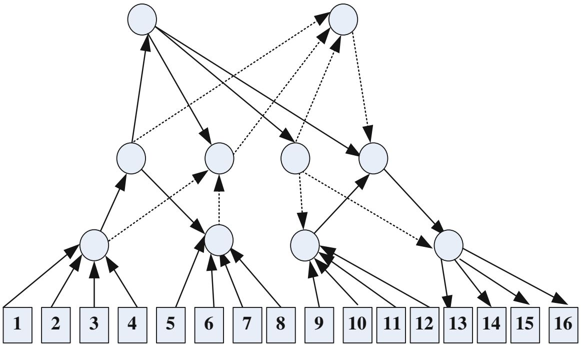
The remainders of this paper are as follows. In section Background, we give a brief introduction of current DCN architectures. Problem formulation and math mode are described in section Problem Formulation and Math Mode. Section Model Solution and Algorithm presents model solution and the algorithm process, followed by section Experiment Evaluation and Discussion; we show the experiment evaluation and the discussion. Finally, we conclude the paper in section Conclusion.

Background

In this section, we give a brief introduction of current data center architectures. Based on how the system is constructed, we classify existing data center network (DCN) architectures into two types, hierarchical one and flat one.

In a hierarchical architecture, servers are arranged in different levels. A higher-level network consists of multiple components of lower-levels. Actually, the most widely used hierarchical DCN architecture is traditional Multi-rooted Tree architecture, known as the typical three-tier architecture. As is shown in Fig. 1, the tree architecture generally consists of switches of three layers. We can see that such topology have 16 servers (has been marked as Tree architecture). At the bottom level, known as the edge layer, where each server connects to one or two edge switches. Each edge switches connect to one or two switches at the layer of aggregation, where all the hosts are placed on the edge layer, forming the leaf nodes within the whole networks, which are responsible for storage and computation. Above all, each aggregation switch connects with multiple switches at the core layer.

Fig. 1 Topology of tree architecture



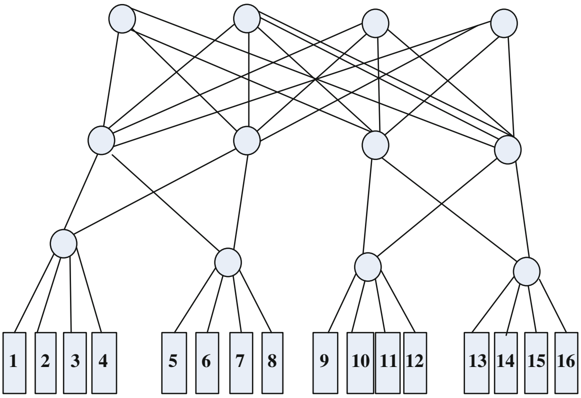


Fig. 2 Topology of VL2 architecture

In contrast to hierarchical architecture, the other type of DCN is flat one. It utilizes a flat organization of servers, which are placed into a single layer and are interconnected by switches. One typical representative of this class is the VL2 Architecture [7] which was introduced by Greenberg et al. In VL2 architecture, a complete bipartite graph was formed by the switches from the layers of aggregation and core. More specifically, interconnection topology of switches follows a folded Clos network where switches of intermediate and aggregation are respectively organized into each side of the graph. Figure 2 shows one example with 16 servers (marked as VL2).

These architectures have been designed independently for different goals in practice. In this paper, we aim to evaluate the performance of these architectures from the perspective of maximum-flow.

Problem Formulation and Math Mode

In this section, we formally propose and define the maximum flow problem and then mainly show the mode in details.

Problem Formulation

Based on the previous topologies in Section Background, we get the main structure of the data center network. Firstly, we choose one of the architecture for example. We use a directed graph G to describe the tree topology. As shown in Fig. 1, all of the nodes have been marked. We assume that the network flow goes through from left to right. As reflected in Fig. 1, The nodes of $\{v_1, v_2, v_3, v_4\}$ are all can be seen as the sources and the network flow eventually flows to the nodes of

$\{v_{13}, v_{14}, v_{15}, v_{16}\}$ which are the sinks in this topology. As maximum-flow theory only supports the graph which exists only one source and one sink, and thus when directly applied in tree architecture, they may suffer from drawback. Therefore, we will show how to convert this problem into a traditional maximum-flow one in the coming section.

The Mode of Maximum-Flow Problem

Transformation of Maximum-Flow Problem

Figure 3 shows how the network from Fig. 1 can be converted into an ordinary flow network with only a single source and a single sink. We first add a supersource s and add a directed edge $(\mathrm{s}, \mathrm{s}_i)$ with capacity $\mathrm{c}(\mathrm{s}, \mathrm{s}_i) = \infty$ for each $\mathrm{i} = 1, 2, \ldots, \mathrm{m}$. And then we create a new supersink t as well as add a directed edge $(\mathrm{t}_i, \mathrm{t})$ with capacity $(\mathrm{t}_i, \mathrm{t}) = \infty$ for each $\mathrm{i} = 1, 2, \ldots, \mathrm{m}$.

Intuitively, any flow in the network in Fig. 1 corresponds to a flow in this network in Fig. 3, and vice versa. The single source s provides the same flow as desired for the multiple sources s_i, and the single sink t as well as costs the same flow as desired for the multiple sinks t_i.

The Math Mode of Maximum Flow

We describe the establishment of the max-flow mode. Let $G = (V, E)$ be a flow network with a capacity function f, where V is the set of nodes and E is the set of edge.

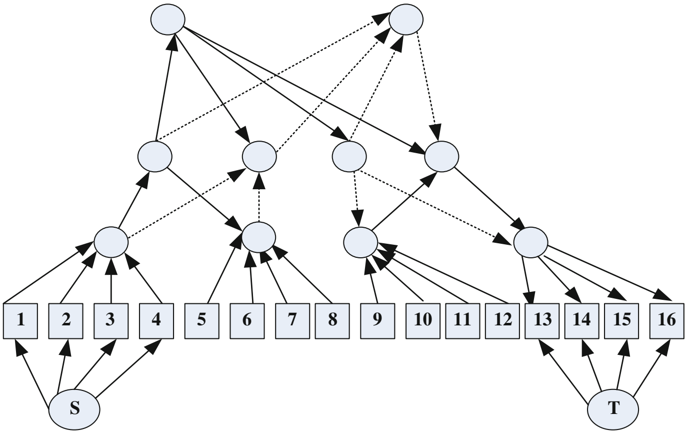


Fig. 3 Maximum-flow for tree

A flow in G is a function of real-valued.

We require that $f(a) : V \times V \rightarrow \mathbf{R}$.

Thus, $\mathrm{f(a)} = \mathrm{C(i,\ j)}$, where $a \in A$.

C (i, j) represents the capacity of network.

f (i, v) represents the total traffic that outflows from node v_i, and f $(v,\ \mathrm{i})$ means the total flow which flows into the node v_i. In the maximum-flow problem, we are given a flow of network G, where s is source node and t is sink node, and our object is to find maximum value of the flow.

According to the conservation principle theory, when the node of v_i is the intermediate node, we can obtained that $f(\mathrm{i}, v) - \mathrm{f}(v, \mathrm{i}) = 0$.

Especially, for the source s and sink t, there will be the formulation,

$$\sum_{j \in N} f(s,j) = \sum_{j \in N} f(j,t)$$

For any network in a (i, j), when the rate of flow increases to the capacity of C(i, j). This path can be seen as saturated path or as unsaturated section. More specifically, if all paths between two nodes contain at least one saturated path, the traffic between two nodes is maximum flow. According to the above analysis, we can establish the maximum flow model.

$$\begin{cases} \max f(s,t) = \sum_{j \in N} f(s,j) \\ f(s,V) - f(V,s) = f(s,t) \\ f(t,V) - f(V,t) = -f(s,t) \\ f(i,V) - f(V,i) = 0, & where\ i \neq s,\ t,\ i \in V \\ \sum_{j \in N} f(s,j) = \sum_{j \in N} f(j,t) & where\ i = sort \\ 0 \leq f(i,j) \leq C(i,j) \end{cases}$$

When the network comes to the maximum-flow, according to maximum flow minimum cut theorem, we can get the critical edges for this flow network.

Model Solution and Algorithm

Model Solution

Based on the formulation which has been proposed in [11], we establish the entire cost matrix for each node in the topology. We show the new expressions of C for the two architectures in our paper.

Firstly, we return the cost function of tree architecture. For the Tree architecture, the cost between two nodes can be expressed as the function of the fan-out of the edge switches (p_0) as well as the fan-out of the aggregation ones (p_1),in addition, when v_i is the source there the weight between v_i and v_j will be ∞. Similarly, when the j is the sink, there the weight between v_i and v_j will be ∞ too. The function of Tree is named as C_{ij}^{Tree}.

$$C_{ij}^{Tree}\begin{cases}\infty & if\ i = s or j = t\\ 0 & if\ i = j\\ 1 & if \left\lfloor\frac{i}{p_0}\right\rfloor = \left\lfloor\frac{j}{p_0}\right\rfloor\\ 3 & if \left\lfloor\frac{i}{p_0}\right\rfloor \neq \left\lfloor\frac{j}{p_0}\right\rfloor and \left\lfloor\frac{i}{p_0p_1}\right\rfloor = \left\lfloor\frac{j}{p_0p_1}\right\rfloor\\ 5 & if \left\lfloor\frac{i}{p_0p_1}\right\rfloor \neq \left\lfloor\frac{j}{p_0p_1}\right\rfloor\end{cases} \qquad C_{ij}^{VL2} = \begin{cases}\infty & if\ i = s\ or\ j = t\\ 0 & if\ i = j\\ 1 & if \left\lfloor\frac{i}{p_0}\right\rfloor = \left\lfloor\frac{j}{p_0}\right\rfloor\\ 5 & if \left\lfloor\frac{i}{p_0}\right\rfloor \neq \left\lfloor\frac{j}{p_0}\right\rfloor\end{cases}$$

Next, we return the function of VL2 architecture. In the VL2, the cost is a function only about the fan-out of the edge switches (p_0), given the traffic that departs the edge switches always passes the core switches. The function is named as C_{ij}^{VL2}.

Maximum-Flow Solution

In this paper, we use Ford-Fulkerson algorithm for the calculation of maximum-flow, which repeatedly increase the flow through the augmenting path until find the maximum flow.

The Ford-Fulkerson method is based on that a flow is a maximum flow if and only if its residual network does not contain the augmenting path.

It is an iterative method. Firstly, it initializes all the flow f to zero. Secondly, we find an augmenting path to increase traffic at every iteration. Repeat this while there exists an augmenting path p unit all the paths have been found. Finally, we will find the value of maximum flow. Specially, use BFS to find each augmenting path at every iteration will greatly improve the efficiency.

The details of the algorithm are described in pseudo-code. It has two major components. (1) Graphmaxflow. (2) Edmonds-Karp Algorithm.

In Alogrithm1, we input the matrix s, t and e, where s and t each represents the source and sink while exists an edge between S and E. In addition, e is the cost matrix for all the nodes in s and t. After the sparse operation, we use: Graphmaxflow (Algorithm 2) for the calculation of maximum-flow.

```
Algorithm1: Maximum-Flow Algorithm
Input: $s$, $t$, $e$, $n$
Output: {Maximum-flow f and Critical Paths p}
matr = sparse (s, t, e, n, n)
[f, p] = Graphmaxflow (matr, s, t)
return {f, p}
```

```
Algorithm2: Graphmaxflow
Input: (G, s, t)
Step1: for each edge(u, v) ∈ E(G)
Step2: do f [u, v] ← 0
Step3: f [v, u] ← 0
Step4: Breadth-first search
while there exists a path p from s to t in the residual network of $G_f$.
Step5: do augment flow f along p
return f[s, t]
```

Experiment Evaluation and Discussion

In this section, we demonstrate a thorough experimental evaluation of the proposed technique on Multi-rooted Tree and VL2 Architectures. The whole experiment project is implemented by Matlab 7.0 on a Windows 7 Operating System with Intel Core i3 Processor 2.10 GHz. Based on the classification of architecture which described in Section Background, we divided into two groups, the former one is Tree architecture, and the other one is VL2 architecture.

Experimental Evaluation

Tree Experiment Design

As shown in Fig. 1, since the nodes within a cluster got the same nature, in order to simplify the complexity of our maximum-flow problem, we choose two nodes in each cluster for our experiment. There are four groups of $\{v_1, v_2, v_3, v_4\}$ $\{v_5, v_6, v_7, v_8\}$ $\{v_{13}, v_{14}, v_{15}, v_{16}\}$ $\{v_9, v_{10}, v_{11}, v_{12}\}$ in the topology. When designing experiments, we only need to choose three groups as group three and group four got the same cost (six nodes for the test, where two nodes in each cluster). Based on the formula cost matrix in Section Model Solution and the Fig. 3, we can get the adjacency matrix for those six nodes of tree topologies. Figure 4 show the initial capacity of Tree architecture. According to the algorithm 1 and algorithm 2, we get the capacity of maximum-flow which is 32, and the corresponding critical edges of $\{e_{24}, e_{25}, e_{26}, e_{27}, e_{34}, e_{35}, e_{36}, e_{37}\}$.

VL2 Experiment Result

Figure 5 shows the capacity figure of the VL2 architecture. The capacity of maximum-flow is 36, and the critical edges are $\{e_{26}, e_{27}, e_{36}, e_{37}, e_{46}, e_{47}, e_{56}, e_{57}\}$

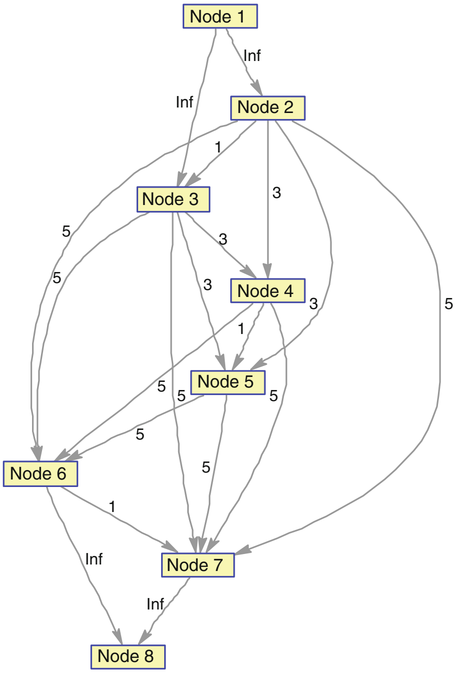


Fig. 4 The capacity of tree architecture

Complexity Analysis and Discussion

Complexity Analysis

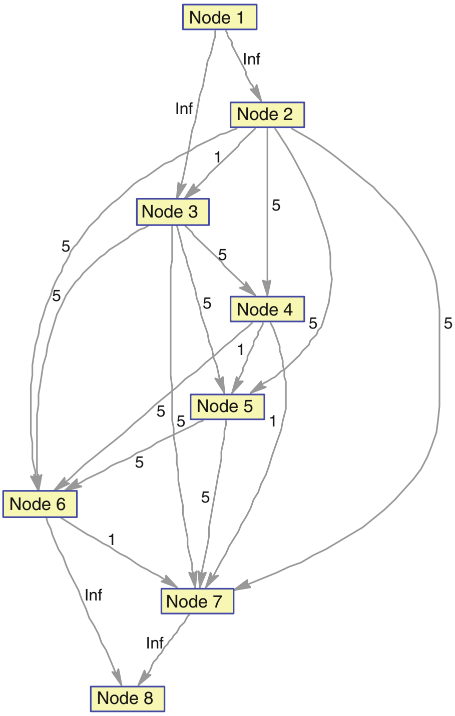
The computational complexity of this Algorithm is determined by Edmonds and Karp algorithm. Implementation is based on a variation called the "labeling algorithm". Time complexity is O (n*e^2), where *n* is the number of the nodes and e is the corresponding edges in the given architecture.

Discussion

As shown in the above experiments (see Section Experimental Evaluation), we can see that our method can find the maximum flow and the corresponding critical edges. The result of complexity analysis also shows that our approach is effective and efficient for the evaluation of the maximum flow. From the perspective of network attack and defense, NSP (Network service provider) should increase the traffic capabilities of the critical links in order to increase the capabilities of whole network.

In this paper, although we choose two kinds of data center network (DCN) topology for our experiments, our method should be generic and can be properly used in any new architecture. Furthermore, we use the cost matrix which has been

Fig. 5 The capacity of VL2 architecture



proposed in the literature [11]. Considering the changes of cost matrix in different scenario, we only need to change the input of the algorithm. Thus, our method can be widely used for the evaluation of maximum-flow for new DCN architectures.

Conclusion

In this paper, we investigate the problem of topology vulnerability in Data Center Network (DCN) Architectures. We abstract two kinds of typical DCN architectures and formulate the performance analysis of architecture as a maximum-flow problem by adding a supersource and a supersink, and then we propose a method to efficiently solve this problem. Furthermore, the experimental results show that our algorithm is effective. In our paper, although we only choose two kinds of architectures while the method and observed issues can be generic and widely used in the newly proposed ones.

Acknowledgments This work is supported by the National Natural Science Fund China under Grant No. 2009CB320406, the National 863 High-tech Project of China under Grant No. 2011AA01A102, Funds for Creative Research Groups of China (60821001) and State Key Lab of Networking and Switching Technology. Ph.D. Programs Foundation of Ministry of Education (20110005130001).

References

1. Vaquero LM, Rodero-Merino L, Caceres J, Lindner M (2008) A break in the clouds: towards a cloud definition. ACM SIGCOMM Comput Commun Rev 39:50–55
2. Peng K, Zou H, Lin R, Yang F (2012) Small business-oriented index construction of cloud data. In: Proceedings in 12th international conference on algorithms and architectures for parallel processing, pp 156–165
3. Guo C, Wu H, Tan K, Shi L, Zhang Y, Lu S (2008) Dcell: a scalable and fault-tolerant network structure for data centers. ACM SIGCOMM Comput Commun Rev 75–86
4. Li D, Guo C, Wu H, Tan K, Zhang Y, Lu S (2009) FiConn: using backup port for server interconnection in data centers. In: Proceedings in 28th conference on computer communications, pp 2276–2285
5. Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S (2009) BCube: a high performance, server-centric network architecture for modular data centers. ACM SIGCOMM Comput Commun Rev 39:63–74
6. Leiserson CE (1985) Fat-trees: universal networks for hardware-efficient supercomputing. Comput IEEE Trans 100(10):892–901
7. Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S (2009) VL2: a scalable and flexible data center network. In: Proceedings of ACM SIGCOMM computer communication review, pp 51–62
8. Liao Y, Yin D, Gao L (2010) Dpillar: scalable dual-port server interconnection for data center networks. In: Proceedings in 19th international conference on computer communications and networks, pp 1–6
9. Shangguang W, Zheng Z, Qibo S, Hua Z, Fangchun Y (2011) Cloud model for service selection. In: Proceedings in 30th IEEE conference on computer communications workshops on cloud computing computer communications workshops, pp 666–671
10. Zhang Y, Su AJ, Jiang G (2011) Understanding data center network architectures in virtualized environments: a view from multi-tier applications. Comput Netw 55:2196–2208
11. Meng X, Pappas V, Zhang L (2010) Improving the scalability of data center networks with traffic-aware virtual machine placement. In: Proceedings in 29th IEEE conference on computer communications, pp 1–9

A Situation-Oriented IoT Middleware for Resolution of Conflict Contexts Based on Combination of Priorities

Z. Cheng, J. Wang, T. Huang, P. Li, N. Yen, J. Tsai, Y. Zhou and L. Jing

Abstract Situation-aware service is recognized as an emerging research issue in ubiquitous computing. It becomes more important and significant with the recent progress in IoT (Internet of Things), since the situations considered in IoT are more complex, become global, and cause more conflict. In this paper, a middleware for management conflict situations was designed, to prompt the development of context-aware services. It is characterized by its ability of situation-oriented, paying attention to relations among users (and situations as well) and smart objects around. Eventually, following issues were solved: (a) a method for detecting (i.e., being aware of) a specific situation, and triggering corresponding service; and (b) an algorithm for conflict situations/contexts management. A diagram of situation state transition (DSST) was proposed to specify states of a situation. A set of situation-oriented ECA rules are presented to reason the situations' states based on sensed data. Policies based on DSST for resolving conflicts were also given. The experiment results demonstrate the feasibility of proposed method, and the performance of proposed situation-oriented policies.

Keywords Internet of things · Context-aware service · Conflict situations · Diagram of situation state transition · Reasonably fair policy

Z. Cheng (✉) · J. Wang · T. Huang · P. Li · N. Yen · J. Tsai · Y. Zhou · L. Jing
School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu City, Fukushima-ken 965-8580, Japan
e-mail: z-cheng@u-aizu.ac.jp

J. Wang
e-mail: j-wang@u-aizu.ac.jp

N. Yen
e-mail: nyyen@u-aizu.ac.jp

J. Tsai
e-mail: jctsai@u-aizu.ac.jp

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_51, © Springer Science+Business Media Dordrecht 2014

Introduction

Situation-aware service is a hot research topic in ubiquitous systems. For examples, a travel information service is presented in [1], which can recommend attractions to users, based on their context information, such as current location, time, and weather etc. It becomes more important and significant with the recent progress in IoT (Internet of Things), since the situations considered in IoT are more complex, become global, and cause more conflict, e.g. many elderly need help in the same time, but the helpers are not enough, especially when a disaster happens in a city.

In this paper, our goal is to design a middleware of management for complex and conflict situations. Especially, we focus on resolution of conflicts among different users' situations, to enhance the development of context-aware applications.

The requests for services by different users may be related or even conflict in the same location and time, in order to provide situation-aware services to each of the user. The conflict often happens, when situation-aware services to different users need common resources. Conflict should be resolved and coordination or scheduling of those services are necessary.

To this end, the following problems have to be solved. (a) a method for detecting (being aware of) a situation, and triggering corresponding services, (b) a resolution method for conflict situations, and (c) a coordination mechanism between related situations. In this paper, we mainly focus on the solutions of the first two problems.

Many approaches have been studied for situation-aware applications [1–6], which can be divided into the following two approaches. The first is applications without using a platform/middleware [1, 2]. Model-based methods are employed to well design applications. However, common functions for handling context have to be realized in each application. The second develops a middleware for dealing with the functions and builds the applications above the platform, to simplify the development of situation-aware services [3–6].

The platform approaches can be further classified into two, depending on whether solutions for the conflicts are provided. Though researches [3, 4] can provide a platform, they cannot be used in conflict situations. On the other hand, [5, 6] can resolve the conflict situations to provide context-aware services. In [5], managing resources is discussed for resolution of conflict between electric devices, e.g., lamp, display and so on. However, the resolution policies are not clear. In real application, the more urgent situation should have more chance to get the services. In [6], Shin and Woo proposed a method based on weights to the historical conflicts. However, it is hard to deal with interruption to a current service, due to lack of sit state manage.

We first present a diagram of situation state transition (DSST) to specify the state of a situation. The diagram shows the phases and states of a situation and the change of them. Executable functions and degree of urgency are bound with each state of each phase. Therefore, the coordination and resolution of conflicts can be simplified by using the DSST. We give policies based on DSST for resolving conflicts, since different states reflect different levels of request urgency. We design a mechanism for reasoning phases and states of the situation based on sensed data. The change of states of a situation is triggered by situation-oriented ECA (Event-Condition-Action) rules, which are an extension of ECA rules for context-aware applications.

In this paper, we first present the Diagram of Situation State Transition (DSST), to describe the change the situation status, and give Situation-oriented ECA rules in order to trigger the creation and management of DSST. Based on DSST, reasonably fair policies are discussed and employed for resolution of the conflict situations of different users. Our advantages of are conflict-free, situation oriented solution, which can deal with interruption of services. In addition, the rate of successful request is reasonably fair, from the viewpoint of urgency and importance, and times of waiting.

The Requirement Model for the Management of Situations

Figure 1 shows the requirement for the management of situations. In Fig. 1, there are elderly or handicapped persons living in their homes. There are also helper(s) and care-robot(s) for support of the persons. Around a person, there are appliances, furniture, etc. It is supposed that some sensors are embedded into appliances, furniture, and rooms. For examples, under the floor, there is RFID sensors, such as U-Tiles [7], which can detect the position and person's ID for collecting the data of situations. A person may wear some wireless sensor device, e.g. a Magic Ring [8], which can be worn on a figure to detect the person's intension for control appliances or a care-robot, and call for the helpers. Some sensors, such as temperature sensor is employed in the room.

By using those sensors, data can be collected, and situation can be reasoned based on the data. For examples, a simple situation could be the handicapped person fell the temperature of the room is too low, the system is aware of his situation, and automatically turn on the heater in the room.

Generally, a situation of a person reflects the user's intension and requirement. The intention or requirement may be directly shown by some gestures, e.g. a person calls the help by shaking his hand, which can be detected by Magic Ring put on her figure. The requirement or needs of the person may not be shown clearly but hidden in his/her activities of life. In other words, they are shown unconsciously.

Another situation might be the person is approaching the table and want to drink his favorite tea; the care-robot will be aware of the requirement and serve the tea for the person. However, there may be a conflict that the user on the bed in the same home may feel bad and ask the care-robot to pick up a medicine, trigged by the signal from her Magic Ring on her finger. Therefore, sometime situations

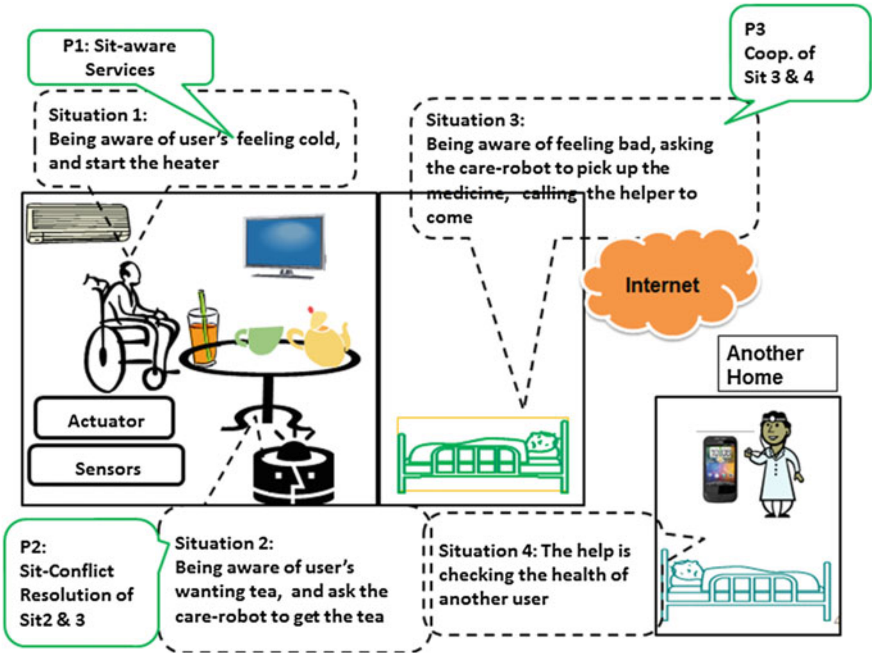

P1: Sit-aware Services
Situation 1: Being aware of user's feeling cold, and start the heater
P3 Coop. of Sit 3 & 4
Situation 3: Being aware of feeling bad, asking the care-robot to pick up the medicine, calling the helper to come
Internet
Another Home
Actuator
Sensors
P2: Sit-Conflict Resolution of Sit2 & 3
Situation 2: Being aware of user's wanting tea, and ask the care-robot to get the tea
Situation 4: The help is checking the health of another user

Fig. 1 Requirement for the management of conflict situations

happened around the two persons may compete for some resources, e.g. the care robot for provision of the situation-aware services. In addition, she may also call the helper for direct service by the helper. However, the helper is taking care of another person in another home. That is to say the situations happened in different homes are related and need coordination.

The Middleware and Conflict Resolution

The Outline of the Middleware

Figure 2 shows the architecture and the position of the middleware in the architecture. There are three layers in the architecture. The lowest one is the sensors and actuators; the middle is the middleware; the highest one is users' applications. The application is situation-awareness, which means they can automatically provide services according to the situations around the users, such as sending a warning signal when a danger may happen, a reminder just before a pre-defined situation happens, or a message to hospital/care center when an elderly person need help. To this end, the sensors are employed to get the data for reasoning the situations, and actuators are employed to perform the services.

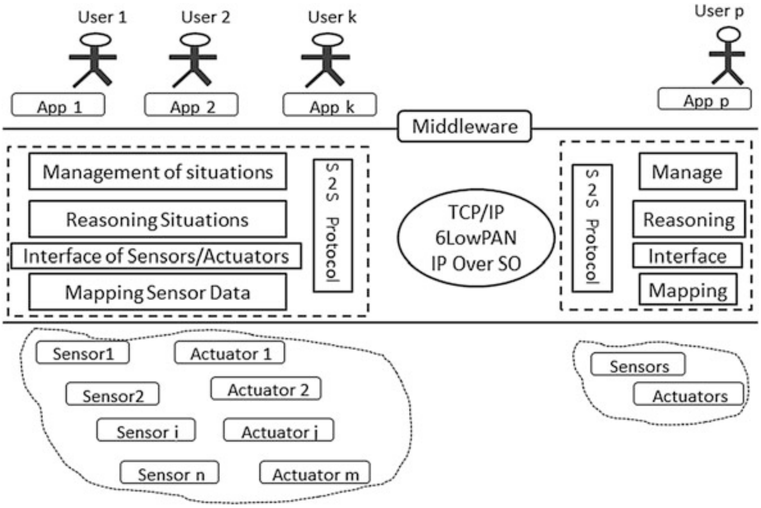


Fig. 2 The architecture and outline of the middleware

The middleware is to help the applications. Many common functions for dealing with the situations can be encapsulated in the middleware, such as creating situations, maintaining a situation, resolving the conflict situations, and coordination of situations. So the developer can concentrate on the design of services and conditions for triggering the services considering real situations.

The middleware is built above the internet currently. It will be built above some core technologies of IoT such as 6LowPAN and IP over smart object in the future. The main functions of the middleware are,

- For each kind of sensors, we assume driver software of the sensors is available to collect the data and transform them into the pre-defined regular data form.
- Reasoning the situations based on those data, which means to explain the meaning the data and relation of the data, by some formal way, such as SoECA.
- Each situation has a life time from start to the end. During life time, there are a series of the states, in each of which different functions or services can be triggered to provide services to user, resolve conflict, etc.
- A protocol called *situation to situation* communication (S2S) is necessary, for dealing with the coordination between related situations happening in different locations. (The protocol will be developed in the future)

Diagram of Situation State Transition

Each situation will be managed by a situation manager, which is called *situator*. There are more than one *situators* corresponding to multiple situations in the middleware. There is a special *situator* called *arbiter*, which is for solving the

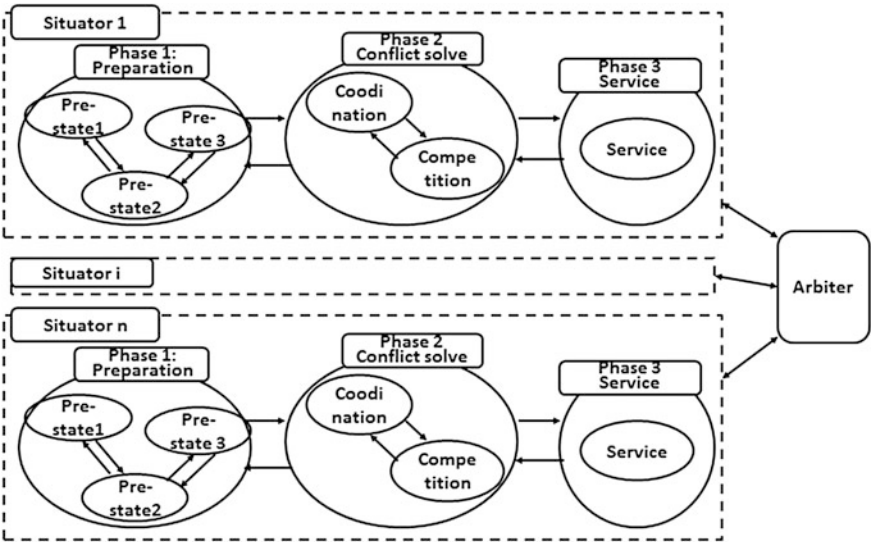


Fig. 3 Diagram of situation state transition

conflicts among the situations. Each *situator* communicates with the *arbiter*, as shown in Fig. 3.

A situator for a situation is created, when at least one condition defined for the situation becomes true, which can be detected by sensors. Each situator has three phases in its life time, i.e. preparation, coordination, and service. The first phase preparation starts, when some data related with a situation is detected by some sensors. Though the conditions for triggering the awareness service defined on the situation are partially satisfied, preparation for provision the pre-defined service, e.g. to find and reserve the necessary device and resource will be performed at this phase.

The second phase is mainly for resolving the conflict with other situations, e.g. completing for a device to provide the services for different users' situations. At this phase, each *situator* sends messages to the arbiter about their requests with the importance and urgency of the requests. The arbiter will make the decision to solve the conflict, based on priority scheme/policies, described later.

The 3rd phase is to provide services to the user, e.g. sending a warning message to avoid danger, etc.

The transition between phases is bi-directional, e.g. when the situator is providing a service, some new requests from other users with higher priority happen, the phase 3 will return to phase 2 for resolving the conflicts. Moreover, if the conditions for triggering a service are changed based on newly detected data, the situator will change its phase to preparation.

A situator will be terminated, when the all services for such situation have been finished and/or all the conditions for the situation become false.

There may be one or more states in each phase, to represent the stage in each phase, such as the level of preparation, degree of urgency for the services, etc. The transition between the states is also bi-directional, and the situator changes its state depending on the changes of situation conditions.

Policies for Resolving Conflicts

The *arbiter* will resolve the conflict based on some policy. Generally speaking, *fairness* is the first policy for resolving the conflict. That is, every user's situation should be allocated the resources in a fair way. For example, a *turn* based resolution is to give the high priority to each user in turn. However, if urgency of the requests from different users is not the same, the *turn* based policy is not really fair and reasonable, in the sense that some users should be paid more attention, since his/her situation is more urgent and crucial. Therefore, we use *degree of urgency*, as a criterion for the resolution. The urgency of the request is different in various situations, when a user need help in his/her everyday life, in an accident, or in a disaster. The user's situation for the services can be detected and reasoned by the system, or directly expressed by the user by some way, such as gestures detected by wearable sensor device. The types of different gestures can express the degree of the urgency, e.g. asking the system to help, call for helper, or SOS. If the degree of urgency of two requests is the same, degree of the importance of the users will be also considered, e.g. the worse the user's health condition is, the higher priority will be given, or the ages of the users (say 90', 80', 70', 60') are used.

The policies are managed based on the situation diagrams. Different states reflect level of satisfaction of conditions for providing services, and degree of urgency for the service. In the same level of state, the more important request will be served with higher priority. Preemption is also employed, which means a request with higher priority will interrupt the service of a lower priority being provided. And the interrupted and pending service will be resumed when the higher priority service is finished. First-come-first-served policy will be used, if the levels of the importance and urgency of two situations are the same.

The arbiter runs the following algorithms.

```
//* Dealing with the request and put it into a queue
When the arbiter receives a request (user's phase and state, etc.) from a situator
Check User Context Check carried by the parameters of the request.
Put the Request into the priority Queue based on the following combined policies
//* Check the requests in the queue, and compare them with the new request
Prior-based Policy
step 1. Check degree of urgency (depending on the phases/states reflecting urgency)
step 2. Check degree of importance (users' profile information such as ages)
step 3. Update request queue, by inserting the request into the proper position.
```

Fairness-Tuning Rule
step 1. Check if taking over request exist (times of waiting over default 5)
step 2. Update the priority of request being taken over.
//* Services Provision
Assign devices to provide services based on the request queue.
Preemption-based Policy is used to interrupt the current service if necessary.
Except a user's services is preempted too many times (default at 5)

Reasoning of Situations

As shown in Fig. 4, there are four main parts for reasoning situations. They can be divided into *specification* and *detection*.

Specification

The description is based on the concepts of situation theory [9], which specifies a situation using 4 elements, i.e. individual, properties, relations, and spatial-time location. The specification is presented in a formal way, such as first-order predicates.

An example of situation could be a handicapped person feels thirsty and wants to drink a cup of the tea, since it is a hot day and high humidity. In this example, the individual is the person as well as the smart objects, such as the cup embedded with sensors to know if it is empty or not. The properties are the person is

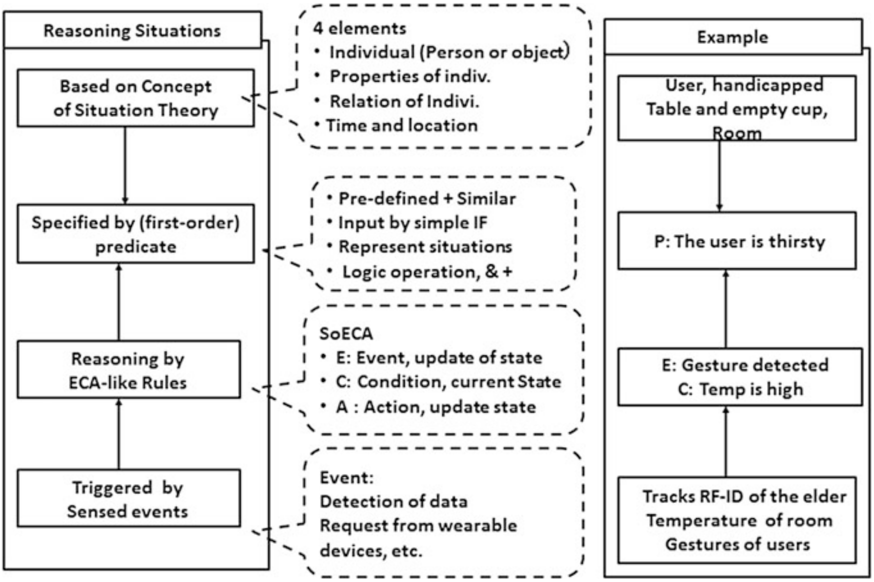


Fig. 4 Reasoning situations

handicapped so that it is hard to get the tea by himself, and the tea cup is empty. The spatial-time location is current room and time. Thus, the situation could be represented as follows

P The user feels thirsty
Q The user is hard to pick up the cup of the tea.

Services actions are defined on a situation, e.g.
P and Q => SA
where, SA (service action) is to ask the robot to get the tea.

Detection (Situation-oriented ECA, simply called SoECA)

Sensors are used to detect the raw data for reasoning a situation, and ECA rules are employed to reason the situation based on the data. The ECA is situation-oriented, which means *Event*, *Conditions,* and *Actions* are related with the situation states. *Event* could be detection of a new data or an update of situation states. *Condition* could be a condition for the service on the situation and current state. *Action* could be a trigger for service, or update of the state of the situation for giving the service later.

SoECA rule:
E an event or an update of situation state
C a condition for a service and current state of the situation
A a service action and/or updating state of the situation.

Some examples of ECA are as follows.

SoECA rule 1:
E Temperature is high than a threshold & humidity high than threshold
C *situator* has not been created
A Create *situator* and set its state to 1st state (S1) of 1st phase with the lowest priority.

SoECA rule 2:
E User1 is approving the table (detected by RFIDs and its position)
C The cup is empty & the current state is S1 of 1st phase
A Ask the robot to get the tea & the state to 2nd phase.

SoECA rule 3:
E User2 did gesture *G*1 for help with picking up medicine
C Situator has not been created
A Creation of the *situator* and set the state to 2nd phase.

SoECA rule 4:
E User2 did a gesture *G*2 for help with getting a drink
C Situator has not been created)
A Creation of the *situator* and set the state to 2nd phase.

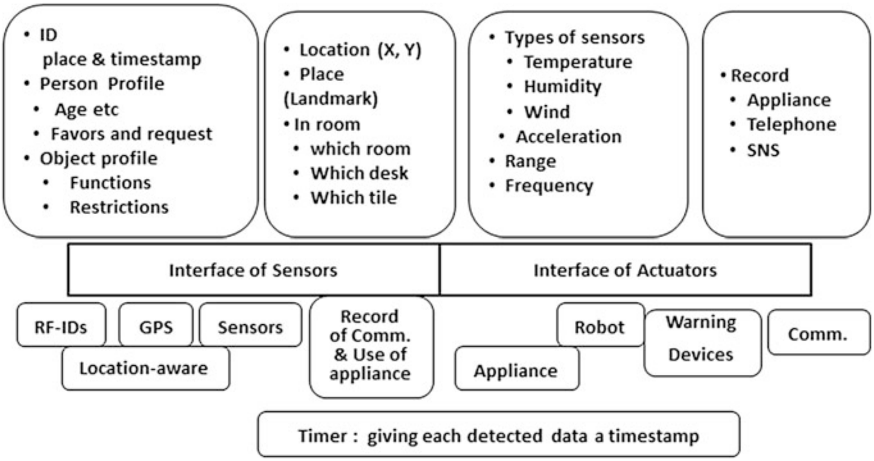


Fig. 5 Interface of sensing/actuation

Generally, a collection of SoECA rules will be designed by developer of the application and/or the services providers who has the experts in the domain, based on domain knowledge and engineering design knowledge.

Ideal Interface of Sensors and Actuators

As shown in Fig. 5, many sensors are available in current technologies, e.g. RFIDs, GPS, various environment sensors, such as temperature, humidity, wind, as well as acceleration, etc. The mobile or wearable devices can detect demands of a user, activities of the user, and vitality data of a user. In addition, the usage of appliance, furniture, care-robots, and other actuators can be recorded and used as historical data. The sensed data can be transformed into a regular form. The time and places when the data is detected is also useful for situation reasoning.

Feasibility Study

Example of Conflict Situations

In this paper, we show how the conflict resolution is realized using an example to investigate the feasibility of the implementation. As shown in Fig. 6, we assume that User1 is a handicapped person in the left room and User2 is an elderly person

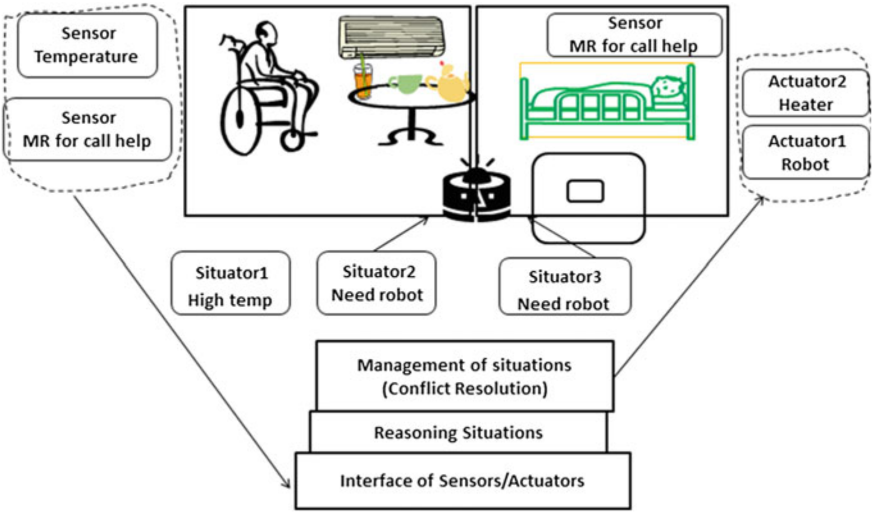


Fig. 6 Implementation example

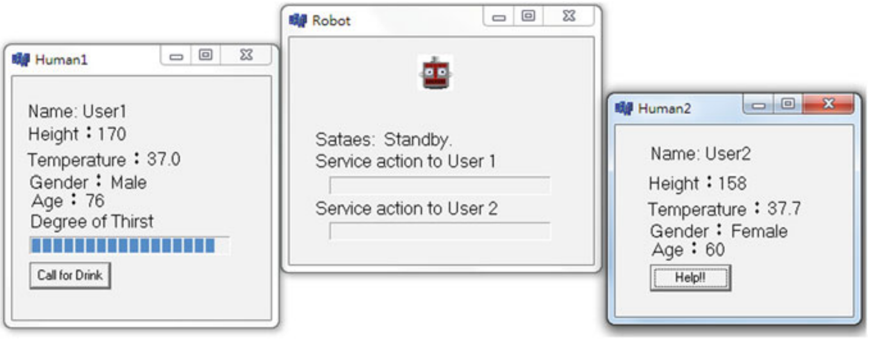
lying on the bed in the right room, and care-robot is employed in the home for services.

We assume that User1's situation is he wants to drink the tea, since the weather is hot and he has not drunk for a while. The temperature can be detected by the sensor, the User1 can be detected by RFID on his body, and the time duration from last drinking can be detected, by a timer and smart cups (with sensor). User2's situation is she wants help by the care-robot to pick up drink or medicine. The request can be sent by using gestures, which can be detected by the figure-worn device Magic Ring [8]. In this example, two gestures are used. Gesture 1 and 2 mean requesting a drink only or medicine, respectively.

Implementation of the Example

We have implemented the example using C++ on Window 7. *Situators* of User1, User2, and the care-robot are implemented as classes. The *arbiter* is implemented as another class, which receives the requests from classes of User1 and User2, and controls the care-robot for the service.

We first show the software for conflict situations. In Fig. 7, three windows show the running information of objects of User1, User2, and the robot classes. Initially, the robot character is on the middle of the window, which means there is no request for it. The *progress bar* in the left window is used to show the thirsty degree. After a drink, the progress bar is set to *full* (*blue*), to mean the user is not thirsty. With time elapses, the blue part is reducing, which means the degree of



Name: User1
Height：170
Temperature：37.0
Gender：Male
Age：76
Degree of Thirst

Sataes: Standby.
Service action to User 1
Service action to User 2

Name: User2
Height：158
Temperature：37.7
Gender：Female
Age：60

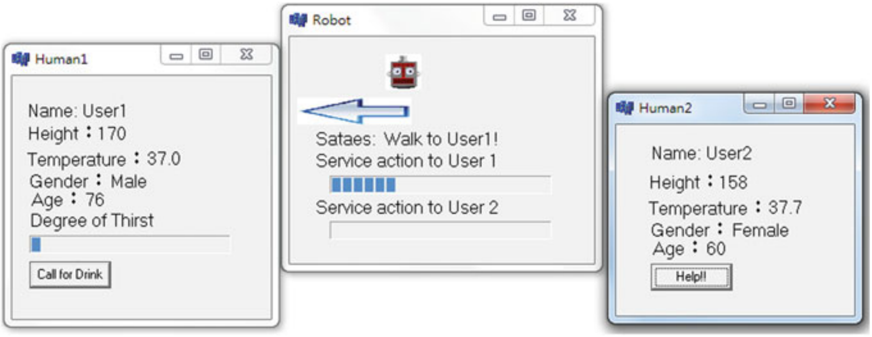
Fig. 7 User1, user2 and robot

thirsty increases. The speed of the reducing is depending on the temperature. User2 can push the button to ask the robot to get help directly, which can be realized by a gesture detected by Magic Ring.

In Fig. 8, when user1 feels thirsty (the progress bar becomes lower than 20 % of the full), the robot will walk to him to provide a service. The progress bar in the middle window show the service by the robot is going on.

In Fig. 9, if User2 uses the button to call the robot, the robot will break the service to user1 and walks to User2 to help her, even when the robot is providing service to User1, since the priority of User2 is higher than User1's. In Fig. 10, the User2 can really use a wireless sensor to make request. The conflict is resolved, same as the above discussed.

To evaluate our method is workable and solve the conflicts reasonably. We use rate of satisfied requests denoted with R_i to represent the percentage of getting the resource when conflict. That is, $R_i = N_i/N$, i belongs to {1, 2}. The number N_i is the times $user_i$ won the competing resource. N is the number of total competing times.

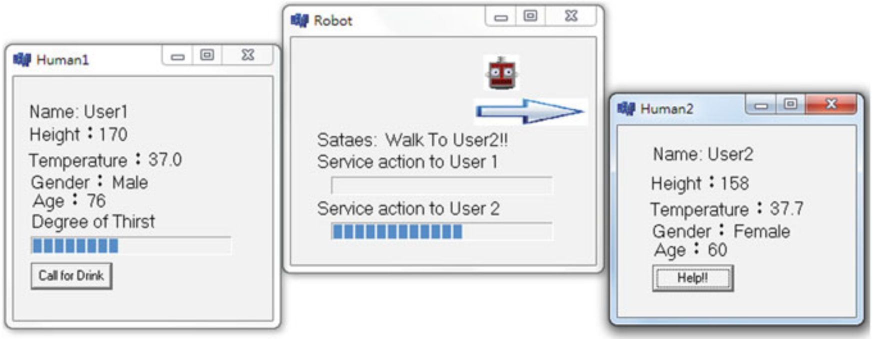


Name: User1
Height：170
Temperature：37.0
Gender：Male
Age：76
Degree of Thirst

Sataes: Walk to User1!
Service action to User 1
Service action to User 2

Name: User2
Height：158
Temperature：37.7
Gender：Female
Age：60

Fig. 8 User1 becomes thirsty and robot comes to user1

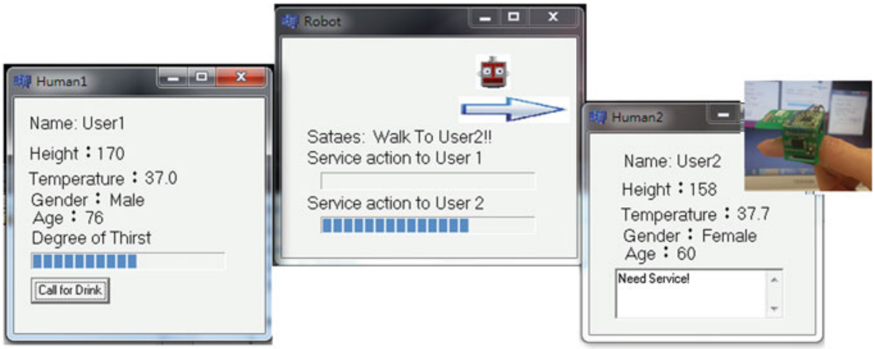


Name: User1
Height：170
Temperature：37.0
Gender：Male
Age：76
Degree of Thirst

Sataes: Walk To User2!!
Service action to User 1
Service action to User 2

Name: User2
Height：158
Temperature：37.7
Gender：Female
Age：60

Fig. 9 User2 asks robot for service



Name: User1
Height：170
Temperature：37.0
Gender：Male
Age：76
Degree of Thirst

Sataes: Walk To User2!!
Service action to User 1
Service action to User 2

Name: User2
Height：158
Temperature：37.7
Gender：Female
Age：60
Need Service!

Fig. 10 User2 uses gestures detected by MR to make request

We compare User1 and User2 in various states. For each case, we run the example mentioned above 1,000 times, record the result, and compute R_i, as shown in Table 1.

The degree of User1's urgency, is in such as a way that no-need < thirty < request < drinking. The urgency of User2's gesture is "for a drink < medicine. User1's request for a tea is almost equal to User2's request for a drink. A pure turn based police means to give the higher priority to each user alternately. Generally, a turn based policy lead to Ri = 50 %. Our method also combines an extension of turn based method. Times of waiting for recourse are used, when the waiting times increase the priority will be increased.

Table 1 Simulation results of conflict resolution

User1's state	Thirsty (%)	Request (%)	Drinking (%)	No-need (%)
User2 requests for a drink	R1: 29.9 R2: 71.1	R1: 54.7 R2: 46.3	R1: 94.8 R2: 5.2	R1: 0 R2: 100
User2 requests for medicine	R1: 16.1 R2: 83.9	R1: 20.7 R2: 79.3	R1: 20 R2: 80	R1: 0 R2: 100

Conclusion

The contributions of this paper are as follows. At first, we proposed architecture of IoT middleware, which described composition of the middleware and relations with situation-aware applications. The main contribution is that we presented the *diagram of situation state transition* (DSST) and conflict resolution policies based DSST. We also presented SoECA rules for reasoning states of situations. Simulation shows our proposal is feasible and effective for solving conflicts.

Out proposal is significant, in the sense that it can well represent changes of user's situation using state transition and degree of urgency in each state, and resolve the conflict between situations, based on the urgency represented by phases and states. The middleware can be a platform for developing situation-aware service application.

In the future, we will complete the middleware and evaluate of the proposal extensively. We will improve DSST in order to manage the situations elegantly. Furthermore, we will give a protocol of situation to situation (S2S) to enhance the coordination of different situations happening in different locations.

References

1. Kapitsaki GM, Prezerakos GN, Tselikas ND, Venieris IS (2009) Context-aware service engineering: a survey. J Syst Softw 82(8):1285–1297
2. Prezerakos GN, Tselikas N, Cortese G (2007) Model-driven composition of context-ware web services using ContextUML and aspects. In: Proceedings of the IEEE international conference on web services 2007 (ICWS'07), Salt Lake City, USA, pp 320–329, 9–13 Jul 2007
3. Zhu J, Oliya M, Pung HK, Wai CW (2010) LASPD: a framework for location-aware service provision and discovery in mobile environments. In: Proceedings of the 5th IEEE Asia-Pacific services computing conference (APSCC 2010), Hangzhou, China, 6–10 Dec 2010
4. Pung HK, Gu T, Xue W, Palmes PP, Zhu J, Ng WL, Tang CW, Chung NH (2009) Context-aware middleware for pervasive elderly homecare. IEEE J Sel Areas Commun 27(4):510–524
5. Hanssens N, Kulkarni A, Tuchida R, Horton T (2002) Building agent-based intelligent workspaces. In: Proceedings of ABA conference, pp 675–681, June 2002
6. Shin C, Woo W (2005) Conflict resolution method utilizing context history for context-aware applications. In: Proceedings of pervasive 2005 workshop, Munich, Germany, pp 105–110
7. Wang J, Cheng Z, Jing L, Ota K, Kansen M (2010) A two-stage composition method for danger-aware services based on context similarity. IEICE Trans Inf Syst E93-D(6):1521–1539, Jun 2010
8. Jing L, Zhou Y, Cheng Z, Wang J (2011) A recognition method for one-stroke finger gestures using a MEMS 3D accelerometer. IEICE Trans Inf Syst E94-D(5):1062–1072
9. Barwise J, Perry J (1983) Situation and attitudes. Center for the Study of Language and Information Stanford, CSLI Publication, California

Data Transmission Mechanism in Cluster-Based Wireless Sensor Networks with Mobile Sink

Ying-Hong Wang, Yu-Jie Lin and Shao-Wei Tsao

Abstract In traditional wireless sensor networks, the sensed data will be forwarded to sink by multi-hop, so that the hot spot problem will happen on the nodes near the sink. The power consumption of these nodes is higher than others, these nodes will be dead sooner. It leads to the network lifetime decreasing. We proposed a data transmission mechanism using mobile sink to solve this problem. We use two threshold values to prevent data overflow. If the data buffer of cluster head exceed to 1st threshold, the node will call sink to come and transmit data. If the sink didn't arrive the transmission range of the node, and the data buffer exceed to 2nd threshold. Instead of forwarding data to sink directly by multi-hop, we use forwarding data to cluster head of neighboring cluster to help buffer the exceeded data. It can decrease the number of node which has to help to help to forward data, so that it can decrease the power consumption of network, and prolong the network lifetime.

Keywords Wireless sensor networks · Multi-hop · Sink · Hot spot · Overflow · Cluster head · Network lifetime

Y.-H. Wang (✉) · Y.-J. Lin · S.-W. Tsao
Department of Computer Science and Information Engineering, Tamkang University,
New Taipei, Taiwan, Republic of China
e-mail: inhon@mail.tku.edu.tw

Y.-J. Lin
e-mail: yeanling319@gmail.com

S.-W. Tsao
e-mail: how4205@yahoo.com.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_52, © Springer Science+Business Media Dordrecht 2014

Introduction

In the operation of the wireless sensor network, how to response the sensed data to the sink is one of the most important issues [1–4]. In the network environment of traditional wireless sensor network, sensors near the sink have to forward the sensed data of other sensors frequently, so the power consumption of these nodes is very seriously. It may lead to sensor dead sooner. This phenomenon is called hot spot problem. To solve the hot spot problem, [5] it has proposed a lot of paper which use mobile sink to solve this problem recently [6, 7]. With the movement of the sink, the neighbor node of the mobile sink will be different, so it can balance the power consumption. In cluster based wireless sensor network system, the sensed data is forwarded to cluster head, and transmit to sink when the sink is in the transmission range of the cluster head. Because the number of sensors in each clusters are different, so the amount of transmitted data in each cluster are different too. However, before mobile sink moves to the destination node, it will lead to data overflow because of the data transmission too much, that will make data loss [6]. To avoid data loss which is because of data overflow, the exceeded data has to transmit to sink before data overflow, but it has to consume additional power to transmit data. So we propose a mechanism to control the data amount of the cluster head. We use two thresholds to achieve the proposed mechanism. First threshold use to call sink to come, and the second threshold use to find the neighboring cluster head to help buffer the exceeded data.

Related Works

Mobile wireless sensor network can be divided into three category, the first one is mobile sink with fixed sensor nodes, the second one is fixed sink with mobile sensor nodes, the last one is both sink and sensor are mobility. Our paper is fall in first category. The existence type of mobile sink with fixed sensor nodes wireless sensor network routing protocols can be divided into two types [7]. The first type is the sensed data through multi-hop transmission to the mobile sink. Another type is the sensed data store in memory of sensor, and waiting for the mobile sink moves to the communication range of the sensor node and then transmit data to the mobile sink directly. By multi-hop, it can response the data immediately, but we can't measure the moving direction of the mobile sink, so we usually use broadcast to forward data, but it will consume lots of energy because of the unnecessary transmission [7]. By one hop, the advantage of this type is the data through other sensor nodes do not have to jump through a multi-point to generation therefore can reduce the wear and tear of the overall network power, but it can't provide real-time transmission, and the sensed data has to store in the memory and waiting for the mobile sink moves to its transmission range, it may lead to data overflow [6].

The Proposed Algorithm

In this paper, we proposed a Non-Real Time Data Transmission Mechanism with Mobile Sink in Cluster-based Wireless Sensor Networks (NTDT). In the proposed mechanism, we use two thresholds to control the data buffer. When the data amount exceed to first threshold, it will broadcast a query to call sink. When the data amount exceed to second threshold, it will broadcast a query to ask neighboring cluster head to help buffer the exceeded data, it can decrease the data loss rate caused by data overflow. The Fig. 1 is the system architecture.

Network Architecture and Assumptions

We deploy hundreds to thousands of sensor randomly in our interested environment and the sensor has fixed on the location. We assume all the sensors are homogeneous, which has the same energy, memory size, computing ability, and transmission range. And every sensor has the following node information table (NIT), as Table 1.

And there is one mobile sink, it has no resource constrain. When the mobile sink moving, it broadcast a time stamp to its one hop neighboring sensor every 2R

Fig. 1 The system architecture

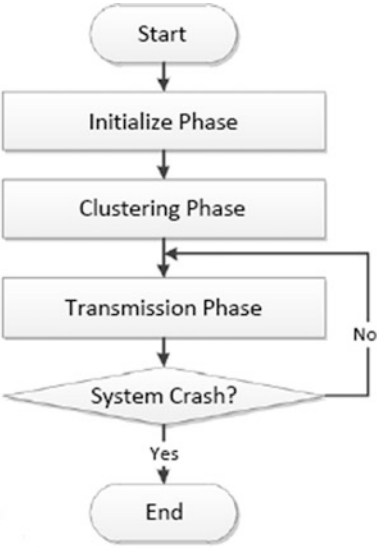


Table 1 Node information table (*NIT*)

Node ID	Cluster ID	N mode	TS

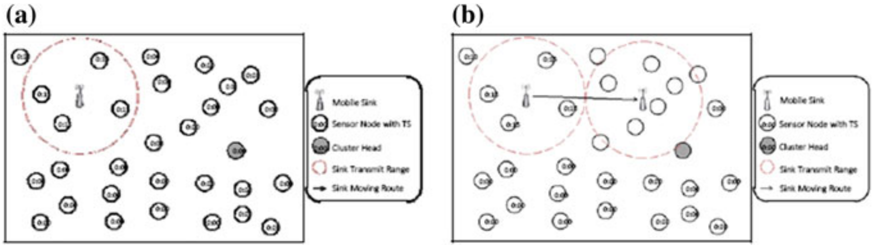


Fig. 2 **a**, **b** Sink broadcasts time stamp

distance (R is the transmission range of the sink). When the sensors receive a time stamp, they will update their TS in NIT. Figure 2a, b show the action of sink broadcasts the time stamp and sensors receive the time stamp, respectively.

The TS value updates by sink broadcasts time stamp, so that the greater TS value indicates that the node is the closer the data collection point.

The Definitions of Message Packets and Record Tables

Except the sensed data, we use some message and record table to execute our mechanism. The first one is Query Message, and there are two query types, the first one is 1st threshold query (Q_1) which is uses to call sink, the second one is 2nd threshold query (Q_2) which is used to ask neighboring cluster head to help buffer exceeded data. The two types of query use the same format packet as Table 2.

When a sensor receives a Q_1 or Q_2 query, if the query hasn't been received, it will be store in Query Record Table (QRT) which is used to avoid the duplicated query. It's shown in Table 3.

After a cluster head receive a Q_2 query which is sent from other cluster, it will reply a Residual Memory Reply (RMR) to the Q_2 query generator. RMR message format is shown in Table 4.

Table 2 Query message

Cluster ID	Q ID	Query type	Q TH

Table 3 Query record table (*QRT*)

Cluster ID	Q ID	Last hop	Query type

Table 4 Residual memory reply (*RMR*)

Cluster ID	RMR ID	TS	Residual memory

Table 5 Ack record table (*ART*)

Next hop	Q ID	Cluster ID	Ack type

Table 6 Residual memory table (*RMT*)

TS	Cluster ID	Residual memory	Last hop

Table 7 Exceeded data packet

Destination cluster	Data payload

Table 8 Query table (*QT*)

Cluster ID	Q ID

When a sensor node receives a message from the other sensor node, it will reply an Ack to the last hop of the message that the sensor nodes have received this message receives the Ack will be recorded in Ack Record Table (ART), the tabular format as shown in Table 5.

When the generator of Q_2 query receives a RMR from other cluster, the RMR will record in Residual Memory Table (RMT), the tabular format as shown in Table 6.

After the generator of Q_2 query receives the RMR, it will forward the exceeded data to the chosen neighboring cluster head according to the TS value in the RMT. And the data packet format shows in Table 7.

When sink receive a Q_1 query, it will be record in Query Table (QT) which is a table built in sink. The format is shown in Table 8.

Data Transmission Phase

After initial phase, it will begin the data transmission phase, we will divide this phase to cluster head, normal sensor, and sink to introduce.

Cluster Head Receive Message Processing Strategy

After cluster head receive a message, first it will determine what kind of the message is, if it is a query, it will execute query dissemination scheme in CH (QDSCH), if the received message is a data, it will execute Data Packet Dissemination Scheme in CH (DPSCH), if the received message is an Ack, it will record in ART, if the received message is a RMR, it will execute RMR Dissemination Scheme in CH (RDSCH), if the received message is a time stamp from sink, it will update the TS value in NIT, and transmit data to sink.

In QDSCH, it will determine the query has received or not by QRT. If not, it will determine the query type, if it is Q_2, it will determine where the query comes from, if it is from other cluster, it will reply a RMR. If it is Q_1, if the Q_{TH} is equal or less than TS in NIT, it will be forwarded continue.

In DPHSCH, after receiving data, it will be stored in memory of sensor, if the data amount exceed to 1st threshold, it will execute Data Buffer Control (DBC), when data exceeded to first threshold, it will broadcast a Q_1 query to sink to ask sink some to transmit data. Before sink arrive, if the data amount exceed to 2nd threshold, it will broadcast a Q_2 query to ask neighboring cluster head to help buffer data.

In RDSCH, it will determine the RM value of the RMR, if the value less than a threshold RM_{TH}, it won't be record in RMT.

Normal Sensor Receive Message Processing Strategy

After normal sensor receive a message, as cluster head, it will determine what kind the message is, if it is a query, it will execute Query Dissemination Scheme in Sensor (QDSS), if it is a data, it will execute Data Packet Dissemination Scheme in Sensor (DPDSS), if it is an Ack, it will be record in ART, if it is a RMR, it will execute RMR Dissemination Scheme in Sensor (RDSS), if it is a time stamp from sink, it will update the TS value in NIT.

In QDSS, it will determine that the query has received or not, if it hasn't received, it will determine what kind the query type is, and determine whether it will be forwarded or not, as the judgment of cluster head. In DPDSS, it will determine where the data come from, if it comes from the same cluster, it will be forwarded to cluster head, if it comes from other cluster, and it will determine that the data packet is on the right route, if it is not on a right route, it will be discarded. In RDSS, it will determine that the RMR is on a right route, if it is, it will be forwarded.

Sink Moving Scheme

Moving scheme of mobile sin in our mechanism, it uses random way point, after receive a Q_1 query it will change its way toward to the Q_1 query generator. When sink is moving into a new cluster, it will find the cluster head of current cluster to transmit data, after receiving data it will delete the cluster information from QT until the QT is empty.

Simulation an Analysis

We use the network simulator tool NS2 2.29 to show our performance. We compare the two papers, "An energy and delay aware data collection for mobile sink Wireless Sensor Networks based on clusters" (HDCA) and "An Efficient Data-Driven Routing Protocol for Wireless Sensor Networks with Mobile Sinks" (DDRP).

We will introduce the assumptions and the parameter setting in Section Assumptions and Parameter Setting. And we will show the comparison with different data generation ratio and different moving speed of the mobile sink.

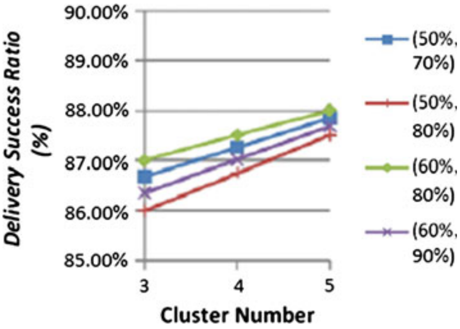
Assumptions and Parameter Setting

The environment settings and the parameters used in our simulation process is as follows.

- Network size is 150*150 m
- Deploy 100 sensors randomly
- Number of mobile sink is 1
- Transmission range of sensor is 15 m
- Initial power of sensor is 2 J
- Memory size of sensor is 50 K Bytes
- Data packet size is 100 Bytes
- Simulation time is 3,600 s.

To find the most appropriate number of cluster, we first compare the data transmit successful ratio between different thresholds and different number of clusters. The moving speed of sink is fixed to 5 m/s, and we set the first threshold as 50 and 60 % will the second threshold different to 20 and 30 %, and the number of cluster set to 3, 4, 5 to do the comparison, the result shows in Fig. 3.

Fig. 3 Data delivery success ratio with different threshold and number of cluster



Because the data in 1st threshold is transmitting by 1-hop, the higher value of the first threshold is, the more complete data is. The data between the 2nd and 1st threshold is transmitted to neighboring cluster head by multi-hop, so that it will lead to data loss caused by the national loss. In different number of the cluster, the less number of clusters is, the greater cluster size is, so that the data will be forwarded by much more node, the data loss ratio will increase.

Simulation and Analysis

In this section, we will compare the performance between, HDCA, DDRP, and our mechanism.

First, we compare the difference of the packet generation interval, and the cluster number fixed to 5, speed of mobile sink fixed to 5 m/s, the difference is shown on Fig. 4. By simulation, we can see the packet loss ratio in DDRP is the highest, 10 % more on average than our proposed NTDT, because DDRP uses multi-hop to transmit data, the major factor is the route time expired. However, if the data generate too fast, it will lead to data overflow, 3 % on average than our proposed NTDT.

Next, we compare different moving speed of mobile sink with packet generation ratio fixed to 3 s., the performance is shown on Fig. 5. By simulation, we can see the packet loss ratio of DDRP is still the highest one. In HDCA, although the packet generation interval is slow, it will happened data overflow caused by moving speed of mobile sink, and the loss ratio is 5 % more than our proposed NTDT.

The last one we compare the system residual energy, moving speed of mobile sink is different, and the packet generation ratio is fixed to 3 s. The performance is shown on Fig. 6.

The major factor of power consumption is data transmission. By simulation, we can see DDRP is the most serious one, because it uses multi-hop to transmit data, and the energy consumption is 15 % more than our proposed NTDT. HDCA use the same type of mobile WSN, as ours, it buffers data in cluster head, and wait sink

Fig. 4 Packet loss ratio with different packet generation interval

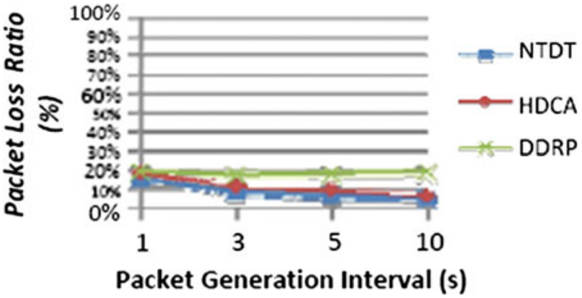


Fig. 5 Packet loss ratio with different moving speed of mobile sink

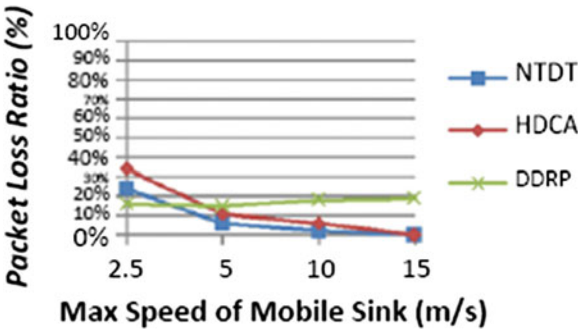
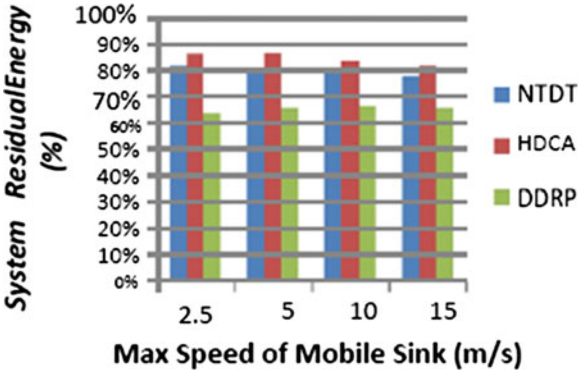


Fig. 6 System residual energy with different moving speed of mobile sink



move to its 1-hop range to transmit data. To avoid data overflow, we forward the exceeded data, so the power consumption of our proposed mechanism is more than HDCA about 5 %.

Conclusion

The most important problem in wireless sensor networks is the energy and the integrity of the data collection; however, the two types of the mobile wireless sensor networks have their own advantages and disadvantages. With non-real time data transmission, the efficacy of the type of data buffered in memory and use one-hop to transmit data to sink is better than the type of data transmitted directly by multi-hop. In this paper, we propose a mechanism, which can avoid data loss caused by data overflow. And we consider the power consumption problem, instead of transmitting the sensed data by multi-hop, we use the query whose packet is smaller than the data packet, to call sink, to decrease the power consumption of the data transmission. By simulation, our power consumption is slightly higher than the other papers, but the difference is not much. We provide the better data protection avoid data loss caused by data overflow.

References

1. Lin CJ, Chou PL, Chou CF (2006) HCDD:hierarchical cluster-based data dissemination in wireless sensor networks with mobile sink. In: Proceedings of the international conference on wireless communications and mobile computing, pp 1189–1194, Jul 2006
2. Lee D, Park S, Lee E, Choi Y, Kim SH (2007) Continuous data dissemination protocol supporting mobile sinks with a sink location manager. In: Proceedings of the Asia-Pacific conference on communications, pp 299–302, Oct 2007
3. Soyturk M, Altilar T (2007) A routing algorithm for mobile multiple sinks in large-scale wireless sensor networks. In: Proceedings of 2nd international symposium on wireless pervasive computing, Feb 2007
4. Heinzelman WR, Chandrakasan A, Balakrishnan H (2002) An application-specific protocol architecture for wireless microsensor networks. IEEE Trans Wireless Commun 1(4):660–670
5. Ye M, Chen G, Wu J (2005) An energy-efficient unequal clustering mechanism for wireless sensor networks. In: Proceedings of the IEEE international conference on mobile adhoc and sensor systems conference, pp 597–604, Nov 2005
6. Yu H, Kuo M (2010) An energy and delay aware data collection for mobile sink wireless sensor networks based on clusters. In: Proceedings of the international conference on computer application and system modeling (ICCASM), pp 536–540, Oct 2010
7. Shi L, Zhang B, Huang K, Ma J (2011) An efficient data-driven routing protocol for wireless sensor networks with mobile sinks. In: Proceedings of the IEEE international conference on communications (ICC), pp 1–5, Jun 2011
8. Chang Sh, Merabti M, Mokhtar HM (2007) Coordinate magnetic outing for mobile sinks wireless sensor networks. In: Proceedings of the international conference on advanced information networking and applications workshops, vol 1. pp. 846–851, May 2007
9. Zhou Z, Xang X, Wang X, Pan J (2006) An energy-efficient data-dissemination protocol in wireless sensor networks. In: Proceedings of the international symposium on a world of wireless mobile and multimedia networks, pp 10–22, Jun 2006
10. Luo H, Ye F, Cheng J, Lu S, Zhang L (2002) TTDD: two-tier data dissemination in large-scale wireless sensor networks. In: Proceedings of the 8th international ACM conference on mobile computing and networking, pp 148–159, Sep 2002

Border Detection of Skin Lesions on a Single System on Chip

Peyman Sabouri, Hamid GholamHosseini and John Collins

Abstract High speed image processing is becoming increasingly important in medical imaging. Using the state-of-the-art ZYNQ-7000 system on chip (SoC) has made it possible to design powerful vision systems running software on an ARM processor and accelerating it from hardware resources on a single chip. In this paper, we take the advantage of accelerating an embedded system design on a single SoC, which offers the required features for real-time processing of skin cancer images. Different edge detection approaches such as Sobel, Kirsch, Canny and LoG have been implemented on ZYNQ-7000 for border detection of skin lesions, which can be used in early diagnosis of melanoma. The results show that the extended 5 × 5 canny edge detection implemented on the proposed embedded platform has better performance in compare with other reported methods. The performance evaluation of this approach has shown good processing time of 60 fps for real time applications.

Keywords Border detection · Edge detection · ZYNQ-7000 · Medical imaging

Introduction

To date, skin cancers have been one of the most common form of cancers particularly in New Zealand [1, 2]. Skin cancers are divided into two main categories: melanoma and non-melanoma. Early diagnosis of malignant melanoma can

P. Sabouri (✉) · H. GholamHosseini · J. Collins
Department of Electrical and Electronics Engineering, Auckland University of Technology,
Private Bag 92006, Auckland 1142, New Zealand
e-mail: psabouri@aut.ac.nz

H. GholamHosseini
e-mail: hgholamh@aut.ac.nz

J. Collins
e-mail: jcollins@aut.ac.nz

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_53, © Springer Science+Business Media Dordrecht 2014

significantly decrease the morbidity, death and cost of the treatments [3]. Dermoscopy is a non-invasive method for diagnosis of melanoma and pigmented skin lesions. Although this device illustrates features of pigmented lesions, it is a challenging task for dermatologist to diagnose melanoma from other skin lesions [4]. Image processing techniques can be applied to skin images for better diagnosis of melanoma. For example, image features such as Asymmetry, Border irregularity, Color variation and regions with Diameter greater than 6 mm (ABCD rule) can be extracted using high performance image processing techniques [5].

In recent years, real-time vision systems have been used in a wide range of applications such as sophisticated medical imaging. Most computer-based vision applications have been developed based on a Graphics Processing Unit (GPU). However, recent developments in the field of powerful, low cost and energy-efficient embedded systems have led to the implementation of image/video applications into Digital Signal Processors (DSPs) and Field Programmable Gate Arrays (FPGAs). Nonetheless, each technique has its advantages and tradeoffs based on the nature of the algorithm, performance requirement, power consumption, cost, productivity, flexibility and design cycle time. In addition, the complexities of these techniques can create bottlenecks for developers. Moreover, working with both hardware and software design components as one system may lead to additional challenges. One of the recent products from Xilinx, ZYNQ-7000, can be considered as a promising solution to overcome these challenges. The ZYNQ-7000 consists of a dual-core ARM processor which is surrounded by a new 7-series Xilinx FPGA-based on 28 nm technology. The close integration between processing system (PS) and programmable logic provides the flexibility of ASIC technology and performance of FPGA technology on a single System on Chip (SoC). The ZYNQ-7000 SoC may be suitable for designing a handheld medical imaging system for skin cancer detection.

In this paper, different edge detection methods have been implemented in the programmable logic of ZYNQ using the VIVADO HLS tool (ver. 12.3). Border detection is widely used dermoscopic image analysis with the aim of detecting the boundaries between melanoma regions and the background.

ZYNG-7000 and System Implementation

The ZYNQ-7000 SoC consists of a dual-core ARM Processor and a Xilinx FPGA based on 28 nm technology [6]. The proposed system consists of the Zynq-7000 AP SoC ZC702 evaluation kit (ZC702). A CCD camera is connected to the board by a high-performance video I/O FPGA mezzanine card (FMC module) with HDMI input and output. Several video soft IP cores, such as defective pixel removal, de-mosaic, color correction matrices and video input/output, are used in order to design an image processing pipeline and border detection. In addition, in the processing system (PS) DDR3 memory is used to create input/output images via the ARM processor using AXI Video Direct Memory Access (VDMA).

Moreover, the VIVADO HLS tool (ver. 12.3) is used to implement the real-time border detection accelerator on the programmable logic (PL) of the ZYNQ AP SoC.

Unlike the traditional FPGA design flows, High Level Synthesis tools (HLS) can improve design quality, accelerate the design and verification tasks by applying optimal synthesis directives to transform C/C++/SystemC code specifications to register transfer level (RTL) implementation [7, 8]. Although C code is used for implementation of the algorithm, several directives such as DataFlow, memory/interface, and for loop optimisations [9] are applied to obtain the required performance and area utilisation in PL. In this article functional verification of the border detection algorithms was performed using C++ testbenches.

Border Detection

Edge detection is one the most important algorithms which is widely used in image analysis with the aim of detection of boundaries between objects and the background in gray level images. Although there are many approaches for border detection and object recognition in medical imaging, the calculation of the gradient level value for each pixel using neighboring masks and comparing with a given threshold, is a common approach. If the calculated value is greater than the threshold, the pixel is considered as an edge. The most important categories of edge detection algorithms are [10]:

- Gradient edge detectors (first derivative such as Sobel, Prewitt and Kirsch)
- Zero crossing (second derivative)
- Laplacian of Gaussian (LoG)
- Gaussian edge detectors (such as Canny)
- Colored edge detectors

In this paper, Sobel, Kirsch, LoG and Canny edge detectors were chosen for border detection of the skin cancer images. It was found that 3 × 3 kernels, which are normally used in edge detection methods, are highly localized. To achieve the best method for border detection, the extended edge operators (5 × 5 kernels) were used to include more neighborhood pixels [11]. These extended operators are described in more details in the following sections.

Sobel Operator

The Sobel operator consists of two vertical and horizontal kernel components, which can be obtained by 3 × 3 vertical and horizontal kernels (Eq. 1). The kernels are applied to each pixel in the image, and then a threshold is applied to create the final output pixel (Fig. 1b). To achieve better results, the Gx and Gy

Kernels are extended to 5 × 5 masks (Eq. 2) and Fig. 1c shows the results of applying this filter to the input image in Fig. 1a.

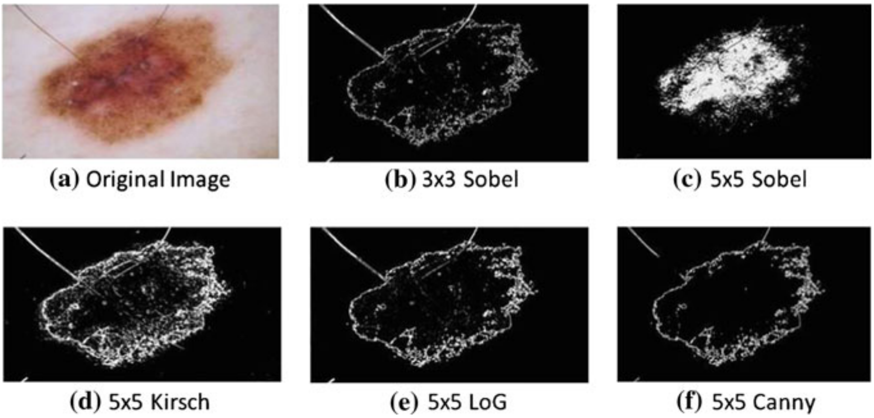
$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{1}$$

$$G_x = \begin{bmatrix} 2 & 2 & 4 & 2 & 2 \\ 1 & 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 & -1 \\ -2 & -2 & -4 & -2 & -2 \end{bmatrix}, \mathrm{G}_y = \begin{bmatrix} 2 & 1 & 0 & -1 & -2 \\ 2 & 1 & 0 & -1 & -2 \\ 4 & 2 & 0 & -2 & -4 \\ 2 & 1 & 0 & -1 & -2 \\ -2 & -2 & -4 & -2 & -2 \end{bmatrix} \tag{2}$$

Kirsch Operator

The Kirsch operator consists of eight basic convolution kernels components [8]. The vertical and horizontal edges can be obtained using the 5 × 5 Gx and Gy kernels in Eq. (3) and Fig. 1d shows the results of applying this filter to the input image.

$$G_x = \begin{bmatrix} 9 & 9 & 9 & 9 & 9 \\ 9 & 5 & 5 & 5 & 9 \\ -7 & -3 & 0 & -3 & -7 \\ -7 & -3 & -3 & -3 & -7 \\ -7 & -7 & -7 & -7 & -7 \end{bmatrix}, G_y = \begin{bmatrix} 9 & 9 & -7 & -7 & -7 \\ 9 & 5 & -3 & -3 & -7 \\ 9 & 5 & 0 & -3 & -7 \\ 9 & 5 & -3 & -3 & -7 \\ 9 & 9 & -7 & -7 & -7 \end{bmatrix} \tag{3}$$



(a) Original Image

(b) 3x3 Sobel

(c) 5x5 Sobel

(d) 5x5 Kirsch

(e) 5x5 LoG

(f) 5x5 Canny

Fig. 1 The resulting images from the border detection IP cores: **a** original image, **b** 3 × 3 Sobel operator, **c** 5 × 5 Sobel operator, **d** 5 × 5 Kirsch operator, **e** 5 × 5 LoG operator and **f** 5 × 5 Canny operator

LoG Operator

The second-order gradient LoG edge detector can be used by convoluting the LoG filter to the image [12]. The mask in Eq. (4) implements the LoG filter and Fig. 1e shows the results of applying this filter to the input image.

$$G_x = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & -2 & 16 & 0 & -1 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix} \tag{4}$$

Canny Operator

The canny edge detector is an optimal approach to find the edges in images using a Gaussian filter to reduce noise when the raw image is convolved with this filter [13]. The masks in Eq. (5) implement 5 × 5 canny edge detection and Fig. 1f shows the results of applying this filter to the input image.

$$G_x = \begin{bmatrix} 15 & 69 & 114 & 69 & 15 \\ 35 & 155 & 255 & 155 & 35 \\ 0 & 0 & 0 & 0 & 0 \\ -35 & -155 & -255 & -155 & -35 \\ 15 & 69 & -144 & -69 & 15 \end{bmatrix},$$

$$G_y = \begin{bmatrix} 15 & 35 & 0 & -35 & -15 \\ 69 & 155 & 0 & 155 & 69 \\ 114 & 255 & 0 & -255 & -114 \\ 69 & 155 & 0 & -155 & -69 \\ -15 & -35 & 0 & -35 & -15 \end{bmatrix}$$

Experimental Results

The original skin cancer image used for testing of border detection algorithms is shown in Fig. 1a. In addition, the results for Sobel, Kirsch, LoG and Canny edge detection implementations are illustrated in Figs. 1b–f. The results show that the extended 5 × 5 canny edge detection has better performance than other methods. After function verification of the algorithms, the IP core is generated using C-to-FPGA (VIVADO HLS) to be implemented into the programmable logic of ZYNQ-7000. Using this feature the development time was significantly reduced during the functional verification stage. The generated core segments border of skin lesions for skin cancer images. Therefore, it can be used for feature extraction step such as asymmetrical shape analysis. Furthermore, the created IP performed better border-detection in real-time (60 fps) using VIVADO HLS in compare with the traditional FPGA design flow.

Conclusions

The results show that the extended 5 × 5 canny edge detection implemented on the proposed embedded platform performs better than other reported methods. Consequently, it can segment border of the lesion before applying image analysis and feature extraction for melanoma detection. The generated IP for lesion border detection presented good processing time of 60 fps for real time applications using VIVADO HLS. Therefore, the proposed ZYNQ-7000 EPP can be considered not only as a high-performance and cost-effective solution for hardware/software implementation of image processing algorithm, but also a promising alternative for computer aided diagnosis (CAD) systems as a single SoC for portable solutions.

References

1. Diepgen T, Mahler V (2002) The epidemiology of skin cancer. Br J Dermatol 146:1–6
2. Kopf AW, Salopek TG, Slade J, Marghoob AA, Bart RS (1995) Techniques of cutaneous examination for the detection of skin cancer. Cancer 75:684–690
3. Goldsmith LA, Askin FB, Chang AE, Cohen C, Dutcher JP, Gilgor RS, Green S, Harris EL, Havas S, Robinson JK (1992) Diagnosis and treatment of early melanoma. JAMA: J Am Med Assoc 268:1314–1319
4. Argenziano G, Soyer HP, Chimenti S, Talamini R, Corona R, Sera F, Binder M, Cerroni L, De Rosa G, Ferrara G (2003) Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet. J Am Acad Dermatol 48:679
5. Xu L, Jackowski M, Goshtasby A, Roseman D, Bines S, Yu C, Dhawan A, Huntley A (1999) Segmentation of skin cancer images. Image Vis Comput 17:65–74
6. Zynq-7000 Extensible Processing Platform. http://www.xilinx.com/products/silicon-devices/soc/zynq-7000/index.htm
7. Chen J, Cong J, Yan M, Zou Y (2011) FPGA-accelerated 3D reconstruction using compressive sensing. In: Proceedings of the ACM/SIGDA international symposium on field programmable gate arrays, pp 163–166
8. Noguera J, Neuendorffer S, Van Haastregt S, Barba J, Vissers K, Dick C (2011) Implementation of sphere decoder for MIMO-OFDM on FPGAs using high-level synthesis tools. Analog Integr Circ Sig Process 69:119–129
9. Cong J, Zhang P, Zou Y (2011) Combined loop transformation and hierarchy allocation for data reuse optimization. In: IEEE/ACM international conference on computer-aided design (ICCAD), pp 185–192
10. Setayesh M, Mengjie Z, Johnston M (2012) Effects of static and dynamic topologies in particle swarm optimisation for edge detection in noisy images. In: IEEE congress on evolutionary computation (CEC), pp 1–8

11. Kekre DHB, Gharge MSM (2010) Image segmentation using extended edge operator for mammographic images. Int J Comput Sci Eng 2:1086–1091
12. Dhawan AP (2011) Medical image analysis. Wiley-IEEE Press
13. Canny J (1986) A computational approach to edge detection. IEEE transactions on PAMI-8 pattern analysis and machine intelligence, pp 679–698

A New Computer Based Differential Relay Framework for Power Transformer

Rachid Bouderbala and Hamid Bentarzi

Abstract A differential relay that is very sensitive relay operating even at its limits may be used for protecting a power transformer. However, this characteristic may lead to unnecessary tripping due to transient currents. In order to avoid this unnecessary tripping, estimated harmonics of these currents may be required which need great computation efforts. In this paper, a new frame work is proposed using PC interfaced with a data acquisition card AD622, which acquires real-time signals of the currents, process them numerically in the computer and outputs tripping signal to the circuit breaker. All algorithms of differential protection function and blocking techniques have been implemented using the Simulink/Matlab. To validate the present work, the performance of developed relay is tested by signals generated by Simulink/MATLAB simulator under different conditions. The test results show that this proposed scheme provides good discrimination between the transient currents and the internal fault currents.

Keywords Differential protection · Fault · Inrush current · Data acquisition card · Real time toolbox

Introduction

Differential protection is one of the most reliable and popular techniques in power equipment protection such as transformer and generator. It is so important for clearing rapidly all types of fault that occur in the power equipment. Differential

R. Bouderbala
Department of Instrumentation, Algerian Institute of Petroleum (IAP), Boumerdès, Algeria
e-mail: sisylab@yahoo.com

H. Bentarzi (✉)
Signals and Systemes Laboratory Lab, IGEE, UMBB University, Boumerdes, Algeria
e-mail: bentarzi_hamid@yahoo. com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_54, © Springer Science+Business Media Dordrecht 2014

protection operates and isolates a faulted part from the power system during an internal fault. However, the differential relay is blocked during the transient conditions due to the magnetizing inrush phenomenon which produces current only in the energized winding side otherwise the differential protection operates unnecessary. Hence, it is unable to distinguish between the transient current and the current that is due to an internal fault. The bias setting is one technique that has been used for overcoming this problem which is not effective. An increase in the protection setting to a value that would avoid operation may lead to make the protection insensitive. However, techniques of delaying, restraining or blocking of the differential element may be used to prevent mal-operation of the protection [1, 2].

In this work, a differential relay has been implemented using Simulink/MATLAB, which ensures security for external fault, inrush, and over-excitation conditions and provides dependability for internal faults, for protecting a three phase power transformer. This work combines harmonic restraint technique with a percentage blocking technique. The harmonic based dual slope characteristic differential relay is modeled using also Simulink. A power transformer Simulink model is used for evaluating the proposed differential relay performance under different operation conditions. The interfacing of this model with the differential relay based on PC is accomplished via acquisition card AD622 associated with real-time toolbox.

Differential Relay Principle Operation

Differential protection compares the currents that enter with the currents that leave a zone or element to be protected. If the net sum of the currents is zero, then the protected equipment is under normal condition. However, if the net sum is different from zero, the differential relay operates due to a fault existing within the equipment and isolates it from the power system.

Even differential protection is relatively simple to be implemented, but it has drawbacks. One of these drawbacks is its unnecessary tripping due to transformer magnetizing current. An inrush current is the surge of transient current that appears in a transformer due to inrush and over excitation conditions. The exciting voltage applied to the primary of the transformer forces the flux to build up to a maximum theoretical value of double the steady state flux plus reminisce as follows,

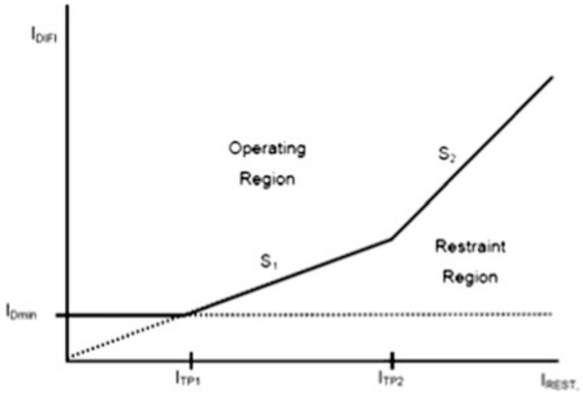
$$\phi_{MAX} = 2\phi_M + \phi_R \tag{1}$$

Therefore, the transformer is greatly saturated and draws more current which can be in excess of the full load rating of the transformer windings.

The slope, magnitude and duration of inrush current depend on several factors [3], such as size of a transformer, magnetic properties of the core material, and way a transformer is switched on (inner, outer winding, type of switchgear).

The most recent technique used for preventing false tripping is the use of second and fifth harmonic restraint. If these harmonic contents of the differential

Fig. 1 Typical restraint and operating characteristic of a differential relay



current exceed pre-defined percentages of the fundamental, inrush current is detected and the relay is blocked from tripping [1].

One study reported the minimum possible level of second harmonic in inrush current is about 17 % [4], a 15 % threshold is a good choice.

The developments in digital technology lead to the incorporation of microprocessors in the construction of relays and to the investigation of protection systems with the capacity to record signals during faults, monitor them and communicate with their peers.

Differential Relay Settings

Low impedance differential protection systems typically have 3–5 settings required to properly define the restraint characteristic of the relay (see Fig. 1). Where, I_{Dmin} = minimum differential current required to operate the relay,

I_{TP1} = turning point 1, I_{TP2} = turning point 2, S_1 = Slope 1 setting,

S_2 = Slope 2 setting, I_{RST} = Total current through the differential system,

I_{DIF} = For a given I_{TOT}, the Mini. Diff. current required to operate the relay.

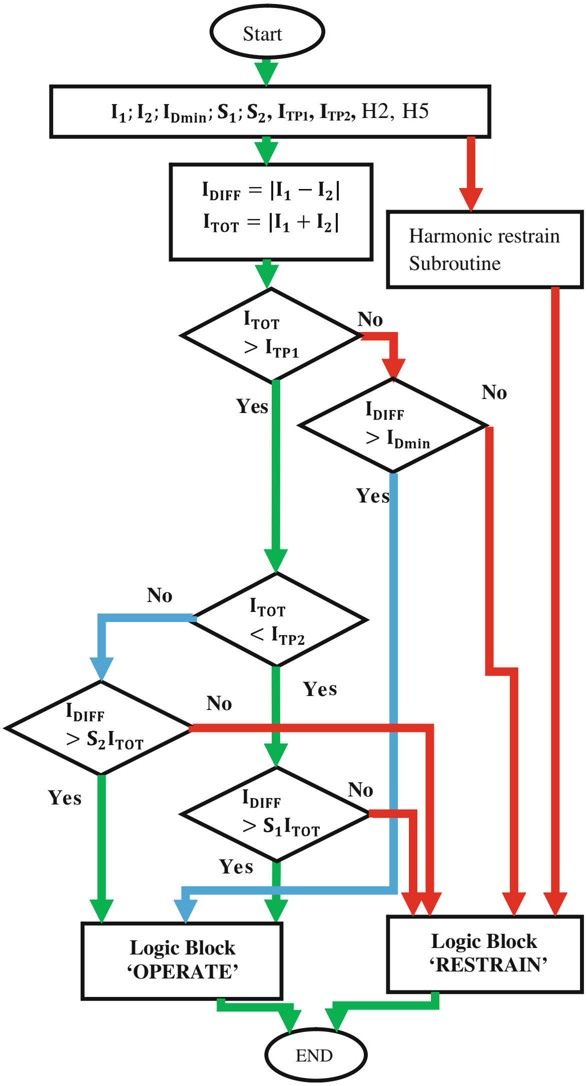
The settings to be considered are I_{Dmin}, I_{TP1}, I_{TP2}, S_1, and S_2. Besides, second harmonic (H2) and fifth harmonic (H5) may be used.

Differential Relay Implementation

Software Structure

Differential protection algorithm, which has been implemented using Simulink/Matlab, its flow chart is shown in Fig. 2. Besides, Graphical User Interface (GUI)

Fig. 2 Differential protection algorithm flowchart



has been developed using the same software tool, the user can select and set the desired parameters, and makes a test by running program and displaying the tripping signal (see Fig. 3).

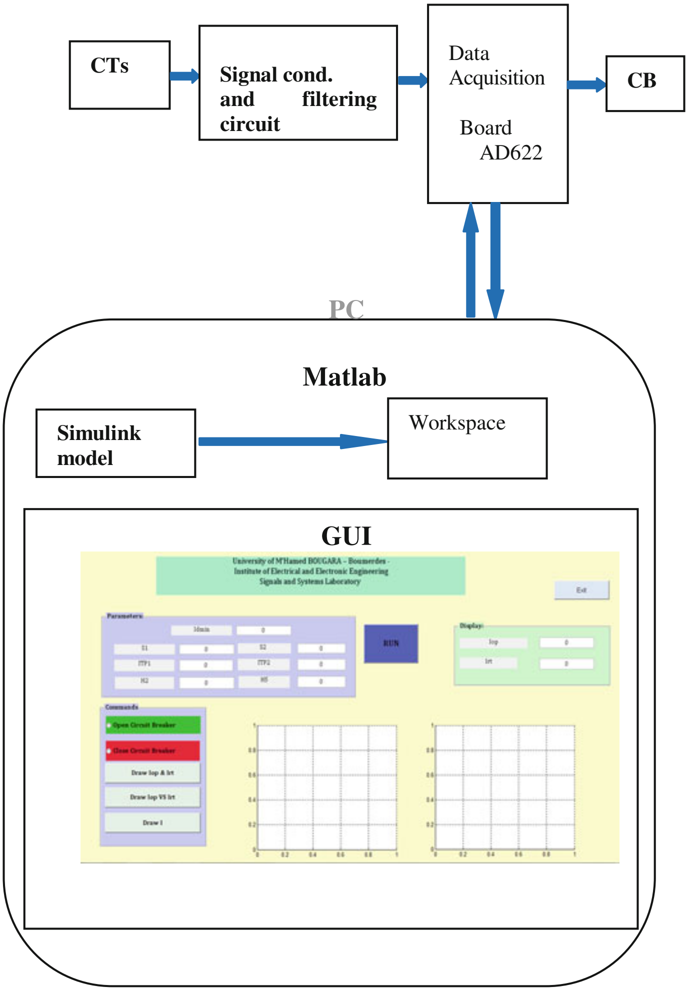


Fig. 3 General block diagram of PC Based differential relay

Hardware Architecture

In protection field, current transformers are used to measure the current and provide the measured quantity as analog voltage signal to the input of protective relay. Circuit breaker is used as actuator.

Thus, the differential protection hardware whose block diagram shown in Fig. 3 consists of:

- Signal transformation: current transformer (CT) transforms currents of the power system into voltage with low safe magnitudes.

- Signal Conditioning and filtering Circuit: the measured values of the power system fed from CTs in analog forms are passed through an anti-aliasing filter (low pass filter).
- Data acquisition boards: Sample and hold circuits and analog multiplexed are used to sample the three different signals of the three phase lines supplied by instrument transformers at the same time. The sampled signals are converted into digital form using ADC.
- PC: the digital signals are fed from data acquisition board AD 622 to the PC where they will be numerically processed as shown in Fig. 3 [5, 6].

Test Results and Discussion

The experimental setup system that may be used for testing the developed differential relay is composed of a 25 kV three-phase voltage source in series with a three-phase power transformer feeding a RL load. One circuit breaker is connected to the primary side of the transformer and takes its tripping signal from the differential relay so that it may open the circuit during faulty condition. The inputs to the relay are the primary side current of the transformer and the current of the secondary side. A simple system is used to check the validity of the proposed algorithm, and it is mainly composed of three phase star/star transformer with saturable core and initial fluxes. The main parameters of the used power transformer are given in Table 1. Two block sets of three phase faults may be applied, one is considered as internal fault and the other is the external fault.

Figure 5 shows the Simulink model of experimental test setup that may be used for generating the current signals under different conditions that may be injected via acquisition card to test the PC based differential relay (Fig. 4) [7].

In order to avoid the source transients, the transformer is connected at time t = 1 s. After running the program, the magnetizing current appears only in the primary side as shown in Fig. 6. This may produce a great difference in the input currents, but, the relay is blocked to operate as shown in Fig. 7.

Table 1 Simulated power transformer main parameters

Parameters	Values
Rated power	250 MVA
Voltage ratio	25,000/250 V
Rated frequency	60 Hz
Primary impedance/phase	0.002 + j0.08 pu
Secondary impedance/phase	0.002 + j0.08 pu
Magnetization resistance	500.2 pu
Connection	*Y/Y*

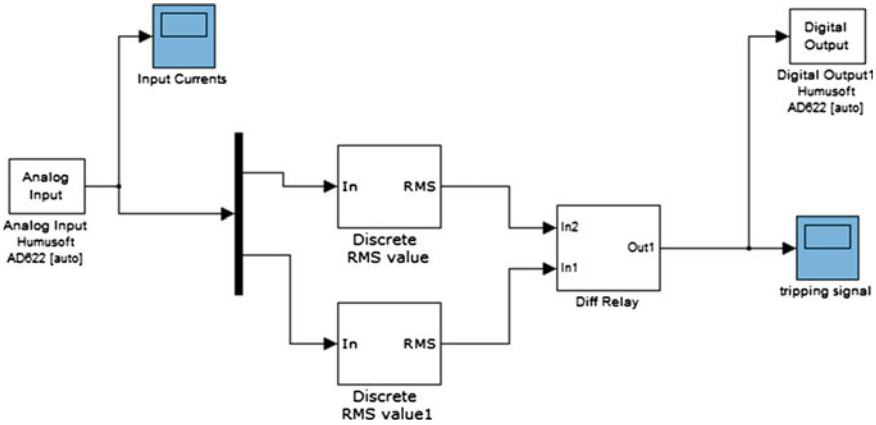


Fig. 4 Differential relay real time Simulink model

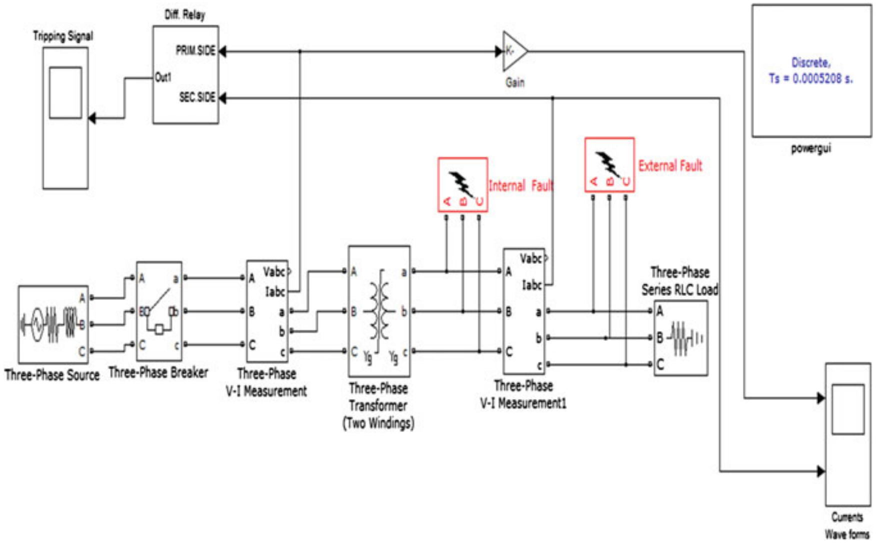


Fig. 5 Simulink model of the three phase transformer for experimental test

However, by applying an internal fault during an interval time $t_1 = 4\,s$ to $t_2 = 5\,s\,5\,s$, a large difference in the current amplitudes of both sides of the power transformer may be produced as shown in Fig. 6 and hence the differential relay operates in this situation as shown in Fig. 7. In the other hand, when an external fault is applied at the time $t_1 = 6\,s\,6\,s$, an increase of the current amplitudes on both sides of the power transformer can be remarked as shown in Fig. 6, in this case the differential relay is blocked to operate.

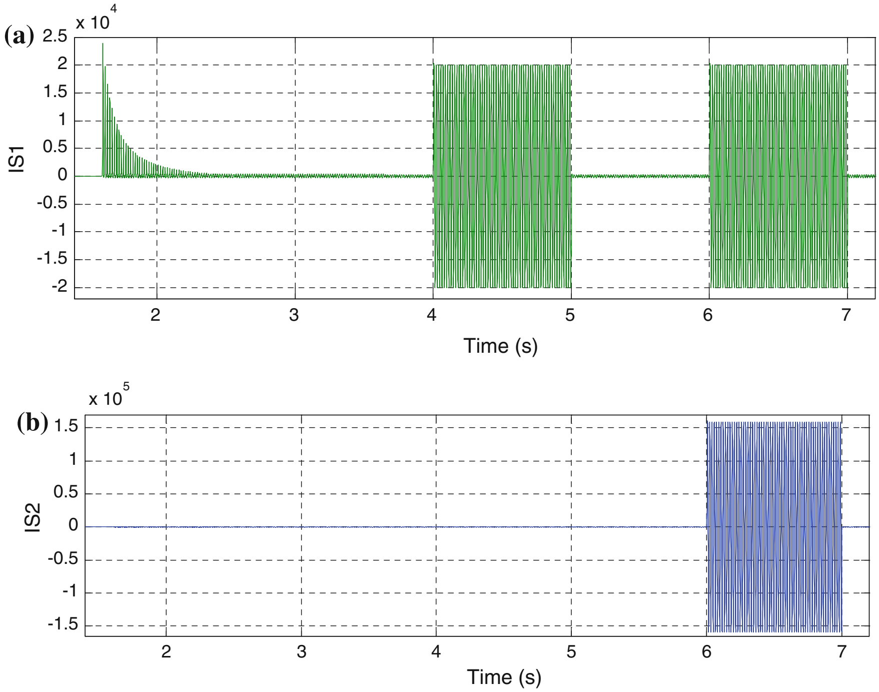


Fig. 6 The currents of (**a**) the primary side and (**b**) the secondary side by applying internal and external faults

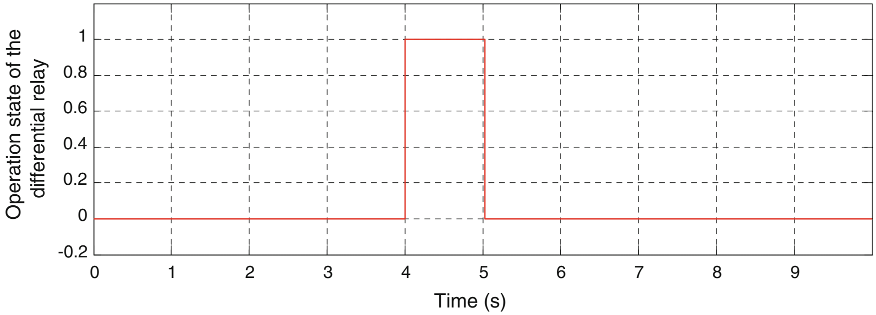


Fig. 7 The differential relay tripping signal

Conclusion

The PC-based differential relay prototype has been realized in this work. Where the algorithm has been implemented using the Simulink/Matlab and interfaced with the real world via the data acquisition card AD622.

After the test, the obtained results satisfy the principle operation of numerical differential relay and its characteristics using this new frame work.

Moreover, it can be concluded that this framework has the advantage; it can easily be implemented and tested using any protection function associated with recent the blocking function.

References

1. Bouderbala R, Bentarzi H, Ouadi A (2011) Digital differential relay reliability enhancement of power transformer. Int J Circ, Syst Sig Process, ISSN: 1998-446, vol 5(1), pp 263–270
2. Bentarzi H Some new aspects of protective relaying in modern electric power system (unpublished)
3. Karsai K, Kerenyi D, Kiss L (1987) Large power transformers. Elsevier, New York
4. Sonnemann WK, Wagner CL, Rockefeller GD (1958) Magnetizing Inrush Phenomena in transformer banks. AIEE Trans 77:884–892
5. IEEE power system relaying committee: understanding microprocessor-based technology applied relaying (2009) WG I-01 report, IEEE Organization
6. AD622 data acquisition card user's manual
7. Chafai M, Bentarzi H, Ouadi A, Zitouni A, Recioui A (2013) PC based testing system for protective relay. Accepted to be presented in 4th IEEE international conference on power engineering, May 13–17, Istanbul, Turkey
8. Blackburn J (1998) Lewis, protective relaying: principles and applications, 2nd edn. Marcel Dekker, New York, pp 275–280

A New Computer Based Quadrilateral Distance Relay Framework for Power Transmission Lines

Abderrahmane Ouadi and Hamid Bentarzi

Abstract In conventional transmission line protection, a three–zone stepped directional distance scheme is used to provide the primary as well as remote backup protection. The voltage and current phasors are needed by the distance relay for determining the impedance. In this paper, a new frame work is proposed using Data Acquisition Card, which acquires real-time signals of the voltages and currents, processes numerically them in PC and outputs tripping signal to the circuit breaker. Algorithm of quadrilateral distance protection has been implemented using the Simulink/Matlab. To validate the present work, the performance of developed relay is tested by signals generated by Simulink/MATLAB simulator under different conditions. The obtained simulation results are satisfactory.

Keywords Quadrilateral distance protection · Numerical relay · Data acquisition card · Power swing · Real time toolbox

Introduction

Power grid protection is the process of making the production, transmission, and distribution of electrical energy as safe as possible from the effects of equipment failures and events that place the power system at risk. When the faults occur in such power grid, protection systems are designed to isolate faulted part of the power grid, and leave the healthy parts of the system connected in order to ensure the continuity of the power supply. The operational security of the power system depends upon the successful performance of the thousands of relays that protect equipments and hence protect the whole system from cascading failures. Thus, the

A. Ouadi · H. Bentarzi (✉)
Signals and Systemes Laboratory Lab, IGEE, UMBB University, Boumerdes, Algeria
e-mail: sisylab@yahoo.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_55, © Springer Science+Business Media Dordrecht 2014

failure of a relay to operate as intended may jeopardize the stability of the entire power system and its equipment.

Disturbances may affect on transmission line relays such as overcurrent, directional overcurrent and distance relays which may respond to the variations of voltage and currents and their phase angle relationship. Under this condition, these relays operate even for stable power swings for which the system can recover [1]. In order to overcome the drawbacks of these traditional relaying systems, many techniques have been developed such as a blinder that blocks the relay to operate during the power swing [2] or by using a quadrilateral distance relay.

The mal-operation of this relay which is generally due to unnecessary tripping during power swing reduces the security of protection system and hence its reliability. The ideal characteristic for preventing this mal-operation is a quadrilateral characteristic but it can be implemented with difficulty.

The basic principle behind distance relay is calculating the line impedance seen by the relay at location using the fundamental frequency components of voltage and current signals. The Fourier transform based method can be used to estimate the fundamental components of the voltage and current. In this paper the basic theory behind the impedance measurement and the DFT algorithms are discussed. Besides, the simulations of the proposed scheme and its implementation using Acquisition card are presented.

Distance Relay Principle Operation

Protections based on distance relaying have been used in the power grid generally and in transmission lines particularly in order to detect the fault rapidly and disconnect the faulted part only. This maintains a reliable operation of the power grid by ensuring continuity of power supply [3–5].

The basic principle governing the operation of a distance relay is the ratio between the voltage V and the current I at the relaying point as shown in Fig. 1. The ratio (V/I) represents the measured impedance Z of the faulty line between the relay location and the point of fault occurrence. Then, the measured impedance is

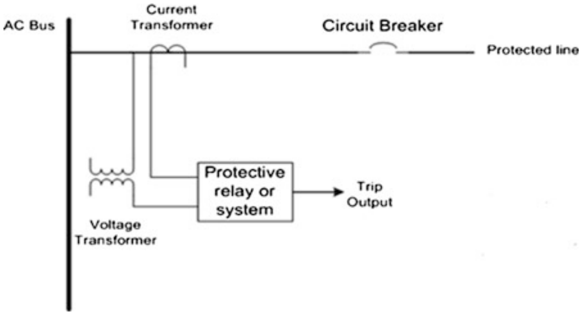


Fig. 1 A typical single line AC connection of a protective distance relay

compared to the set impedance, and if this Z is within the reach of the relay then the fault will be cleared. However, zone-three relay may be affected by the power swing because its impedance enters the characteristic of the relay and its time delay is longest.

A distance relay can be set to protect different zones of a transmission line either in forward direction (mho characteristic) or in both forward and backward direction (Offset mho characteristic). Figure 2a shows relay characteristics of an offset Mho and Fig. 2b illustrates Mho characteristics.

Faults are permanent compared to power swings that are more gradual events and disappear after a short period. This fact is used in the distance relay to distinguish between short circuit faults and stable or unstable power swings [6].

Traditional impedance-based characteristics for detecting power swings on a transmission system are shown in Fig. 3. Almost of these methods involve measuring apparent impedance and introducing a time delay between two measuring elements.

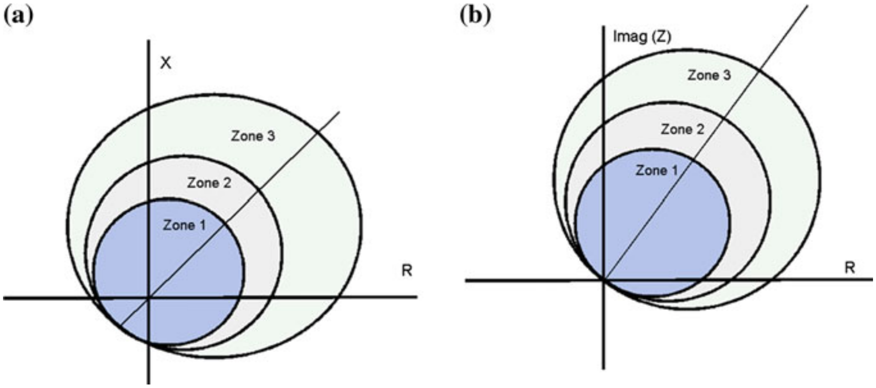


Fig. 2 **a** Offset Mho, and **b** Mho relays characteristics

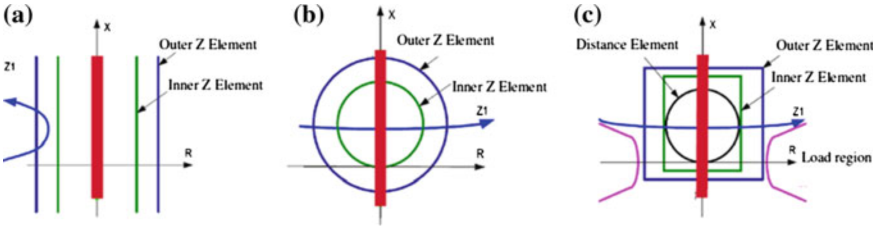


Fig. 3 Different power swing protection schemes. **a** Double blinder, power swing. **b** Offset mho, power swing characteristic. **c** Quadrilateral, power swing characteristic

However, when a quadrilateral characteristic is used, a blinder and a time delay are not required for avoiding the power swing as shown in Fig. 3c.

Impedance Measurement Algorithm

Generally, Fourier analysis can be used to calculate the fundamental frequency components of the voltage and current which in turn they will be used for estimating the impedance. In digital algorithms, sampled values are used instead of continuous variables. By taking N samples within time interval T which immediately precedes the sampling instant in question. The digital domain is given by [7],

$$x(k) = A_0 + \sum\nolimits_{\mathrm{n}=0}^{(N-1)/2} \mathrm{A_n} \cos\left(\frac{2\pi}{\mathrm{N}} nk\right) + \sum\nolimits_{\mathrm{n}=0}^{(N-1)/2} \mathrm{B_n} \sin\left(\frac{2\pi}{\mathrm{N}} nk\right) \tag{1}$$

where, n = 0,1,2,…….(N-1)/2 and k = 1,2,3,……..(N-1). The discrete from of the integrals are given by:

$$\begin{aligned} \mathrm{A_n} &= \frac{2}{\mathrm{N}} \sum\nolimits_{k=0}^{(N-1)} \mathrm{x(k)} \cos\left(\frac{2\pi}{\mathrm{N}} nk\right) \\ \mathrm{B_n} Z &= \frac{2}{\mathrm{N}} \sum\nolimits_{k=0}^{(N-1)} \mathrm{x(k)} \sin\left(\frac{2\pi}{\mathrm{N}} nk\right) \\ \mathrm{A_0} &= \frac{1}{\mathrm{N}} \sum\nolimits_{k=0}^{(N-1)} \mathrm{x(k)} \end{aligned} \tag{2}$$

where x(k) represents individual samples within the sampling window using the discrete Fourier series. The amplitude and the phase of the nth harmonic are given by,

$$|\mathrm{C_n}| = \sqrt{\left(A_n^2 + B_n^2\right)}, \tag{3}$$

$$\theta_n = \tan^{-1}\left(\frac{B_n}{A_n}\right) \tag{4}$$

The fundamental frequency components are given by $\mathrm{A_1}$ and $\mathrm{B_1}$. The discrete form of Eq. (1) is given by

$$\mathrm{X}(n) = \frac{1}{\mathrm{N}} \sum\nolimits_{k=0}^{(N-1)} \mathrm{x(k)} e^{-j\left(\frac{2\pi}{\mathrm{N}} nk\right)} \tag{5}$$

where k, n = 1,2, …N. Using Eq. (5), the fundamental components of the voltage and current can be obtained.

$$|\mathrm{C_1}| = \sqrt{\left(A_1^2 + B_1^2\right)}, \tag{6}$$

$$\theta_1 = \tan^{-1}\left(\frac{B_1}{A_1}\right) \tag{7}$$

Using Eqs. (6) and (7), the voltage C_V, θ_V and current C_I, θ_I fundamental phasor components can be obtained. The impedance and phase angle can be derived by:

$$|\bar{Z}| = \frac{c_V}{c_1} \tag{8}$$

$$\theta_z = \theta_V - \theta_1 \tag{9}$$

The above computation is carried out at each sampling interval. This will give an impedance value based on the voltage and current samples of the previous cycle.

Implementation is made through the use of Quadrilateral technique considered as the best suitable one that can be implemented only in the Numerical relays.

Simulink/MATLAB is used to model the distance relay components such as ADC and digital filters. Figure 4 shows the block diagram of the developed distance relay. The voltage and current data are derived using the power system simulator (power system block set for SIMULINK/MATLAB) as illustrated in Fig. 5. This system may include three phase source, transmission line (represented using π model), current transformers, voltage transformers and voltage and current measurement units. The voltage and current input signals are inserted in the distance relay block.

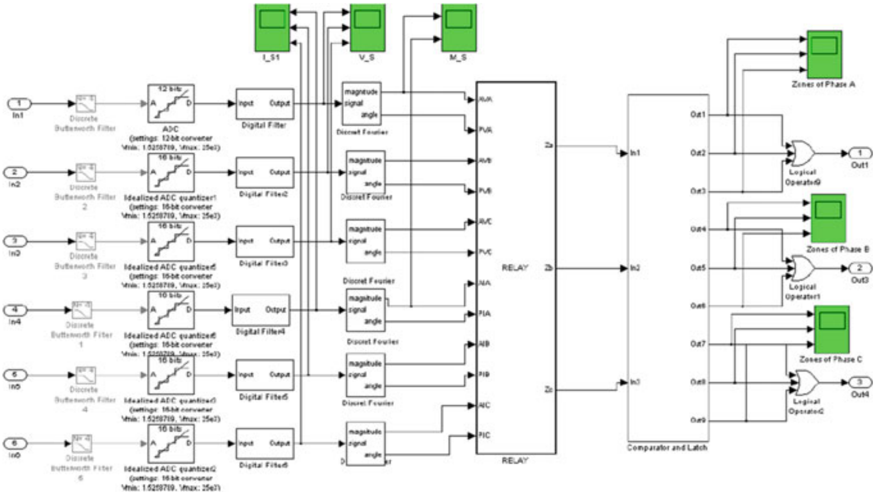


Fig. 4 General bock diagram of the designed quadrilateral relay

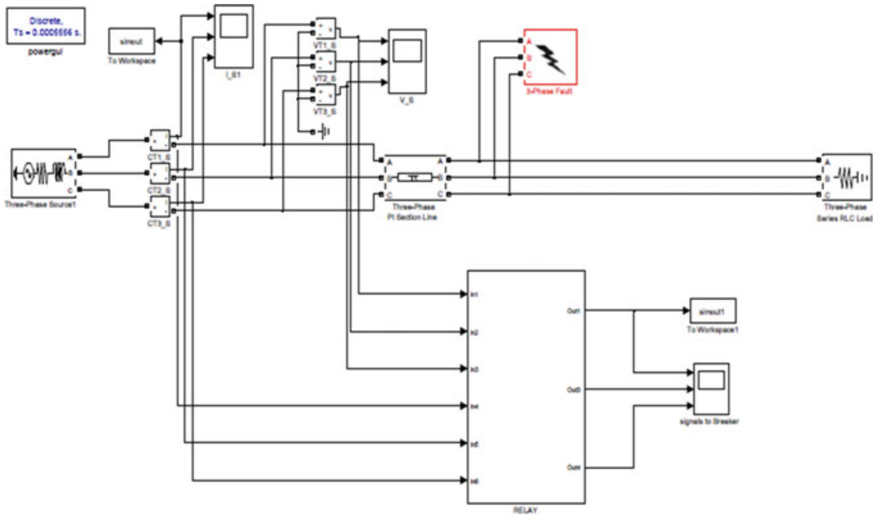


Fig. 5 Simulink model of the experimental test setup

Distance Relay Implementation

Software Structure

A quadrilateral characteristic is best suited for the numerical protection of HV transmission lines as it possesses an ideal distance relay characteristic. It excludes all the conditions for which the tripping is undesirable such as power swings, fault resistance and overloads. In the present work, is designed and implemented.

The quadrilateral distance relay algorithm that is implemented, its flowchart is illustrated in Fig. 6 [8, 9].

Hardware Architecture

In protection field, current transformers and potential transformer are used to measure the current and the voltage and provide the measured quantity as analog voltage signal to the input of protective relay. Circuit breaker is used as actuator.

Thus, the distance protection hardware whose block diagram shown in Fig. 9 consists of:

- Signal transformation: current transformer (CT) transforms currents of the power system into voltage with low safe magnitudes.
- Signal Conditioning and filtering Circuit: the measured values of the power system fed from CTs in analog forms are passed through an anti-aliasing filter (low pass filter).

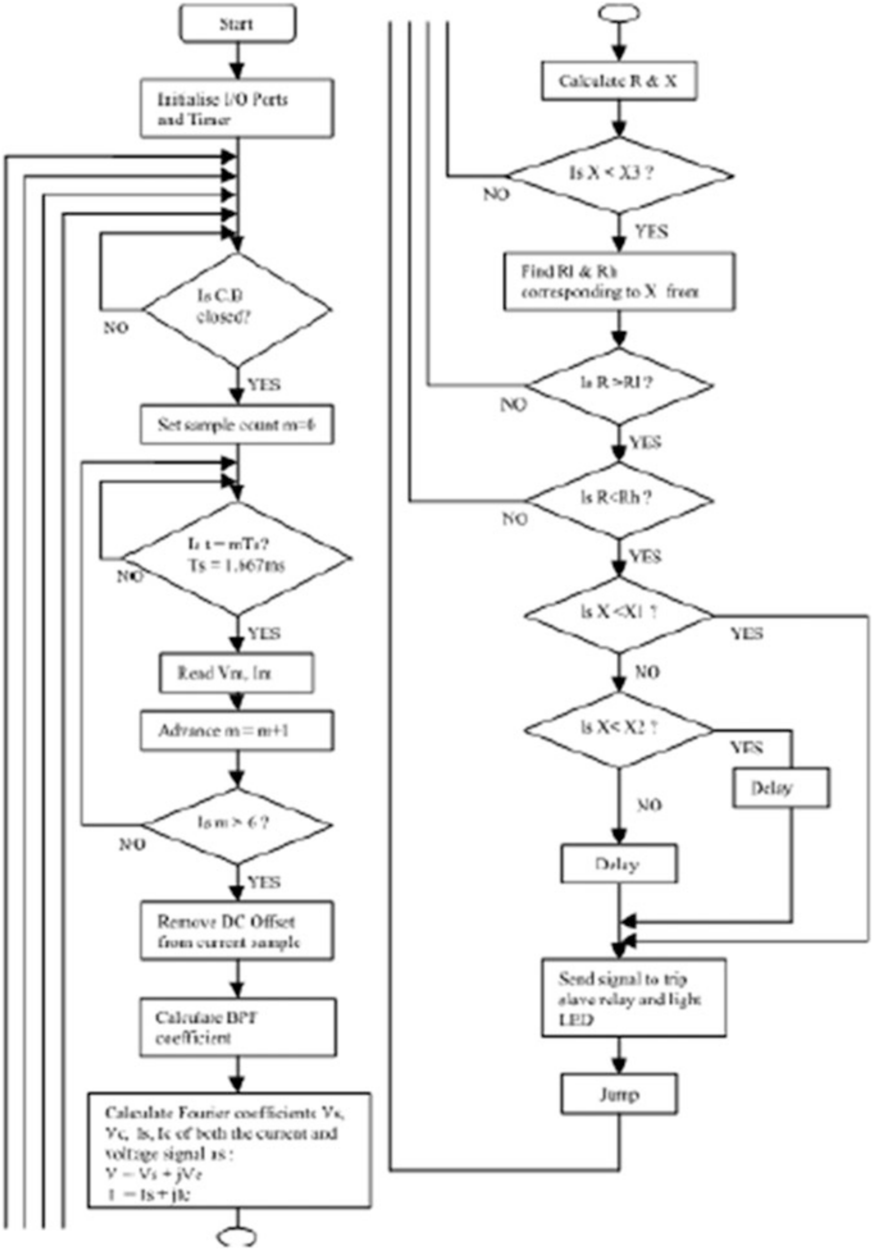
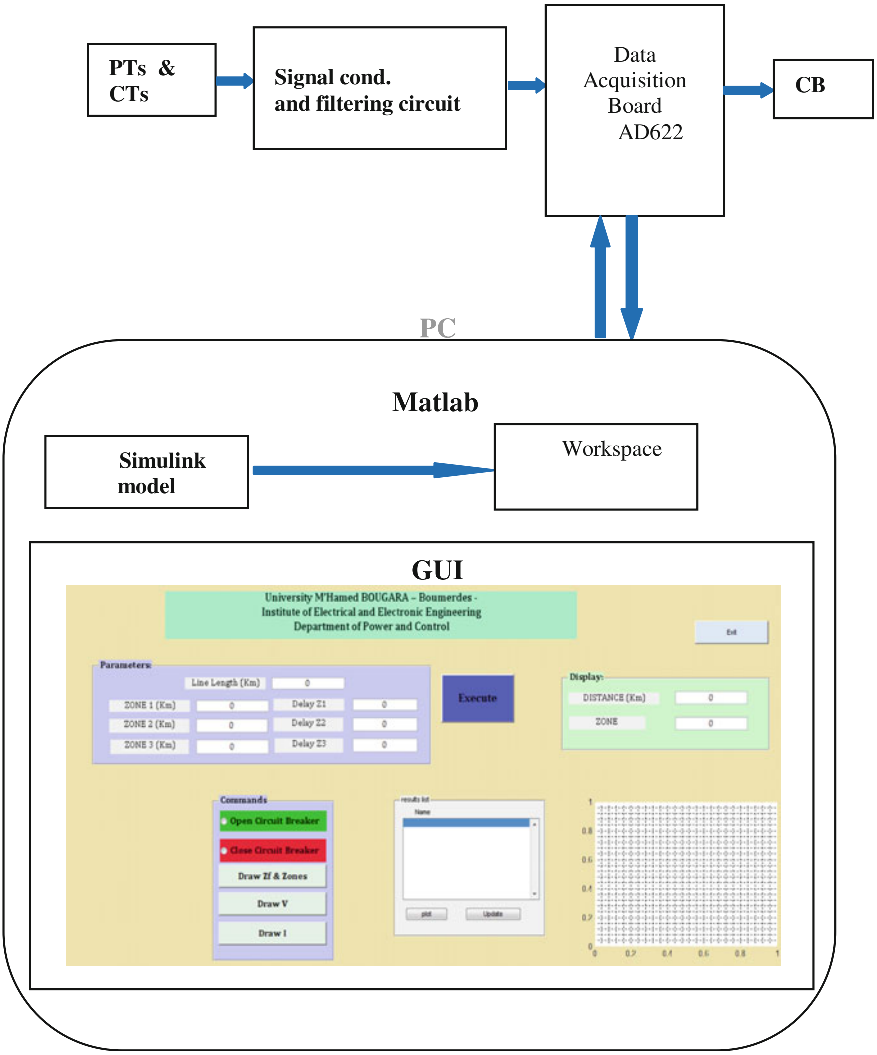

Start
Initialise I/O Ports and Timer
Is C.B closed?
NO
YES
Set sample count m=6
Is t = mTs?
Ts = 1.667ms
NO
YES
Read Vm, Im
Advance m = m+1
Is m > 6 ?
NO
YES
Remove DC Offset from current sample
Calculate BPF coefficient
Calculate Fourier coefficients Vs, Vc, Is, Ic of both the current and voltage signal as :
V = Vs + jVc
I = Is + jIc
Calculate R & X
Is X < X3 ?
NO
YES
Find Rl & Rh corresponding to X from
Is R >Rl ?
NO
YES
Is R<Rh ?
NO
YES
Is X <X1 ?
YES
NO
Is X< X2 ?
YES
Delay
NO
Delay
Send signal to trip slave relay and light LED
Jump

Fig. 6 Flowchart of quadrilateral distance relay algorithm

- Data acquisition boards: Sample and hold circuits and analog multiplexed are used to sample the three different signals of the three phase lines supplied by instrument transformers at the same time. The sampled signals are converted into digital form using ADC.
- PC: the digital signals are fed from data acquisition board to the PC where they will be numerically processed.

The developed distance protection relay has been implemented in PC associated with acquisition card AD 622 using Real time toolbox of Simulink as shown in Fig. 4 [6, 10].



University M'Hamed BOUGARA – Boumerdes -
Institute of Electrical and Electronic Engineering
Department of Power and Control

Parameters:
Line Length (Km) 0
ZONE 1 (Km) 0 Delay Z1 0
ZONE 2 (Km) 0 Delay Z2 0
ZONE 3 (Km) 0 Delay Z3 0

Display:
DISTANCE (Km) 0
ZONE 0

Fig. 7 General block diagram of PC based differential relay

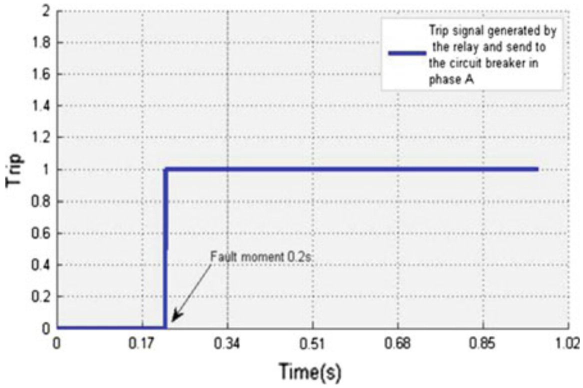


Fig. 8 Trip signal generated by the distance relay to the circuit breaker during the fault

Testing Results

The experimental setup system that may be used for testing the developed distance relay is shown in Figs. 5 and 7.

Figure 8 shows the trip signal generated by the relay when the fault occurs in phase A.

Conclusion

The PC-based distance relay prototype has been realized in this work. Where the quadrilateral characteristic algorithm has been implemented using the Simulink/ Matlab and interfaced with the real world via the data acquisition card AD622.

After the test, it can be noticed that the obtained results satisfy the principle operation of numerical quadrilateral distance relay and its characteristics using this new frame work.

Moreover, it can be concluded that this proposed framework has the advantage; it can easily be used implemented and tested even when complex protection algorithms such as quadrilateral characteristic are used.

References

1. Power system relaying committee of the IEEE power engineering society: power swing and out-of-step consideration on transmission lines. IEEE PSRC WG D6 (2005)
2. Bentarzi H, Ouadi A, Chafai M, Zitouni A (2011) Distance protective system performance enhancement using optimized digital filter. In: Proceedings of CSECS '11 the 10th WSEAS international conference on circuits, systems, electronics, control and signal processing, Montreux University, Switzerland, 29–31 Dec

3. Sachdev MS, Baribeau MA (1979) A new algorithm for digital impedance relay. IEEE Trans Power Apparatus Syst 98:2232–2240
4. Isaksson A (1988) Digital protective relaying through recursive least-squares identification. In: IEE proceedings, Part C: generation, transmission, and distribution, vol 135, pp 441–449
5. Phadke AG, Hlibka T, Ibrahim M (1976) A digital computer system for EHV substation: analysis and field tests. IEEE Trans Power Apparatus Syst, vol PAS-95, pp 291–301
6. Internal PNR report (2012)
7. Ouadi A, Bentarzi H, Maun JC (2011) Phasor measurement unit reliability enhancement using real-time digital filter. Int J Circ, Syst, Sig Process 5(1):1–8
8. Shrivastava K, Vishwakarma DN (2007) Microcontroller-based numerical quadrilateral relay for the transmission line protection, electric power components and systems, pp 1301–1315
9. Ziegler G (2006) Numerical distance protection, 2nd edn. Public is Corporate Publishing, Siemens, Erlangen
10. Ouadi A, Bentarzi H, Maun JC (2009) A new computer based phasor measurement unit framework (published conference proceedings style, IEEE explorer). In: Proceedings 6th international conference SSD’09, Djerba, Tunisia, pp 1–6
11. Gadgil K (2010) A numerical protection relay solution, texas instruments application report, SLAA466—Sept

A Study for a Low-Power Way Predictor for Embedded Data Caches

Yul Chu

Abstract This paper introduces an enhanced predictor to reduce power consumption for a way-prediction cache used for embedded systems. The proposed predictor shows better prediction accuracy and lower power consumption compared to any conventional data caches. In addition, two representative cache replacement policies, *LRU* (Least recently Used) and *random*, are examined for low-power data caches; simulation results show that *random* reduce power consumption more than *LRU* for highly-associative way-prediction caches. SimpleScalar and Cacti simulators are used for these simulations with SPEC benchmark programs.

Introduction

During the last decade, low-power consumption has been a critical issue for embedded systems, especially for hand-held devices. For most embedded systems, microprocessor and cache memory might consume most of the power in the system. According to [1–3], modern cache memories occupy more than 60 % of the microprocessors' die area and cause more than 40 % of total power dissipation. To reduce power dissipation, it is necessary to design low-power cache memory by lessening access times to charge bit-lines of cache memory [3, 4]. Typically, on-chip caches in mobile devices are highly-associative, which is greater than 16-way. Therefore, a cache miss results in a costly access to the memory. Highly-associative caches in [5–7] are specifically designed for low-power embedded systems to provide better performance by reducing the conflict misses, due to imperfect allocation of entries in a cache. However, they significantly increase power

Y. Chu (✉)
Department of Electrical Engineering, University of Texas Pan American, 1201 W. University Dr, Edinburg, TX 78539, USA
e-mail: chuy@utpa.edu

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_56, © Springer Science+Business Media Dordrecht 2014

consumption because of simultaneous accesses to all the banks in parallel, i.e., the n-way cache has n banks to access simultaneously. This paper aims to reduce the power consumption by accurately predicting only one bank out of n banks (i.e., n-way) and by restricting access to the predicted bank only. This not only saves the power consumption but also reduces the latency since the cache is accessed as a direct-mapped cache.

The rest of the paper is organized as follows: section Related Works explains some related works done in this area; section Dual-Access Way-Prediction Cache presents the proposed way predictor; section Simulations and Performance Metrics deals with simulations and performance metrics; section Experimental Results provides the experimental results and discussion; and finally the conclusions are provided in section Conclusions.

Related Works

Much research has been proposed how to reduce the power consumption for highly associated cache memories. One popular approach is to use a phased cache, which divided into two caches, tags and data caches [8]; in the phased cache, the tag bits for all tag banks are powered, active to be accessed, and compared with a referenced memory address; on a hit from a bank, the bank is powered and accessed during the next cycle. Even though the phased cache can reduce power consumption, it has a fixed latency to all cache accesses. Another approach is to use a way-prediction cache [9]; the way predictor logic is added on top of a conventional cache scheme, such as a 2-way, 4-way, etc. This logic predicts a way, which should be accessed; on a miss, all the rest of the banks are accessed. *Therefore, the way-prediction can reduce power consumption more effectively than the phased cache* [9]. However, the prediction logic might increase power consumption slightly compared to a conventional cache memory. The way predictor logic implemented in [9] is a simple ‘most recently accessed logic’, having the same number of index table entries as the banks in a cache. This results in better power savings than a conventional cache. The power savings from such a design is entirely dependent on a prediction-hit rate of a way predictor. Such a scheme also has the advantage of reducing latency [10].

Dual-Access Way-Prediction Cache

This section proposes an enhanced way-prediction cache called Dual-Access Way-Prediction (DAWP) to select one of banks in a highly-associative data cache to reduce power consumption.

Figure 1 shows a conventional way-prediction (WP) cache: The way-prediction cache speculatively chooses one way from the index table, and then accesses the predicted bank as shown in Fig. 1. If the prediction is a hit in the bank, the cache access will be completed. Otherwise, the other remaining banks will be accessed using a normal cache access process, which is a case of prediction miss.

Figure 2 shows the proposed DAWP cache, which has two major parts. The first part is a direct-mapped index table, which stores previously accessed bank (way) address. The referenced output (as a predicted way) in the table goes to a multiplexer. The second part of this scheme consists of a global history register and a small fully-associative cache. The register stores a group of cache ways [e.g., 0001 (bank 1) and 0010 (bank 2)], which accessed recently. The small fully-associative cache has a fixed number of lines and each line holds a selected global history and a predicted way for the given global history. The output of this also goes to the multiplexer. The global history and fully-associative cache are rarely enabled, *valid* in Fig. 2; the *valid* will be '1' only when there are constant trashing of data in the index table; after then, if the global history is matched, the predicted way is accessed to see if the data is present in it. If the data is not present, the rest of the banks will be checked. On a cache miss, the data is retrieved from the memory and updated to the cache, and the way information is updated in the index table and/or the fully-associative cache. On a way-prediction miss but a cache hit, i.e., if the data is found in one of unpredicted banks, the correct bank is updated to the index table.

This scheme works based on the locality of data. A sequence of data, spatial locality, can be located and accessed in a cache line of a bank (way). As we discussed, the way information is updated via memory address (index) and global history. To reduce power consumption, it is required to reduce conflicts in the table and cache; highly biased accesses can be filtered out using a global history; in addition, the hit time can be reduced while accessing only one way instead of accessing n-way caches [9] for a prediction hit.

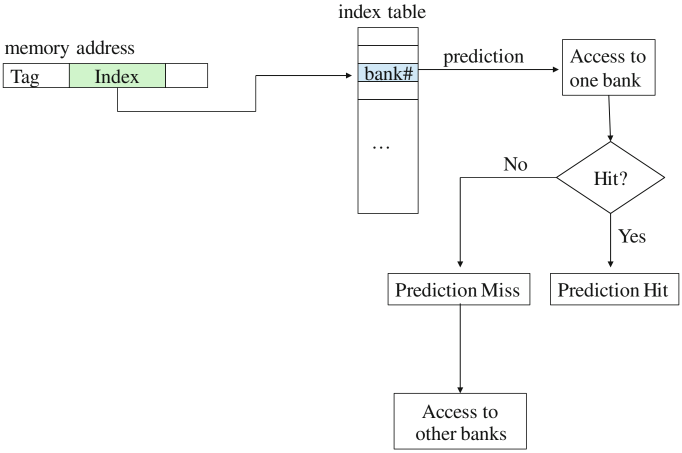


Fig. 1 Conventional way-prediction cache (WP)

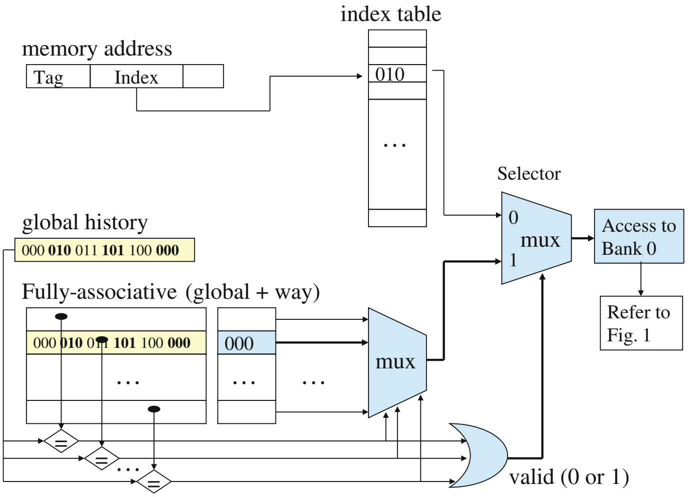


Fig. 2 Dual-access way-prediction (DAWP) cache

Simulations and Performance Metrics

The simulations are done on 16 KB, 32 KB and 64 KB cache sizes with configurations of direct mapped, 2-way, 4-way, 8-way, 16-way, and 32-way set-associative cache memories. SimpleScalar is used for these simulations with 9 SPEC2000 programs (art, ammp, equake, mesa, mcf, vpr, vortex, gcc, and gzip) [11].

For the simulation, the size of the index table is optimized as 512 bytes since we found that it is a good compromise between prediction-hit rate and power consumption through our simulation results.

To compare the miss rates, we use one more metrics, latency. In order to calculate this, the cache accesses are grouped into three categories: 1) the first one is the case of a correct prediction. For this case, the latency and power dissipation is based on accesses to the index table and the predicted bank like a direct mapped cache; 2) the second one is the case of a wrong prediction but a cache hit. In this case, the predictor predicts a wrong way but the data is found in a different bank (way). So after the data access, the correct way needs to be updated to the index table. The latency and power dissipation is due to an access to the table, an access to one way of the cache, accesses to other banks, and an access to update the index table; and 3) the last case is a cache miss, which results in the above accesses and a costly access to memory to fetch data.

Cache latencies and power dissipations are calculated using Cacti [10]. The memory latencies and power dissipations are estimated from [12]. All the hit rates, prediction misses and cache miss rates are obtained from Simplescalar simulator [11].

Experimental Results

The way-prediction (WP) cache described in [8] has an index table with the same number of entries as the number of sets in a cache. In Fig. 3, index size of 1X represents the WP cache and other sizes (2X to 16X) are based on the DAWP cache.

As the associativity of the cache increases such as 16-way or 32-way, the number of sets in a cache reduces drastically; then, the reduced entries of the index table degrade prediction rate because of conflicts in the table. For example, index table of 16-way is 1/16 of the direct-map (one-way) index table size. For this reason, the number of entries in the index table is scaled up by multiplication factors of 1X, 2X, 4X, 8X and 16X. Since each entry of the index table is ranged only from 1 bit to 5 bits, the scaling up of the index table does not affect total power consumption and latency significantly. Our experimental result shows prediction miss rates according to variable index sizes for a benchmark program (swim); the prediction miss rates for a 32 KB and 16-way cache are decreased as the index table is scaled up. However, the prediction rate is saturated to 5 % after 16X. Therefore, we select a scaling factor of 16X as an optimal value for DAWP cache to run all the benchmarks with the cache size of 512B.

The next optimizing parameter for our simulation is to determine the replacement policy for data caches, *LRU* (least recently used) and *random* replacement policies. *LRU* replacement policy conventionally provides low miss rates at the cost of having more complex hardware. For WP and DAWP caches, *random* replacement policy provides better performance than *LRU* according to our experimental results.

Figure 4 compares the total power dissipation using arbitrary unit (a relative unit of measurement) for 16- and 32-way cache memories between *LRU* and *random* with SPEC2000 benchmark programs. The figure shows *random* can reduce more power consumption than *LRU*; therefore, *random* will be used for the simulations for WP and DAWP caches in this paper.

According to [8] and Fig. 3, the DAWP cache gives better results than a conventional cache, which has the same number of banks (e.g., n banks for n-way

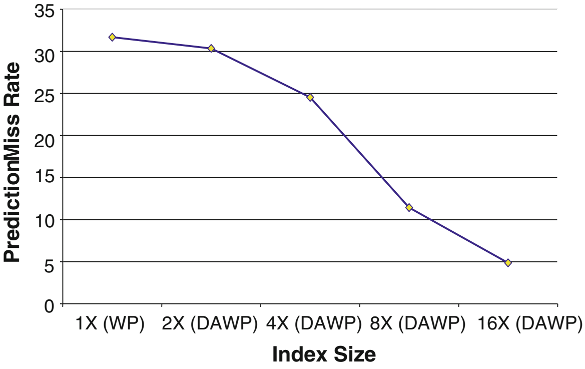
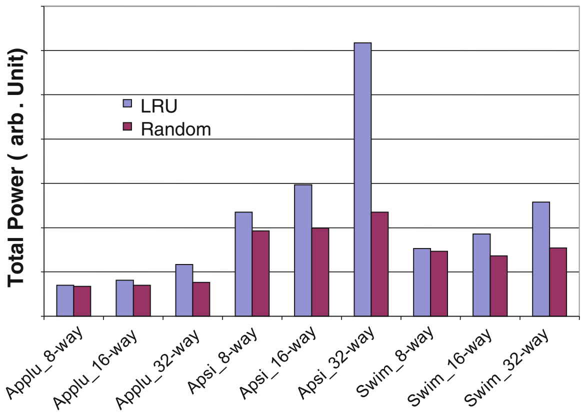


Fig. 3 Prediction miss rates vs. index table size

Fig. 4 LRU vs. random power dissipation

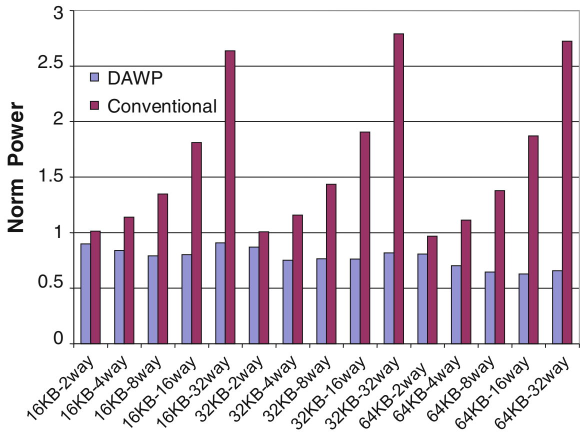


set-associative and n-way way-predictive cache) for SPEC benchmark programs. Therefore, we implement DAWP cache instead of WP cache in this paper.

In all the benchmarks, the DAWP cache has equal or lesser latency than conventional caches. In addition, power consumption savings for a highly-associative DAWP cache is much lower than any conventional caches, such as 16-way or 32-way.

Figure 5 shows total power consumption for SPEC2000 benchmark programs. The power consumptions are normalized to that of a direct-mapped cache in the figures. That means the total power dissipation of the nine SPEC2000 benchmark programs normalized to a direct mapped cache; power consumption for a direct-mapped cache is '1'. From our experimental results, we found that the DAWP caches offer lower power consumption than the highly-associative conventional caches because the DAWP cache is accessed as a direct-mapped cache for prediction hits; this leads to both low latency and power consumption especially when the prediction miss rate is low.

Fig. 5 Harmonic mean of power for SPEC2000



Conclusions

An enhanced way-prediction cache, DAWP, is implemented and simulated using SPEC 2000 benchmark programs. The WP and DAWP caches are optimized for the best prediction rates by scaling the index table and choosing the *random* replacement policy. The DAWP cache has equal or lesser latency than a conventional cache for all the SPEC benchmark programs. This cache also reduces power dissipation compared to any conventional caches, making it ideal for embedded systems. Our simulation results also show that a 16-way DAWP cache is found to have the best compromise for low-power consumption.

References

1. Montanaro J, Witek RT, Anne K, Black AJ, Cooper EM, Dobberpuhl DW, Donahue PM, Eno J, Hoeppner W, Kruckemyer D, Lee TH, Lin PCM, Madden L, Murray D, Pearce MH, Santhanam S, Snyder KJ, Stephany R, Thierauf SC (1996) A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor. IEEE J Solid-State Circ 31(11):1703–1714
2. Flynn MJ, Hung P (2005) Microprocessor design issues: thoughts on the road ahead. IEEE Micro 25(3):16–31
3. Zhang C (2006) A low power highly associative cache for embedded systems. IEEE international conference on computer design (ICCD), San Jose, CA, pp 31–36, Dec 2006
4. Powell M, Agarwal A, Vijaykumar TN, Falsafi B, Roy K (2001) Reducing set-associative cache energy via way-prediction and selective direct-mapping. IEEE 34th international symposium on microarchitecture (MICRO), Austin, TX, pp 54–65
5. Furber S et al (1989) ARM3—32b RISC processor with 4 kbyte on-chip cache. In: Musgrave G, Lauther U (eds) Proceedings IFIP TC 10/WG 10.5 international conference on VLSI, pp 35–44. Elsevier (North Holland)
6. Santhanam S et al (1998) A low-cost, 300-MHz, RISC CPU with attached media processor. IEEE JSSC 33(11):1829–1838
7. Intel XScale Technology. http://www.intel.com/design/intelxscale. Accessed on April 2013
8. Inoue K, Ishihara T, Murakami K (1999) Way-predicting set-associative cache for high performance and low energy consumption. In: Proceedings of international symposium on low power electronics and design, pp 273–275
9. Batson B, Vijaykumar TN (2001) Reactive-associative caches. Parallel architectures and compilation techniques. In: Proceedings of international conference on 8–12 Sept 2001 pp 49–60
10. Shivakumar P, Jouppi N (2001) CACTI 3.0: an integrated cache timing, power, and area model, Compaq, Palo Alto, CA, WRL Res. report 2001/2
11. Burger DC, Austin TM (1997) The SimpleScalar tool set, version 2.0, computer architecture news, vol 25(3), pp 13–25, June 1997
12. Contreras G, Martonosi M (2005) Power prediction for Intel Xscale processors using performance monitoring unit events, ISPLED ‘05, San Diego, CA, August 2005

A Density Control Scheme Based on Disjoint Wakeup Scheduling in Wireless Sensor Network

EunHwa Kim

Abstract Wireless sensor networks consist of many nodes with sensing, computation, and wireless communications capabilities. Due to difficulty of recharging battery, energy efficiency is essential problem in wireless sensor network. We propose a density control scheme based on disjoint wakeup scheduling which can provide a full connectivity to sink node with a minimum set of active nodes in a highly dense network to prolong the network lifetime.

Keywords Wireless sensor network · Scheduling · Connectivity · Energy efficiency · Density control

Introduction

Advanced sensor technologies and wireless communications have enabled the development of wireless sensor networks which can be used for various applications such as earth-quake, forest fire, the battlefield surveillance, machine failure diagnosis, biological detection, home security, smart spaces, inventory tracking, etc. [1, 2]. A wireless sensor network consists of tiny sensing device which has capability of detecting some phenomena.

In a large-scale sensor network, we need to deploy numerous sensor nodes densely in the sensing field and maintain them in proper way. One of the most important issues in such high-density sensor networks is density control [3, 4].

It can save energy by turning off redundant nodes and it can prolong the system lifetime by replacing the failed nodes with some sleeping nodes. Density control is

E. Kim (✉)
YongIn University, Yongin, South Korea
e-mail: ehkimanna@yongin.ac.kr

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_57, © Springer Science+Business Media Dordrecht 2014

very important for energy efficiency in topology control with wireless sensor networks.

In GAF [5], total network is divided a smaller virtual grid cell enough to communicate with its neighbors and one sensor node is selected as active node in a virtual grid cell. OGDC [6] prove that if communication radius is greater than twice of sensing radius, the connectivity is guaranteed by full coverage problem. In ASCENT [7], sensor node participate the network topology with parameter as the number of neighbors and packet loss ratio. Joint Scheduling method [8] awakes sensor nodes with random time slot which provide statistical coverage ratio and it awakes the other extra time slot that belongs to its downstream neighbor for connectivity to the sink node.

In this paper, we propose a density control scheme based on disjoint wakeup scheduling to improve energy efficiency while satisfying the requirements for sensing coverage and network connectivity. Our scheme works in a distributed manner at each sensor node and does not require the location information.

The rest of this paper is organized as follows. In paper A Vocabulary Learning Game Using a Serious-Game Approach, we propose our scheme, and evaluate it through simulation in paper An Improved Method for Measurement of Gross National Happiness Using Social Network Services. Finally, we conclude the paper in Advanced Comb Filtering for Robust Speech Recognition.

Scheduling Algorithm

In this paper, we propose a density control scheme based on disjoint scheduling algorithm which can provide a full connectivity to sink node with a minimum set of active nodes to prolong the system lifetime. It is assumed that every sensor node initially knows the hop count to the sink node. The best way to deliver data to sink node is forwarding its data to one of its upstream neighbor nodes. Upstream neighbors of a node are defined as nodes with smaller hop count by one than itself within its transmission range. A sensor node has only to communicate with upstream neighbor nodes to deliver data to sink node or receive a query from to sink node. In our algorithm, only one of nodes with the same hop count within transmission range (hereinafter, we call them ‘peer neighbors’) is allowed to wake up. To elect such an active node, each sensor node initially turns on its radio and keeps listening, and performs a random back-off. Once the back-off process is finished, it sends a *GREETING_MSG* to declare that it wants to become an active node and thus prevents other peer neighbors from activating. When a node hears a *GREETING_MSG* from its peer neighbors, it immediately stops the back-off process and goes to sleep. In this way, we can minimize the number of active nodes in the network field. However, it cannot guarantee path connectivity from a node to the sink node if there is no active upstream neighbor node around it.

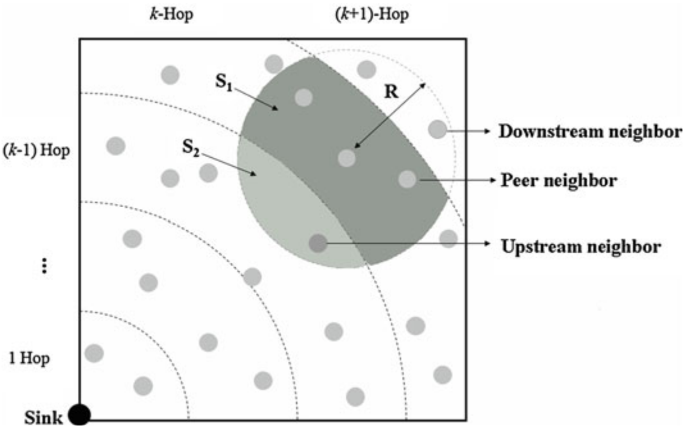


Fig. 1 Wireless sensor network

As shown in Fig. 1, the average number of active upstream neighbor nodes in our disjoint scheduling method is given by

$$N_{active\_up} = \frac{S_2}{S_1} \quad \text{where} \quad \begin{aligned} S_1 &= 2\int_{-\frac{1}{2}R}^{\frac{1}{2}R} \left(\sqrt{R^2 - y^2}\right) dy * \frac{A(N_{adj})}{\pi R^2} \\ S_2 &= 2\int_{\frac{1}{2}R}^{R} \left(\sqrt{R^2 - y^2}\right) dy \end{aligned}$$

where *A(k)* means area of a regular k-polygon, N_{adj} means the average number of adjacent active nodes, and R denotes communication radius. From [9], N_{adj} is given by

$$N_{adj} = \frac{\pi}{\theta} \approx 3.3 \quad where \quad \cos\theta = \frac{{}^{7}\!/_{6}R}{R} = \frac{7}{6}$$

So, we have $N_{active\_up} = 0.64$ approximately. This indicates that there is a possibility of 0.36 that an active node can not find another active neighbor in the upstream direction toward the sink. Actually, the number of upstream neighbors required to additionally become active is not so many that it is not much burden to the network. In our algorithm, if an active node does not hear a *GREETING_MSG* from upstream neighbor nodes, it selects an upstream neighbor node at random, and wakes up it by sending a *SELECT_MSG*.

Performance Evaluation

To evaluate our algorithm, we carried out experiments to measure coverage and connectivity of the network. To validate our scheme, we implemented a simulator in C/Java and compare it with GAF method and Joint Scheduling method [8].

Fig. 2 Coverage ratio

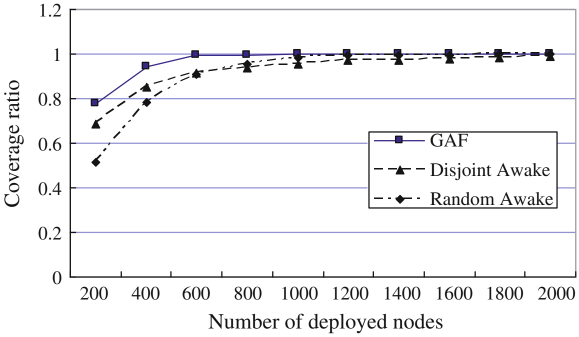


Fig. 3 Number of active nodes

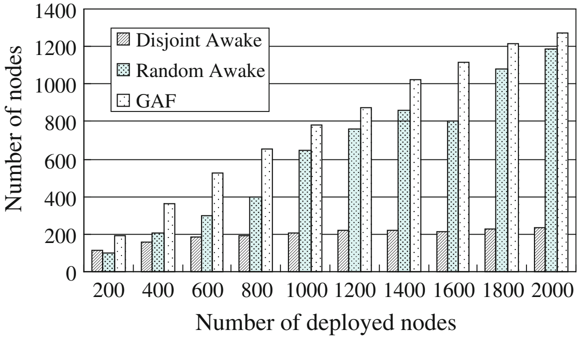
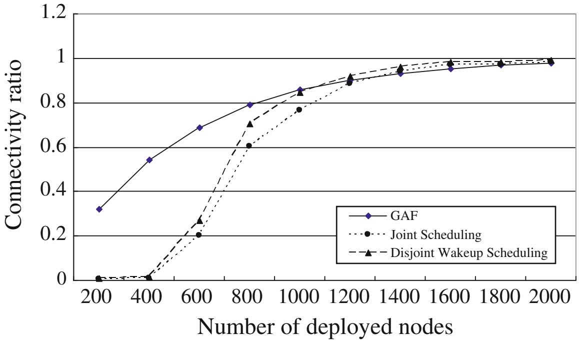


Fig. 4 Path connectivity ratio



Sensor nodes are deployed uniformly in a random fashion in a 200 * 200 square region with a communication radius of 10. The sink node is located at the corner of network.

Figure 2 shows that our method gives a satisfactory coverage performance. GAF and random awake method also provide full coverage but they require more active nodes as shown in Fig. 3. In a dense network, our method performs well with the coverage problem.

Fig. 5 Number of active nodes

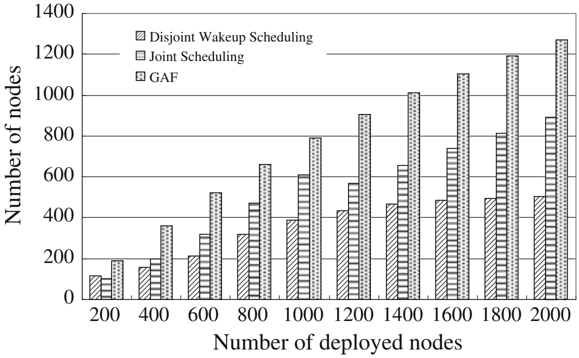


Figure 4 shows that our method provides successful connectivity in a dense network. It also awakes smaller nodes than GAF and Joint Scheduling method as shown in Fig. 5. In a dense network, our method performs well more apparently in terms of coverage and connectivity performances.

Conclusion

In this paper, we propose a disjoint scheduling algorithm for density control in wireless sensor network. It selects a small set of working nodes to avoid wasting excessive energy by turning off too many redundant nodes. It selects a small set of active nodes to improve energy efficiency while providing path connectivity to sink node.

Simulations showed that our scheme achieves the desired robust coverage as well as satisfactory connectivity to the sink with a smaller number of working nodes in an energy efficient fashion than GAF and Joint Scheduling method.

References

1. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. In: Computer networks, vol 38, pp 393–422
2. Chong C, Kumar SP (2003) Sensor networks: evolution, opportunities, and challenges. In: Proceedings of the IEEE, vol 91, pp 1247–1256
3. Ye F, Zhong G, Lu S, Zhang L (2003) Peas: a robust energy conserving protocol for long-lived sensor networks. In: ICDCS
4. Zhang H, Hou JC (2003) Maintaining sensing coverage and connectivity in large sensor networks. Technical report UIUC, UIUCDCS-R-2003-2351
5. Xu Y, Bien S, Mori Y, Heidemann J (2003) Topology control protocols to conserve energy in wireless Ad Hoc networks. Technical report CENS 0006
6. Zhang H, Hou JC (2003) Maintaining sensing coverage and connectivity in large sensor networks. Technical report UIUC, UIUCDCS-R-2003-2351

7. Cerpa A, Estrin D (2004) ASCENT: adaptive self-configuring sensor networks topologies. J IEEE Trans Mob Comput 3:272–285
8. Liu C, Wu K, Xiao Y, Sun B (2006) Random coverage with guaranteed connectivity: joint scheduling for wireless sensor networks. J IEEE Trans Parallel Distrib Syst 17(6):562–575
9. Li H, Yu D (2002) A statistical study of neighbor node properties in ad hoc network. In: Proceedings of ICPPW, pp 103–108

Part VII
Multimedia Computing

A Human Voice Song Requesting System Based on Connected Vehicle in Cloud Computing

Ding Yi and Jian Zhang

Abstract Traditional on-board vehicle computer (OVC) systems do not support speech recognition and do not have massive song library due to their limited storage capacity and computing power. Furthermore, it is very inconvenient to update the song library to keep it up-to-date. The advent of cloud computing in conjunction with mobile computing has enabled human voice song requesting and mass song playing from on-board computer in a vehicle. A system of human voice song requesting from OVC connected to the music cloud has been proposed. Self-adaptive speech recognition and speech recognition technology based on keywords from song titles has been adopted. By recognizing the voice on a song title in a cloud environment, the digital content of the selected song stored in the cloud is streamed to and then played back in the OVC. As a result, the system improves the safety of driving and enables automobile entertainment more humane.

Keywords Networking of vehicles · On-board vehicle computer (OVC) · Speech recognition · Human voice song requesting (HVSR) · Cloud computing · Music cloud

Introduction

Vehicle drivers tend to listen to songs while driving. However, it is inconvenient for a driver to operate keyboard, even a remoter while steering the vehicle wheel. Thus, drivers cannot choose songs on demand while driving a traditional vehicle.

Traditional on-board vehicle computer (OVC) system has small word bank and small linguistic model. It does not support speech recognition and does not have

D. Yi (✉) · J. Zhang
School of Computer Engineering, Shenzhen Polytechnic, Shenzhen, Guangdong, China
e-mail: yiding300@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_58, © Springer Science+Business Media Dordrecht 2014

massive song library due to their limited storage capacity and computing power. Furthermore, it is very inconvenient to update the song library to keep it up-to-date. The advent of cloud computing in conjunction with mobile computing has eliminated the above limitations, enabling human voice song requesting (HVSR) and massive song playing in OVC.

In this paper, a system of HVSR based on connected vehicle in cloud computing has been proposed and implemented. Self-adaptive speech recognition technology in combination with song title keyword speech recognition has been adopted. By recognizing the voice of a song title in the cloud environment, the digital content of the selected song stored in the cloud is streamed to and then played back in the OVC. Vehicle drivers can request songs by only speaking out its song title. As a result, the system improves the safety of driving and renders automobile entertainment more humane as well.

General Scheme

A HVSR system based on connected vehicle in cloud computing (as shown in Fig. 1) consists of OVC, HVSR management cloud server, music cloud servers and networking of vehicles. Among these components, the networking of vehicles and music cloud servers are built from existing online resources while the design and implementation of the OVC and the HVSR management server are our own innovations.

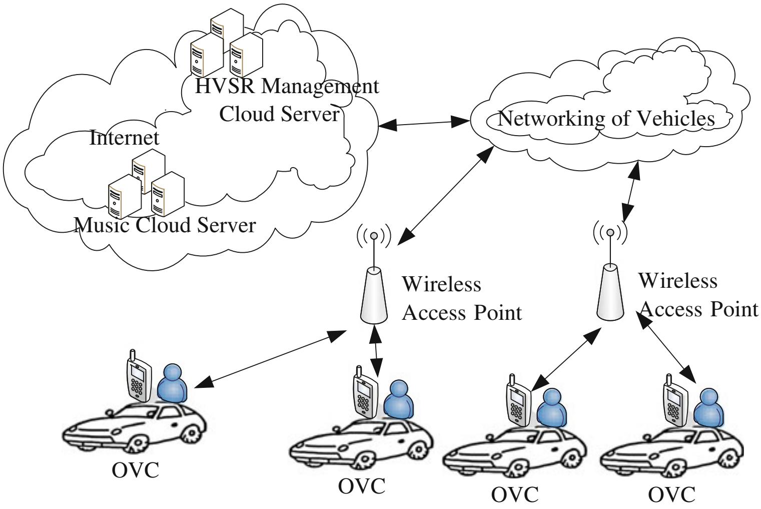


Fig. 1 General scheme

The networking of vehicles, provided by network service providers, is the access network by which an OVC is connected to the Internet.

The music cloud servers located in the Internet are the existing public resources. These servers host and control audio streaming data that is served as audio streaming source. In the Internet there have been many music servers that possess massive audio streaming files. An OVC needs to playback these audio streaming data from these public music cloud servers.

The HVSR management cloud server provides the service of speech recognition and song titles management. The speech recognition and song titles management is realized in cloud computing.

The client, OVC, installed in an automobile consists of hardware and software. It receives the service of speech recognition and song titles management from cloud services, and obtains audio streaming data from the cloud services to playback.

The Workflow and Dataflow

There is a one-push button in the OVC. When the button is pushed down, it starts to run, and when pushed again it stops working. The following is the workflow when the system starts:

- The OVC accesses the Internet via the networking of vehicles. It establishes connection to the HVSR management cloud server.
- The OVC receives song requesting voice spoken by a driver, and then sends the digital voice data to the HVSR management cloud server;
- The HVSR management cloud server receives the digital voice data of song requesting from the OVC and identifies the song's title through speech recognition. Then the music server that hosts the song audio streaming data is queried and the corresponding network address is sent back to the OVC.
- The OVC receives the above network address, establishes a connection to the music server of the given address. It then receives and plays the audio streaming from this music server [1].

Corresponding to the above workflow, the dataflow is shown in Fig. 2.

Fig. 2 Dataflow

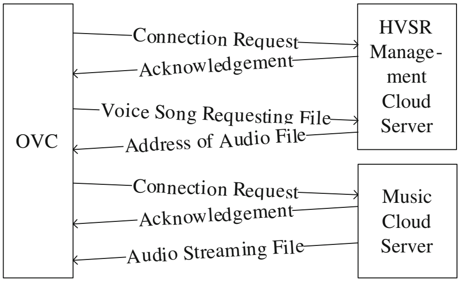
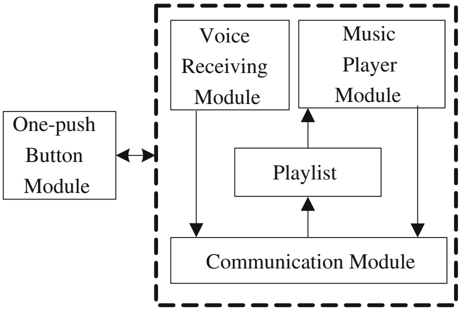


Fig. 3 OVC structure



The On-board Vehicle Computer

The client, OVC, installed in an automobile consists of both hardware and software [2]. It obtains the service of speech recognition and song titles management from cloud services, and receives audio streaming from the cloud services.

The OVC provides the following functions:

- Access to the Internet through the networking of vehicles,
- Receive users' voice song requesting,
- Communicate with HVSR management cloud server,
- Communicate with public online cloud music servers,
- Manage the playlist,
- Play the audio streaming data from cloud music servers.

The OVC consists of one-push button module, communication module, voice receiving module, playlist and music player module (as shown in Fig. 3).

The one-push button module in the OVC is used to start/stop the system.

The communication module in the OVC has a unique network address with which the vehicle is connected into the Internet. It is responsible for accessing to HVSR management cloud server and music cloud servers.

The voice receiving module in the OVC is responsible for detecting voice, and recording effective song requesting voice. The communication module sends this recording file to the HVSR management server.

The communication module in the OVC receives the song audio streaming file's network address from the HVSR management server, and then appends this address to the playlist.

The music player module in the OVC retrieves the server's network address attached to the current selected song from playlist. Then the communication module establishes connection to that music server, and receives the audio streaming data from the music server. Once the streaming data is received, the player module plays this audio streaming data.

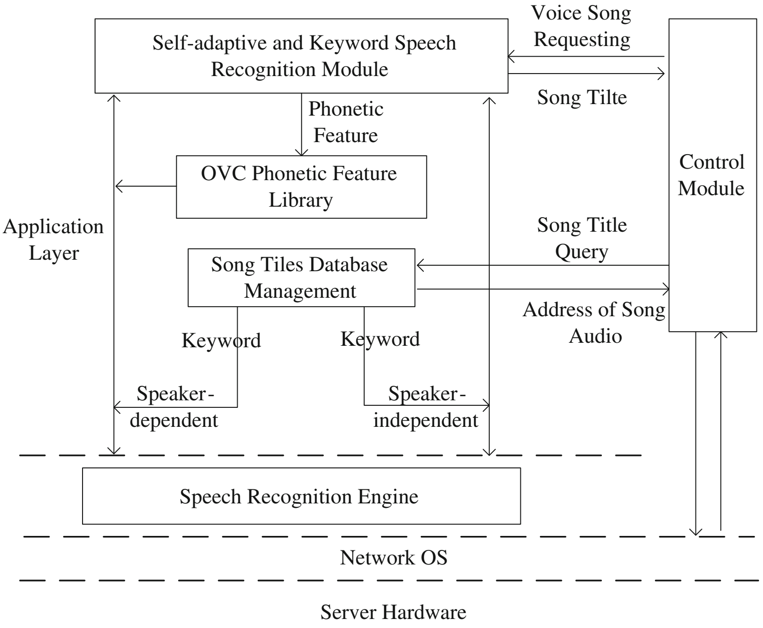


Fig. 4 HVSR management cloud server

The HVSR Management Cloud Server

The HVSR management cloud server (as shown in Fig. 4) provides the service of speech recognition and song titles management [3]. It adopts third party professional speech engine product that directly runs on top of the operating system [4, 5]. Speech engine is a middleware with the capabilities of both speech recognition and synthesis. Leveraging its services of speech recognition and synthesis in the application layer, the legacy keyboard operations such as text input, menu selection, etc. are now substituted by a user's voice requesting, which implements truly man–machine dialogue.

The service of speech recognition and song titles management that works in the application layer is mainly discussed in the paper.

The Service of Speech Recognition and Song Titles Management

The service of speech recognition and song titles management works in the application layer of network architecture (Fig. 4). It is realized by self-adaptive speech recognition and song titles keyword speech recognition technology module, song titles database management module, OVC phonetic feature library and control module.

The control module sends HVSR digital file from the OVC to the speech recognition module. The “song title” text is identified in the speech recognition module. Then the song audio streaming file’s network address in the song titles database is queried using this text as keyword. Finally the control module sends the server address back to the OVC.

The Song Titles Database Management

Each record in song titles database includes at least three fields: the song title, the singer name, the network address of the song audio streaming file.

There are many music servers in the Internet that provide massive song audio streaming files. Therefore it is sufficient to store the network address of song audio streaming files instead of complete audio streaming files in song titles database.

The song titles management implements the real-time maintenance of song titles database, renews the fields of the database in time. This is to ensure that new songs are appended up-to-date and the song audio streaming file’s network address is valid.

The ‘song title’ and the ‘singer name + song title’ in song titles database are searched as keywords in keyword speech recognition. So the song titles database has two fundamental functions: keyword speech recognition, and the query of the network address for a song audio streaming file.

The Song Title Keyword Speech Recognition

Applying keyword speech recognition technology, rather than applying continuous natural language speech recognition technology, is more effective in the HVSR application. For a continuous natural language, the same semantics can have different ways of expression. If continuous natural language speech recognition technology is adopted, the system becomes difficult to implement, and the rate of recognition accuracy will be low. Aimed at the special application environment of the HVSR, sufficient information is collected by using ‘song tile’ or ‘singer name + song title’ spoken by the user without the full natural language expression. Therefore, the technology of keyword speech recognition simplifies the problem and is easy to realize with higher accuracy rate of speech recognition.

The OVC Phonetic Feature Library

The OVC phonetic feature library includes OVC address field and its phonetic features field. It is used to support speaker-dependent speech recognition.

Every networking OVC has a unique address. In most cases, an automobile has one fixed driver or multiple drivers. Therefore, the phonetic characteristics from the same OVC are limited. These limited features can be stored in the phonetic features library. If an OVC accesses the system for the first time, the system creates its phonetic characteristics information record. Later new phonetic features from the same OVC are appended to the record.

Speaker-Independent/Speaker-Dependent Self-adaptive and Song Title Keyword Speech Recognition Technology

During speech recognition, the song title keyword in speaker-dependent speech recognition technology is used first [6]. If succeeded, it means that the phonetic feature of the OVC has already been stored in the library. If failed, it could be the case that this is a new access, thus there is not such phonetic feature recorded in the phonetic feature library. In such case, the system automatically switches to use speaker-independent speech recognition and song title keyword speech recognition technology to identify the selected song. After a song is identified, the new phonetic feature is appended into the record of the phonetic feature library.

Conclusions

While driving a vehicle driver needs to control the steering wheel. Thus it is inconvenient to operate buttons. In this paper, a system of HVSR based on OVC connecting to the music cloud has been proposed. A driver can request a song by only speaking song title. The song title is then recognized by speech recognition technology in a cloud environment. Then the digital audio content of the selected song stored in the cloud is streamed to and finally played back in the OVC.

The OVC installed in automobile is a combination of hardware and software. It obtains the services of speech recognition and song titles management from cloud services, and accesses audio streaming from cloud services as well.

Using speaker-independent/speaker-dependent self-adaptive in combination with song title keyword speech recognition technology in cloud computing is an important characteristic of this system.

The advent of cloud computing in conjunction with mobile networking of OVC has eliminated the traditional limitation, making HVSR and mass song playing from OVC into reality. The methodology and the implemented system in the paper allow a driver to request a song by only speaking the song title. We believe this renders automobile entertainment more humane while considerably improving the safety of driving as well.

References

1. Wang Z (2004) Research in the application of streaming media technology. Comput Eng 5:34–36
2. Chen X (2005) The design of embedded streaming media player. Zhejiang University, pp 17–19
3. Zhao B, Yan F, Zhang L, Wang J (2012) The construction of dependable cloud computing environment. China Comput Fed Commun 8(7):28
4. Windows Azure blog: http://blogs.msdn.com/b/azchina/
5. Microsoft Tellme Speech Innovation: http://www.microsoft.com/en-us/Tellme/default.aspx
6. Li N, Xu S, Ma Z, Shi L (2011) Simulation research on adaptive speech recognition algorithm. Comput Simul 8:181–185

Access Control System by Face Recognition Based on S3C2440

Yongling Liu, Yong Lu and Yue Song

Abstract Face recognition technology has become more and more important in the field of biometric identification because of its advantages. A face recognition system combining hardware and software based on S3C2440 is presented in this paper. We adopt S3C2440 embedded development board as the hardware platform. This paper creates an embedded Linux software platform and our system is divided into four sections by their function: graphical user interface, image capture, face detection and face recognition. The whole system is tested as an access control system in practical situation. Experimental results show that our system has a good performance in both accuracy and efficiency of face recognition in access control.

Keywords S3C2440 · Face recognition · Access control system · Linux

Introduction

Face recognition has been widely used in access control and authentication because of its benefits such as accuracy and fast identification, high usability and security. Previous face recognition system was designed under the platform of personal computer at the cost of portability. Therefore, with the rapid development

Y. Liu · Y. Lu · Y. Song (✉)
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing, China
e-mail: sycat7@gmail.com

Y. Liu
e-mail: 10120009@bjtu.edu.cn

Y. Lu
e-mail: ylu@bjtu.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_59, © Springer Science+Business Media Dordrecht 2014

of embedded technology, an embedded face recognition system is more popular. An embedded face recognition system would entail many advantages including low cost, integration with other technologies and optimization for real-time operations.

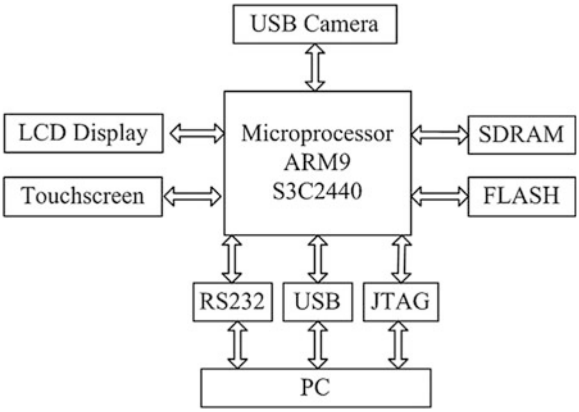
Now, access control system has evolved as an intelligent management system of Entry/Exit Control instead of physical key management. With such a smart access control system, it can ensure more security on the condition that the people lose or forget the ID card. Therefore, it is meaningful and useful to carry out research on embedded face recognition system for future security and convenience.

System Design

Hardware framework of the access control system is shown in Fig. 1. The face images are extracted from the USB camera and then processed with S3C2440 embedded processor by means of face recognition algorithm. SDRAM and FLASH work together as memory devices which host temporary files data and permanent data. Besides, the human–computer interaction module consists of LCD and touch screen. The RS232 serial port or JTAG port connects the access control system and PC to communicate each other.

And our system is divided into four sections by their software function: graphical user interface, image capture, face detection and face recognition.

Fig. 1 Framework of the access control



Hardware Platform

Our system is based on FL2440 FORLINX embedded development board. The major descriptions of FL2440 are following:

- CPU: S3C2440A processor made by Samsung, main frequency 400 MHz.
- Memory: SDRAM, 4 Banks * 4 Mbits * 16 bits.
- Flash Memory: 256 M Nandflash, 4 M Norflash.
- System clock: 12 MHz system external clock source.
- LCD: 3.5 inch touchable screen supports TFT type.
- Interface: 2 pieces of RS232 serial ports and 4 pieces of USB host interfaces.

Besides, we add a USB camera into the system to capture face images. This camera (ZC301) is created by VIMICRO Corporation.

Software Platform

Our platform is based on Linux operating system. To complete the platform, the following work is been done:

1. Building up cross compiling environment

In this paper, the operating system for the prototype machine is Red Hat 9. We install cross compiler of version 3.4.1.

2. Portability of Boot loader and kernel

Our system adopts the U-Boot of version 1.3.3 as boot loader and Linux version 2.6.12 as our kernel. An executable file named U-Boot.bin has been generated finally. Besides,

3. Root file system

Our system chooses the document system with yaffs2 format based on NAND FLASH. We complete the file system by means of BusyBox to generate an executable file.

4. Device drivers support

There are two cases in Linux device drivers. One is the hardware supported by kernel. The other one is the hardware not supported by kernel such as the USB camera used in our system. The drivers can be available by the third party and dynamic load into the kernel.

Software Design for Face Recognition System

1. User interface Module

Our user interface is designed by QT. First, we download the source code of QT-x11 and QT-embedded. The QT-x11 will generate QT development tools and qvfb by which could enable embedded development environment simulation to develop QT projects under the situation of x86. Particularly, the QT-embedded is optimized for embedded development. Then, QT development environment will be established by compiling, installing and configuring.

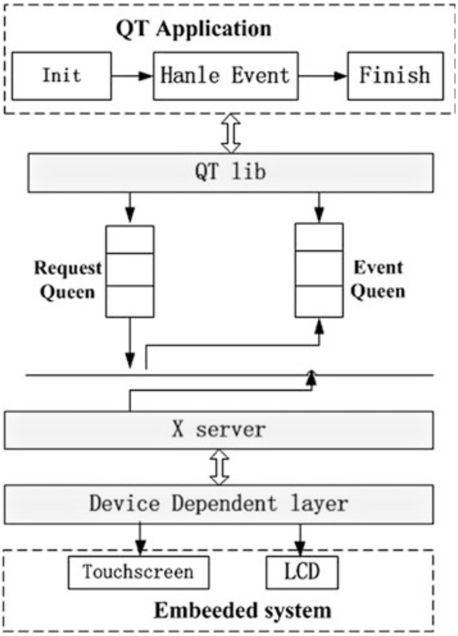
The flow chart of user control interface is followed as Fig. 2.

The system interface is in two parts: window button and display area. When a user click the button, the system will call program such as pre-processing and face extracting to process face images and then output results in the display area.

2. Image capture Module

Our system will collect images by means of Video4Linux (hereafter referred to as V4L). The image capture module is designed based on API functions offered by V4L. The program of image capture is realized as Fig. 3 and the capture result in real solution is shown in Fig. 4.

Fig. 2 The flow chart of user control interface



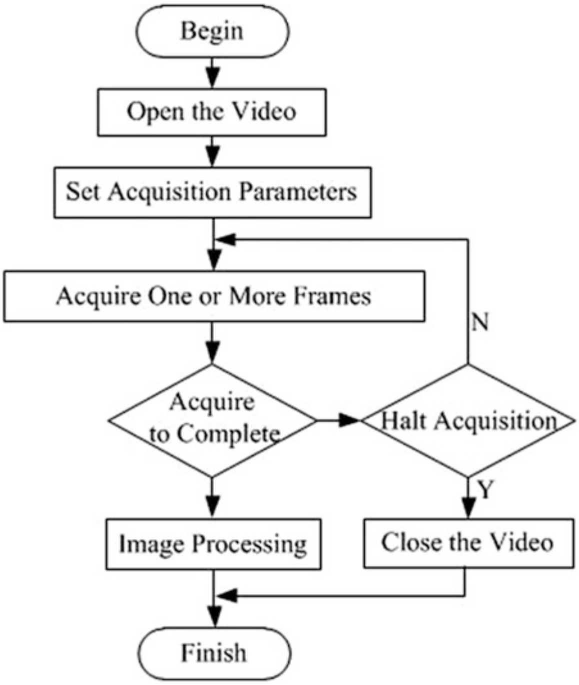


Fig. 3 The flow chart of image captures

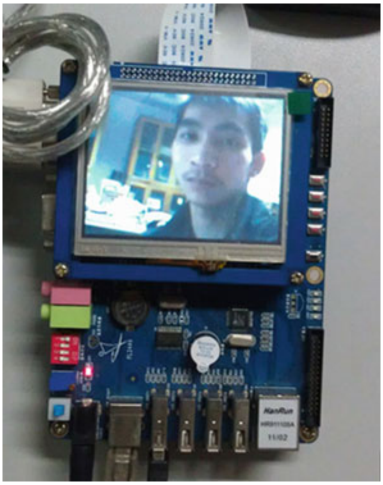


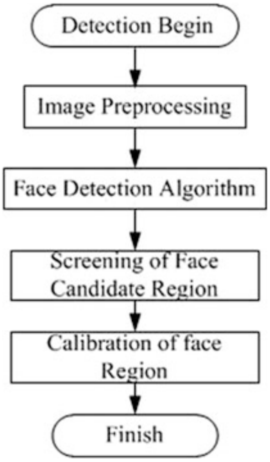
Fig. 4 Capture results

3. Face detection Module

The major function of face detection module is to locate the face in a captured image from the USB camera. The flow chart of face detection is followed as Fig. 5.

In this paper, we adopt method of haar+Adaboost as detection algorithm. Calculating the features of haar-like by such a method will be more efficient.

Fig. 5 The flow chart of face detection



Furthermore, Adaboost learning algorithm could combine weak classifiers into one strong classifier with feature selecting and classification training. We use the algorithm based on cascade classifiers instead of skin color because in comparison to skin-color-based algorithm, cascade-classifier-based method will offer a more accurate location in the situation that the background is similar to skin color or skin on other parts of body exposed besides human face. Major steps of detection are following:

- Select a proper classifier.
- Read images or videos.
- Detect human faces.
- Display the target and locate the human face.

From the Fig. 6, we can see the detection performance from a dynamic image captured by the USB camera.

4. Face recognition Module

Now we need to extract features describing the face for recognition by face recognition module. Principal Component Analysis (PCA) and 2 Dimension Principal Component Analysis have a wide range of applications for feature extraction algorithms. Compared to PCA, 2DPCA constructs covariance matrix by its image matrix directly. Therefore, the 2DPCA algorithm has a better performance in both accuracy and computation speed. However, the advantages of 2DPCA are obtained at the cost of a high memory usage during the period of image reconstruction. In this paper, we combine PCA and 2DPCA to face recognition. That is a two-stage strategy that we apply 2DPCA to reduce image's dimensionality at first stage and then extract features from prior outputs based on PCA algorithm. This strategy will help the system to save memory and reduce computational cost.

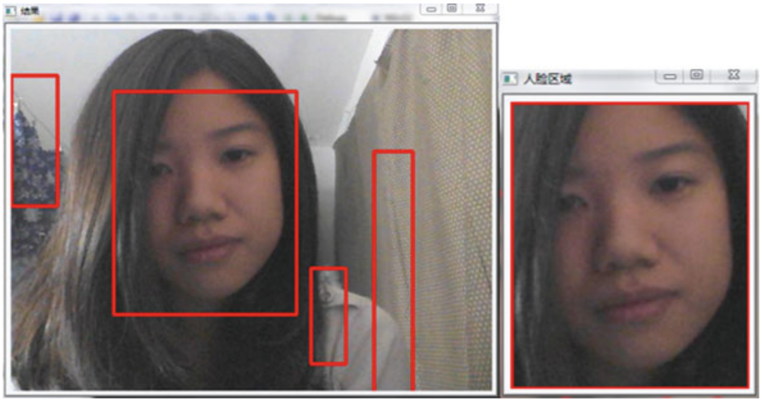


Fig. 6 Face detection performance

Besides, our face recognition algorithm is based on gray scale images while the captured images are color images of RGB. We must make pre-processing such as grey level transformation and normalization before training. In our system, the image resolution is 65 × 80 and the training images after pre-processing are shown in Fig. 7.

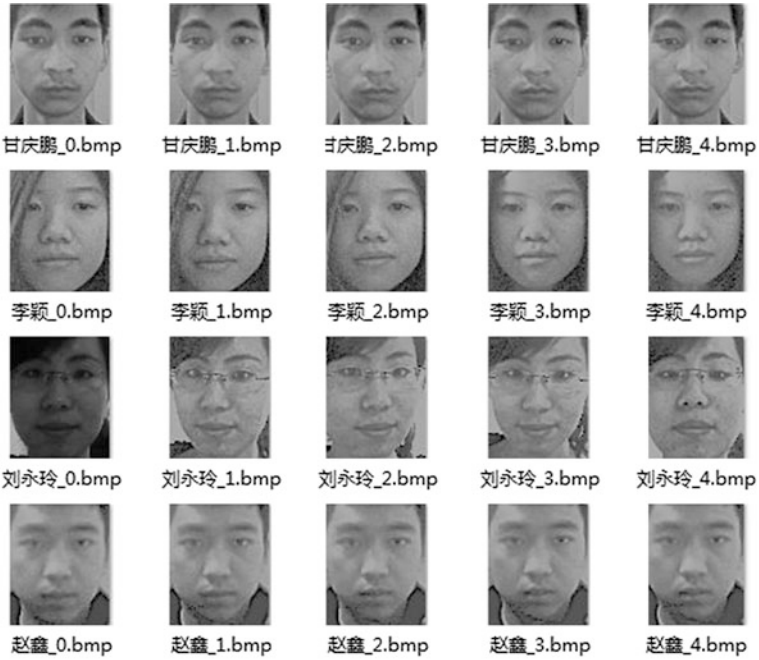


Fig. 7 Normalized face images

There are two parts in face recognition: the training process and the matching process.

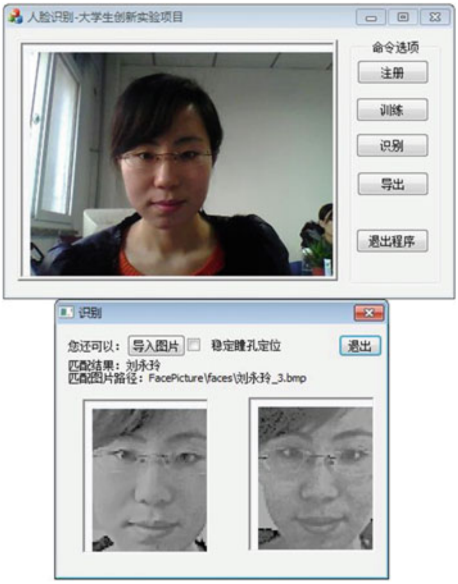
Firstly, the training procedure works as following:

- Create feature subspaces from the training set by function prototype shown below.
  ```
  bool CreateEigenFaceSpaces(IMG *src, MAT *dest, MAT *EigenVect, int num);
  ```
- The eigen subspaces can be constructed by image projection. The function prototype is:
  ```
  bool ProjectSubEigenFace(IMG *src, MAT *dest, MAT *SubSpace, int num);
  ```

In addition, the recognition procedure is following:

- Project the test image into the eigen subspace by function:
  ```
  bool ProjectSubEigenFaceForSingle(IMG *src, MAT *dest, MAT *SubSpaceForSingle);
  ```
- Compare it with training images. Then output the matching result.
  ```
  int Compare_Ma(MAT *SubSpaceForSingle,MAT *SubSpace, MAT*EigenVect,int num);
  ```

From the Fig. 8, we can see the face recognition system performs well. In the window, the left one is detected human face while the right one is registered face image. If the two images are matched, the system will display the registered information of the person and the authorized person will be allowed to enter the secure area.



匹配结果：刘永玲
匹配图片路径：FacePicture\faces\刘永玲_3.bmp

Fig. 8 Face recognition

Table 1 The results of face recognition

Training samples of per person	1	2	3	4	5
Recognition rate	0.80	0.83	0.85	0.87	0.91

System Test

We can set up a database by 100 photos to do a test with this system. These photos are taken of 20 people in the lab, each with 5. The dimensionality is 55. We need to observe the changes of recognition rate when the samples are changed from one to five. When we conduct the experiment, we need to ensure that light is much enough and that the faces of the tested person stay a little while in front of the camera. We can come to the conclusion in the following sheet1 (Table 1).

The more the samples are, the higher the recognition rate is. After a series of experiments based on the five samples, the recognition rate can be increased to 90 % or higher. But if the light is not enough, shooting angle change is big, or facial expression changes great, it is easy for us to make mistakes on recognition. Therefore, when doing a registration, we need to try to take different photos from different angles with different facial expressions, which benefits increasing recognition rate.

Conclusion

This thesis mainly covers a research on face recognition access control system based on S3C2440. During the process, I divided the whole system into four functional parts, and respectively design and fulfill each of the tasks, including developing graphic user interface, image collection, face detection, and the algorithm of face detection rate. The results manifest that the system needs to be improved in the following aspects:

1. Algorithm needs to be improved or new algorithm needs to be proposed, in order to get a better effect on face detection.
2. Fully take advantage of the resources of the system, for example, increasing network transmission patterns.

References

1. Zhang J, Yan Y, Lades M (1997) Face recognition: eigenface, elastic matching and neural nets. Proc IEEE 85(9):1422–1435
2. Meng L, Nguyen TQ, Castanon DA (2000) An image-based Bayesian framework for face detection. In: Proceedings of IEEE conference on computer vision and pattern recognition, Hilton Head Island, South Carolina, USA, pp 302–307

3. Lu H, Shi W (2005) Accurate ASM for human face image search. The 17th IEEE international conference on tools with artificial intelligence, pp 642–647
4. Xin X, Lansun S, Kongqiao W (2000) Automatic human face detection. ISO/IECJTCI/SC29/WGll MPEG99/M6144, Beijing, China, July 2000
5. Hsu RL, Abdel-mottaleb M, Jain AK (2002) Face detection in color images. IEEE Trans Pattern Anal Mach Intell 24(5):696–706
6. Zhang J, Yan Y, Lades M (1997) Face recognition: eigenface, elastic matching and neural nets. Proc IEEE 85(9):1422–1435
7. Hadid A, Pietikinen M, Ahonen T (2004) A discriminative feature space for detecting and recognizing faces. In: IEEE conference on computer vision and pattern recognition (CVPR)

A Secure Digital Watermark Skeleton Based on Cloud Computing Web Services

Jian Zhang and Ding Yi

Abstract More and more digital watermark systems provide the services via the internet. The watermark embedding and detection technology components are the parts of the web services of the sites. This paper designs a secure digital watermark skeleton based on cloud computing, pushing these web services to the cloud. The entire skeleton is based on cloud computing architecture, including service delivery, data storage and security, data lifetime, etc. The skeleton combines the digital watermark and cloud computing technologies with the certain reference value, it also provides efficient digital watermark within the large number of complicating requests to protect digital products in the marketing of multimedia network environment.

Keywords Digital watermark · Cloud computing · Web service

Introduction

Digital watermark is an important direction of the multimedia technology within the field of information security. It protects the copyright of the original data by embedding secret information, such as the watermark in the original data. This watermark can be embedded into the original data as a paragraph of text, logos, serial numbers, etc. The watermark is with raw data (images, audio, video, etc.) and hides them closely. It can be retrieved via the specific algorithm so that it can protect the copyright of the multimedia products.

With the internet technology and object-oriented component technology's rapid development, it promotes the internet within the distributed computing technology

J. Zhang (✉) · D. Yi
School of Computer Engineering, Shenzhen Polytechnic, Shenzhen, China
e-mail: jzha930@szpt.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_60, © Springer Science+Business Media Dordrecht 2014

changes. In this environment, the web services-based digital watermarking technology, as a new distributed computing systems, has been proposed and so that enterprises in property rights protection, quickly deploy solutions and explore new opportunities. There for, many of the original system designs and developments are getting re-deployed through again as a web services system there. Some of digital watermarking systems are web-based. They can provide the services of watermark embedding and detecting via the internet [1, 2].

According to the characteristics of web services, this paper provides a cloud computing web services based secure digital watermark skeleton. It combines the digital watermark and cloud computing technologies, providing the watermark embedding and detecting services via the cloud. The author of the multimedia products just uploads the products to the web site, the cloud will call the relevant applications to embed the watermark or detect the watermark. In the course of using the services there are many aspects of security problems should be concerned: the safety of the web services environment and the security of digital watermarking itself. These are service delivery, data storage and security, data lifetime, etc. This paper will analyze these issues in detail and propose a workable solution.

The Background of the Project

There is a company which provides a network platform for picture trading. The system is embedding the lightweight watermark into the picture when authors upload their own products to ensure non-destructive and products of digital watermarking robustness, and the system can provide the detecting services [2]. The company is getting very successful with the picture trading. Now the company wants to provide the professional digital watermarking services to many commercial entities via the internet. According to their marketing records, they found it is difficult to provide the complicating requests services via the traditional web services. They want to deploy the web services and combine them as a cloud so that they can complete the mission.

Web Services System and Its Key Technology

Web services are self-contained self-describing modular components which can be publishing, locating and calling through the Web. Web Services system in a number of existing technologies (such as HTTP, XML) added some new standards. Here are the four key technologies [3].

- Simple Object Access Protocol, SOAP, is XML based for a decentralized, distributed environment, a lightweight protocol for exchanging information.

SOAP in between requestor and provider objects defines a communication protocol.

- Web Services Description Language, WSDL, provides a description of the service mode. Through a complete description of the service, the service requester can access the service methods and know the specific location; service developers also can use this interface is a standard compatible development services.
- Universal Description, Discovery, and Integration, UDDI, provides Web Services Publishing and discovery methods. Through open standards, Internet on the entity to find each other and call each other's Web Services.
- Web Services Security [4]. In the W3C XML Signature and XML Encryption based on the specification provides three mechanisms to extend the SOAP message: trustlike transmission, message integration and message confidentiality. WS Security and other Web services protocols can work together to meet a variety of types of application security needs.

In the environments of dynamic e-commerce, the original digital watermarking system migrates to web services system and releases it as a service and registration. The process will use SOAP, WSDL, UDDI and other key technologies. Because the digital watermark itself and security issues are closely related, so how to ensure that web services-based digital watermarking system security is a particularly important issue.

Cloud Computing Skeleton and Its Key Tehchnology

According to the characteristics and key technologies of web services, this paper proposes a digital watermark skeleton based on cloud computing web services. This skeleton combines the digital watermark and cloud computing technologies. It provides the watermark embedding and detecting services from the cloud. Innovations include:

- The watermark embedding and watermark detection technology as web services components in the form of a "cloud" released over the internet.
- The whole system is based on cloud computing architecture Web service platform, through the Simple Object Access Protocol (SOAP) exchange service requestor and service information between providers.
- Digital certificates, XML encryption and digital signatures, and other security technology to ensure the security of information exchange.
- Data storage, security and life time will be the key aspect of the skeleton.

The skeleton proposed to reduce the cost of the development and use of digital watermarking system, effectively protect the property rights of digital media products in an open network environment.

Paper [2] discuss the detail of the web services-based security model for digital watermarking. The following will discuss the key technologies of this skeleton involves service delivery, data storage and security, data lifetime, etc.

Service Delivery

The model is based on cloud computing and SaaS technology. The core of the model is service level. It will provide better availability management, capacity management, service level management, financial and risk management, continuity management in the field of SMS/MMS/CRBT/OSS/BSS. It delivers the five management processes and basic infrastructure [5].

Figure 1 shows the features of the A&S Cloud structure, they are: Business logic and calculation of storage are separated; Resources abstraction and sharing; Intelligent and automatic control. These features will be the direction of future research.

Data Storage and Security

Cloud computing usually consists of front-end users who possess clients devices and back-end cloud servers. This paradigm empowers users to pervasively access a large volume of storage resources with portable devices in a distributed and cooperative manner. During the period between uploading and downloading files

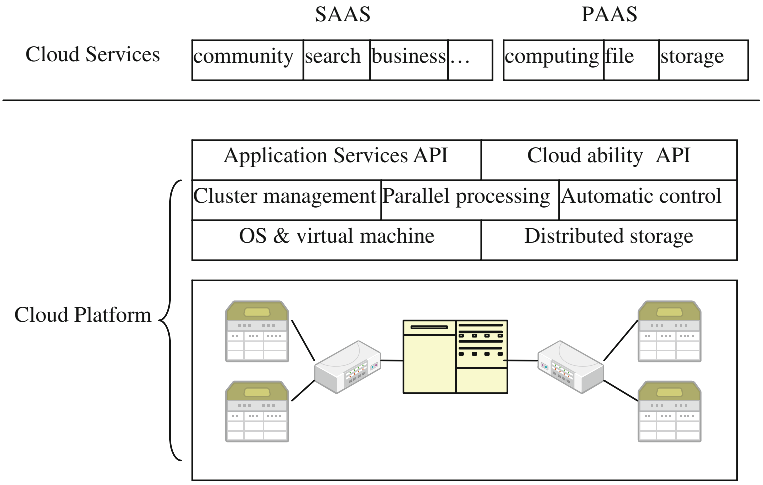


Fig. 1 A&S cloud structure (*Source* Huawei A&S cloud computing solutions report, p. 11)

(data), the privacy and integrity of files need to be guaranteed. To this end, paper [6] proposed a family of schemes for different situations. All schemes are lightweight in terms of computational overhead, resilient to storage compromise on clients devices, and do not assume that trusted cloud servers are present. Corresponding algorithms are proposed in detail for guiding off-the-shelf implementation. The evaluation of security and performance is also extensively analyzed, justifying the applicability of the proposed schemes.

As cloud providers have priority access to data, it is difficult to guarantee the confidentiality and integrity of users data. For this problem the paper [7] presents an architecture to protect user data security by encryption. Searching cipher text and integrity verification will be used to ensure the availability and integrity of users data. This mechanism provides reliable security support for the massive data in the cloud computing system.

Data Lifetime

Data privacy protection is one of the primary concerns and major challenges for online services, such as cloud computing and outsourced data center. The concern is getting serious with the computing practices shifting towards cloud computing. Once users data is uploaded, end users are hard to guarantee that the data is protected and can be completely destructed by any means. Users can only rely on blind trust on the online service vendors. However, the privacy of user data can be compromised in multiple ways including careless operations of cloud administrators, bugs and vulnerabilities inside cloud infrastructure and even malicious cloud vendors. Paper [8] seeks to provide users with a concrete way to protect or destroy uploaded data. It utilizes the technique of trusted computing as the trusted root in the hardware layer, and the hypervisor as the trusted agent in the software layer. The trusted hypervisor is responsible for protecting sensitive user data or destructing them at users command. Even administrators of the cloud cannot bypass the protection. This paper presents Dissolver, a novel system that keeps the data privacy in the whole life time and ensures the destruction at the users command. Performance evaluation shows that the prototype system imposes reasonably low runtime overhead.

Web Services Cloud Computing-Based Digital Watermarking Model and Security Analysis

The purpose of the skeleton is to migrate the web services environment to cloud level. Any service requester can publish service requests by searching through the UDDI information, the service interface description and SOAP protocol. From the service requester perspective, embedding a watermark to the image request to the server needs to provide three aspects:

- Original image data: it is the carrier of the watermark.
- Watermark: It is the identification of the copyright owner of images and image basis for legitimacy.
- Watermark extended information: In the detection of watermark, digital watermarking service using the extended information to identify the real watermark, the watermark and the extracted contrast, to determine whether the watermark has been tampered with.

The algorithm of the digital watermark has been achieved. The development of corresponding interfaces, the original components of the digital watermark embedding and watermark detection as a service released [1, 2]. According to paper [1] discussed security-based digital watermark over the internet, the scheme can provide the encryption and digital signature information in the secure transmission of services over the internet. Paper [6–8] proposed relevant algorithms to secure the data up in the cloud.

Conclusion

Digital watermark embedding and watermark detection over the cloud can reduce development and spending costs. The single web service of prototype digital watermarking systems using the Java language has been implemented on the Apache server, and is in trial operation [2]. The system uses digital watermarking for multimedia services, online services and e-work applications such as instances of rights management to provide a safe and feasible solution. The system will effectively protect digital property rights in an open network environment. Once the different web services from different company upload over the cloud, the followings will be the key issues to study: the security of the digital watermark, cloud computing security and compliance, web service dynamic balance.

References

1. Hu W, Tong C, Chen D, Li Z, Kou W (2004) Secure digital watermark scheme based on web services. J Zhejiang Univ (Eng Sci) 38(11):1441–1445. Zhejiang University, Hangzhou
2. Zhang J (2011) A web services-based security model for digital watermarking. In: 2011 international conference on multimedia technology (ICMT 2011), pp 4805–4808. IEEE press, New Jersey
3. Booth D, Haas H, McCabe F, Newcomer E, Champion M, Ferris C, Orchard D (2011) Web services architecture. http://www.w3.org/TR/ws-arch/, 6 Jun 2011
4. Atkinson B, Della G, Hada S (2002) Specification: web services security (WS-Security). http://www-106.ibm.com/developerworks/library/ws2secure/, 4 May 2002

5. Huawei Technologies Co., Ltd (2009) Huawei A&S cloud computing solutions. In: Huawei technical report
6. Wei R, Linchen Y, Ren G, Feng X (2011) Lightweight and compromise resilient storage outsourcing with distributed secure accessibility in mobile cloud computing. Tsinghua Sci Technol 16(5):520–528. Tsinghua University, Beijing
7. Dapeng Z, Ke C, Min Z, Zhen X (2011) The research of cloud computing data security support platform architecture. J Comput Res Dev 48(Suppl):261–267. Institute of Computing Technology, Beijing
8. Fengzhe Z, Jin C, Haibo C, Binyu Z (2011) Lifetime privacy and self-destruction of data in the cloud. J Comput Res Dev 48(7):1155—1167. Institute of Computing Technology, Beijing

Video Transmission Quality Improvement Under Multi-Hop Wireless Network Architecture

Chih-Ang Huang, Chih-Cheng Wei, Kawuu W. Lin and Chih-Heng Ke

Abstract As wireless local area networks have been widely deployed and smart devices are becoming common in human everyday life, the demanding for video applications, such as video conference, video on demand, and video games is still increasing. However, the bandwidth in wireless networks is limited and channel quality is varying. As a result, how to provide a better delivered video quality over wireless networks is challenging. Also, the current solutions to better video transmission are almost only considering single hop transmission. When those solutions are applied in multi-hop wireless networks, such as wireless mesh network or VANET, the delivered video quality may not be good as one hop transmission. Therefore, we propose a new mechanism to improve video transmission in multi-hop wireless networks. The new mechanism contains two methods. One is to privilege the packets that have already traversed more hops. The other is control the number of retransmission for video packets with different importance. Through NS2 simulations, the results show the effectiveness of proposed mechanism.

Keywords Multi-hop wireless networks · Video transmission · MPEG · TTL

Introduction

Advances in wireless network technology with increased bandwidth and the benefits of unwired connectivity have overcome numerous inconveniences commonly experienced in cabled networking. For example, through wireless

C.-A. Huang (✉) · C.-C. Wei · K. W. Lin
Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan, Republic of China
e-mail: smallko@gmail.com

C.-H. Ke
Department of Computer Science and Information Engineering, National Quemoy University, Kinmen, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_61, © Springer Science+Business Media Dordrecht 2014

distribution systems (WDSs), a wireless network might continually extend its coverage by joining two or more WDS-capable nodes. The network architecture that enables data transmission by relaying packets from one node to the next between the sending and receiving end is termed a multi-hop wireless network. Most studies regarding multi-hop wireless networks have focused on increasing the volume or speed of data transmission [1–6], but few have examining improving the transmission quality of videos. Therefore, in this paper, the transmission quality of video streams through IEEE 802.11 [7, 8] and a multi-hop wireless network is examined.

Related Work

In wireless local area network communications, the IEEE 802.11 protocol is employed to access a shared microwave band of the transmission medium. Because of this characteristic of multi-hop wireless networks, preceding nodes and subsequent nodes compete for access to the channel, as shown in Fig. 1. Packet collisions resulting from this competition might cause latency and additional back-off time, thereby continually reducing the network bandwidth. Because of the competition between nodes, multi-hop wireless networks provide less bandwidth and longer latencies than single-hop wireless networks do. In this paper [9], a novel node-based priority calculation mechanism is presented to ease competition by granting higher priorities of channel access to packets that are sent earlier than later ones. When packets must be transmitted to a destination through a node, packets that are sent earlier compete with subsequent packets for channel access. However, if subsequent packets obtain the channel first, the utility is minimal because they cannot start transmission before the earlier packets finish transmission. Therefore, by granting channel access to earlier packets, the latency can be reduced and the bandwidth increased. Therefore, regarding access to the communication channel, the method presented here assigns a higher priority to earlier packets. In conventional techniques, all packets have equal priority for packet transmission, and early packets might compete with those sent later for channel access. The method proposed in this study increases the priority of the packets after passing through each node to reduce the competition between packets. In other words, the transmission priority is determined by the distance to the

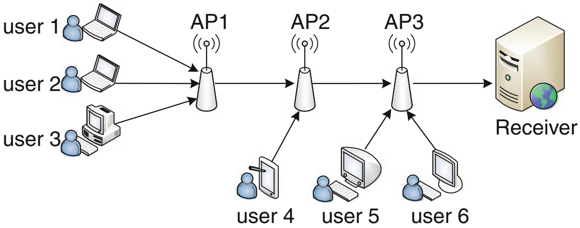


Fig. 1 Network environment of multi-hop wireless network

receiving end: Packets with shorter travel distances are assigned higher priority. By specifying a smaller contention window, packets closer to the receiving end gain access to the channel more easily. Similarly, by specifying a larger contention window, packets farther from the receiving end have a lower priority of access to the channel, and gaining access to the channel becomes more difficult. In this manner, higher transmission quality and operational efficiency in multi-hop wireless networks are achieved.

The Proposed Method

In addition to existing IEEE 802.11 specifications, this paper proposes adding two new mechanisms to improve the quality of video transmission in multi-hop wireless networks. The two mechanisms are described as follows (Tables 1 and 2):

Table 1 TTL and *retransmission* mechanism algorithm

TTL mechanism algorithm	Retransmission mechanism algorithm
If (qlen>threshold)	*I frame packet*
{	set RetryLimit = normal;
△TTL = default_TTL – current_TTL;	*P frame packet*
prob = ((qmax-qlen)/qmax)*(△TTL/maxhops);	if(qlen>threshold)
tmpx=random(0.0 ~ 1.0)	set RetryLimit = less;
if(prob≥tmpx)	otherwise
enque(packet)	set RetryLimit = normal;
else	*B frame packet*
drop(packet)	if(qlen>threshold)
}	set RetryLimit = min;
	otherwise
	set RetryLimit = normal;

Table 2 Description of mechanism parameters

Parameter	Description
qlen	Current total queue length
threshold	Threshold of network congestion
default_TTL	Default TTL value of packet
current_TTL	Current TTL value of the packet
prob	Probability that the packet enters the queue
qmax	Total queue length
maxhops	Total number of nodes
△TTL	Total number of nodes the packet passes through
Tmpx	Randomly generated probability value (0.0–1.0)
RetryLimit	Maximum number of packet retransmissions
normal	Normal number of retransmissions
less	Fewer numbers of retransmissions
min	Minimum number of retransmissions

- **Time to Live (TTL) Mechanism**

Time to live is employed to protect packets traveling through numerous hops. Having passed through multiple contentions and approaching the receiving end, such packets should be granted higher priority through protective measures to increase the opportunity to enter the queue and the success rate for transmission, otherwise the video transmission bandwidth is affected and transmission quality compromised.

- **Retransmission mechanism**

Vital packets are protected by altering the number of retransmissions. Because the space for node queuing is limited, if lower-order packets are retransmitted continually, the packet queue might overload and lose the higher-order packets because of its inability to accommodate them in the queue. The loss of higher-order packets might have severe consequences on video decoding and transmission quality.

TTL Mechanism

TTL refers to the longest time that a packet can live in a network. The TTL value of each packet varies in different operating systems. TTL primarily works by assigning a value to the packet transmitted by the sending end; the value decreases by 1 each time the packet passes through a node. When the TTL value drops to zero, the packet is discarded by the node that receives it. Specifying a TTL value prevents packets from random continuous travel among various network nodes, which is caused by unpredicted factors, before reaching the intended destination; this avoids affecting the overall network bandwidth and quality.

This mechanism first determines the number of nodes the packet has traveled through, and based on an algorithm, it assigns a higher priority to packets that have traveled through more nodes (i.e., by assigning a smaller TTL value) so that these packets have a greater opportunity to enter the queue for transmission. By assigning varying weights to packets through TTL values, packets that have traveled more nodes have a greater possibility of being transmitted to the receiving end. This mechanism is based on the concept that in a wireless network environment, each packet must pass through at least one node to obtain a transmission opportunity before it is transmitted again. In a multi-hop wireless network, all packets must pass through multiple nodes before reaching the receiving end; in other words, the packets must compete for priority in each node to be successfully transmitted to the receiving end. Conversely, when the wireless network environment is congested, the ability of the packets to travel through multiple nodes and successfully arrive at the receiving end decreases, thereby reducing the quality of the video transmission. Therefore, the use of the TTL mechanism could protect packets that have traveled through numerous hops by prioritizing them in the transmission queue, thus increasing the chance of successful transmission and greatly enhancing the bandwidth and video transmission quality.

According to the TTL mechanism algorithm, when the queue length surpasses a set threshold of network congestion, the number of nodes that a packet has passed through is equal to the default TTL value minus the current value. The probability of the packet entering the queue is determined by two factors: (1) the ratio of vacant space to the total length of the queue; a smaller remaining space represents a smaller probability of the packet being granted entry to the queue; and (2) the ratio of the number of nodes the packet has passed through to the total number of nodes (△TTL). The more nodes that a packet passes through increase the probability of the packet entering the queue; conversely, fewer nodes reduce the probability of a packet entering the queue. In addition, a probability value (tmpx) is randomly generated, with a value ranging between 0.0 and 1.0. When the probability of entering the queue (prob) is greater than or equal to the tmpx, the packet enters the queue for subsequent transmission; otherwise, the packet is discarded.

Retransmission Mechanism

Both IEEE 802.11 Distributed Coordination Function (DCF) and IEEE 802.11e [10] Enhanced Distributed Channel Access (EDCA) consist of a retransmission mechanism that might appear in two forms: the time limit and retry limit. This study is based on the retry limit mechanism because after compression by MPEG-4, video streams are encoded into I-frames, P-frames, and B-frames, which are subsequently encapsulated into frame packets. During transmission, these frame packets may experience unsuccessful transmission because of collisions or other factors. Under such circumstances, the working node determines the number of packet retransmissions based on the retransmission mechanism. For example, suppose an I-frame packet has a retry limit set to four. In the event that the frame packet transmission fails, the working node competes for its right to transfer again and retransmits the packet once retransmission permission is granted. If the frame packet cannot be transmitted to the next node after four attempts, the working node would discard the frame packet, thereby protecting the priority rights of other frame packets.

The retransmission mechanism is primarily based on adjusting the number of packet retransmissions. When a packet transmission fails, the so-called retransmission mechanism activates; when a packet fails to be correctly transferred to the receiving end, the node initializes packet retransmissions. In the current IEEE 802.11 configuration, the number of retransmissions is fixed and a priority is not assigned based on packet importance. In the present study, the varying effects of frames with differing importance levels on video transmission are considered; therefore, different probabilities for retransmission are assigned to frame packets based on their priority to enhance the quality of video transmission. When the network is busy, frame packets with higher priority are assigned more retransmissions, and lower priority frame packets have fewer opportunities for retransmission. In this manner, the probability of losing frame packets with higher priority is significantly reduced, and the likelihood of successfully transmitting frame packets with higher priority is increased; thus, the overall quality of video transmission improves considerably. Because the size of the queuing buffer at each node is limited, if frames of varying priorities are provided the same level of protection, the node would be busy retransmitting low-priority frame packets, and its queue would have insufficient space for high-priority packets, thereby resulting in poor performance. Therefore, this study adjusted the priority level of various packets, and reduced the number of retransmission attempts for low-priority packets when the network is in congestive or highly competitive conditions. The packets are discarded once the maximum retransmission number is attained so that subsequent packets can enter the transmission queue. The probability of successfully transmitting packets is significantly improved, as is the transmission bandwidth and quality of video transmission.

According to the retransmission mechanism algorithm description, the number of packet retransmissions is adjusted based on the priority of various packets. Important I-frames are always assigned four retransmissions, regardless of the wireless network condition; for second-level P-frames, the number of retransmissions is set at two in busy wireless network conditions and at four in slower periods. For low-priority B-frames, the number of retransmissions is set at one in congested wireless network conditions and four in slower periods. Therefore, the proposed retransmission mechanism is activated only when the wireless network is busy to assign more retransmissions to the higher order packets, thereby increasing the possibilities of successful transmission for higher order packets and effectively improving the video transmission quality.

Results and Analysis of the Simulated Experiment

Settings for the Simulated Parameters

In this section, a simulated experiment is conducted using NS2 [11, 12] network simulation software. In the experiment, a multi-hop wireless network environment is simulated, which consists of four wireless network nodes. Suppose that the sequence of these four nodes is AP 1, AP 2, AP 3, and AP 4. Direct data transfer may occur between adjacent nodes, and non-adjacent nodes do not interfere with one another. In other words, if data are to be transferred from AP 1 to AP 4, they must first be transferred AP 2, then to AP 3, and finally to AP 4; if data are to be transferred from AP 1 to AP 3, they must be transferred to AP 2 first, and then transferred from AP 2 to AP 3. Data can be directly transferred from AP 1 to AP 2 because these two nodes are adjacent nodes and capable of direct data transfer. The simulative parameters of the experiment are set as shown in Table 3.

Table 3 Simulated parameter settings

Parameter	Description	Set value
threshold	Threshold of network congestion	25
default_TTL	Default TTL value	32
qmax	Total queue length at the node	50
normal	Normal number of retransmissions	4
less	Fewer number of retransmissions	2
min	Minimum number of retransmissions	1

Results and Analysis of the Simulated Experiment

In the experiment, a data stream was added to AP 1, which will be transferred to AP 4; data streams are also added to AP 2 and AP 3, which act as sources of interference during data transfer. After the data are transferred, the video frames transmitted in this experiment are decoded. By analyzing the peak-to-signal noise ratio (PSNR) of the decoding process, the proposed mechanism can be assessed for its improvement of the transmission quality of MPEG-4 compressed videos.

This simulated experiment generates three sets of data. The first set of data is obtained through the process of video transmission using the existing IEEE 802.11 mechanism; the second set is obtained from the experiment that incorporates the TTL mechanism into the 802.11 mechanism; and the third set of data is obtained from the simulated experiment that incorporates both the TTL and retransmission mechanisms.

Table 4 presents the Foreman video data. A total of 400 frames were extracted from the source video, including 51 I-frames, 83 P-frames, and 266 B-frames. The statistics of the video frames transmitted through the IEEE 802.11 mechanism (Table 5) show that 21 I-frames were received by the receiving end, 30 I-frames were lost; 50 P-frames were received and 33 were lost; 144 B-frames were received and 132 were lost.

The statistics of video frames transmitted by using the TTL mechanism (Table 5) show that after the TTL mechanism was incorporated, 31 I-frames were received by the receiving end, which is 47.6 % higher than the transmission with IEEE 802.11 mechanism. The number of P-frame transmissions increased to 55 frames from the previous 50 frames using IEEE 802.11, which is a 10 % increase. Finally, 168 frames of B-frame packets were received compared to the 134 frames transmitted by using IEEE 802.11, indicating a 25.4 % increase.

The statistics of video frames transmitted using TTL and the retransmission mechanism (Table 5) show that after the TTL and retransmission mechanism was

Table 4 Foreman video data

Video	Format	Number of frames			Total number of frames
Foreman	QCIF	I	P	B	400
		51	83	266	

Table 5 Statistics of video frames transmitted using (a) IEEE802.11, (b) TTL mechanism and (c) TTL and retransmission mechanism

Mechanisms	IEEE802.11			TTL			TTL and retransmission		
Categories	I-frames	P-frames	B-frames	I-frames	P-frames	B-frames	I-frames	P-frames	B-frames
Total transmissions	51	83	266	51	83	266	51	83	266
Total receptions	21	50	134	31	55	168	35	50	154
Number of losses	30	33	132	20	28	98	16	25	112

Table 6 Average PSNR of the foreman video

	IEEE 802.11	TTL mechanism	TTL and retransmission mechanism
Average PSNR	21.292	23.363	23.518

used, 35 I-frames were received by the receiving end, a 66.7 % increase over the transmission using the IEEE 802.11 mechanism; the number of P-frame transmissions increased to 58 from the 50 frames transmitted using IEEE 802.11, showing a 16 % increase; finally, 154 B-frame packets were received. Only 134 frames were received when using IEEE 802.11, indicating a 14.9 % increase.

The results of the average PSNR presented in Table 6 show that the average PSNR obtained by using IEEE 802.11 was 21.29 after the discarded frames were counted and compared to the source data. After incorporating the TTL mechanism, the transmission in the same simulation environment generated an average PSNR of 23.363; a 9.7 % improvement over IEEE 802.11. After the TTL and retransmission mechanism was added, the average PSNR was raised to 23.518 under the same experimental conditions, indicating a 10.5 % improvement over the IEEE 802.11 mechanism.

Overall, the TTL and retransmission mechanism proposed in this study can effectively improve the quality of video transmission.

Conclusion and Research Suggestions

This study proposed two mechanisms for improving the quality of video transmission in multi-hop wireless networks: the queuing mechanism implemented by adjusting packets based on TTL analyses, and adjustment of retransmissions based on the varying priorities of differing packets. The analytical results of a simulated experiment conducted on an NS2 network confirmed that, compared to the traditional IEEE 802.11 mechanism, the two mechanisms proposed in this study can effectively improve the quality of video transmission in multi-hop wireless networks.

In this study, fixed nodes were employed in a multi-hop wireless network. Future studies should simulate actual scenarios occurring in multi-hop wireless networks, such as environments in which multiple nodes randomly move within a certain range, or in which specific nodes remain static but others move.

References

1. Shugong X, Saadawi T (2001) Does the IEEE 802.11 MAC protocol work well in multi-hop wireless ad hoc networks? IEEE Commun Mag 39(6):130–137
2. G. Bianchi (2001) Performance analysis of the IEEE 802.11 distributed coordination function. pp 535–547

3. Hsieh H-Y, Sivakumar R (2002) IEEE 802.11 over multi-hop wireless networks: problems and new perspectives. In: IEEE 56th VTC 2002
4. Hsieh H-Y, Sivakumar R (2001) Improving throughput and fairness in multi-hop wireless networks. In: Proceedings of ICN, Colmar, France, July 2001
5. Nguyen LT, Beuran R, Shinoda Y (2007) Performance analysis of IEEE 802.11 in multi-hop wireless networks. In: Proceedings of the 3rd international conference on Mobile ad-hoc and sensor networks, ser. MSN'07. Springer, Berlin, Heidelberg, pp 326–337
6. Mustapha I, Jiya JD, Monguno K (2010) Throughput analysis of IEEE 802.11 MAC protocol in multi-hop wireless Ad-Hoc network. J Sci Technol Res 9(1):113–122
7. IEEE (1999) Wireless LAN media access control (MAC) and physical layer (PHY) specifications. IEEE Std. 802.11
8. Crow BP, Widjaja I, Kim LG, Sakai PT (1997) IEEE 802.11 wireless local area networks. IEEE Commun Mag 35(9):116–126
9. Lee SH, Yoo C (2010) Contention avoidance with hop based priority in 802.11e multi-hop network. Consumer electronics (ICCE), 2010 digest of technical papers international conference, January 2010, pp 151–152
10. IEEE Std 802.11e-2005 (2005) Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 8: medium access control (MAC) quality of service enhancements
11. The network simulator ns-2 [Online]. Available: http://www.isi.edu/nsnam/ns/
12. Evaluation of video stream quality over IEEE 802.11e EDCF [Online]. Available: http://140.116.72.80/smallko/ns2/ns2.htm

Mapping IDCT of MPEG2 on Coarse-Grained Reconfigurable Array for Matching 1080p Video Decoding

Guoyong Li, Leibo Liu, Shouyi Yin, Changkui Mao and Shaojun Wei

Abstract Coarse-grained reconfigurable array (CGRA) can achieve flexible and highly efficiencies for computing-intensive application such as multimedia, baseband processing and etc. MPEG2 is a popular multimedia algorithm which suits for CGRA. IDCT takes around 29 % of total time for MEPG2 Decoding, which is one of main parts of MPEG2. IDCT belongs to computation-intensive which fits for CGRA. The paper explores the parallelism of IDCT algorithm, mapping it on coarse-grained reconfigurable array. The simulation result shows 693 clock cycles are needed to complete 8 × 8 IDCT on REMUS, the cycles needed is just 36 % of XPP, just 24.7 % of ARM. The method improves performance for MPEG2 decoding. The performance fulfils MPEG2 decoding for 1080p @30 fps streams when employs 200 MHz clock frequency.

Keywords IDCT · MPEG2 · CGRA · Mapping · REMUS

Introduction

Many applications are computing-intensive algorithm. In the traditional way, two approaches are used for implementation: one is using a general-purpose processor that it is flexible but not efficient. The other is using ASIC which is efficient, however, has no flexibility. To achieve flexibility and high efficiencies, coarse-grained reconfigurable array (CGRA) has been raised, like XPP [1], Adres [2], REMUS [3]. This architecture can be used to process wide arrange of application in high efficient. In this architecture, different sets of configuration information

G. Li · L. Liu (✉) · S. Yin · S. Wei
Research Center for Mobile Computing, Tsinghua University, Beijing 100084, China
e-mail: liulb@mail.tsinghua.edu.cn
Institute of Microelectronics, Tsinghua University, Beijing 100084, China

C. MaoInstitute of Microelectronics, Xi'an Jiaotong University, Xi'an, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_62, © Springer Science+Business Media Dordrecht 2014

called context are used to make CGRA to realize different functions. Decoding progress of MPEG2 [4] consists of several sub algorithms, VLD\Inverse Scan\Inverse Quantization\IDCT\MC\AddBlock. They can sort into computing-intensive algorithm and control-intensive algorithm. VLD\Inverse Scan\Inverse Quantization belong to control-intensive. And IDCT\MC\AddBlock belong to computation intensive, these algorithms can be pipelined and parallelized. This character fits for coarse-grained reconfigurable array. Doing these algorithms on coarse-grained reconfigurable processor can obtain high performance. Just as shown below, the whole figure is MPEG2 algorithm, data intensive algorithm (IDCT\MC\AddBlock) are processed on coarse-grained reconfigurable array.

The overall application behavior of MPEG2 is researched [5]. The Figs. 1, 2 shows the relative performance of decoder for a 9 MB/s case. IDCT takes around 29 % of total time for MPEG2, it is one of the main parts of MEPG2. By speeding up this part can improve decoding performance effectively. There are some ways to speed up IDCT. Swamy et al. [6] use ASIC to accelerate IDCT, which executes very fast, but needs a special module and has no flexibility. Fang et al. [7] maps IDCT on GPU. Winger [8] use CPU SIMD execute IDCT, which reduces 22 % execution time. XPP [1] maps IDCT on CGRA. This paper implements sub-algorithm IDCT of MPEG2 on REMUS which is one kind coarse-grained reconfigurable SOC. Section Architecture of REMUS and RPU introduces the architecture of REMUS and RPU. Section Implementation IDCT on REMUS do task partion of MPEG2 on REMUS and analyses IDCT algorithm of MPEG2. Section Result proposes techniques to mapping IDCT on REMUS and does the simulation to get results. Section Conclusion gets the conclusion.

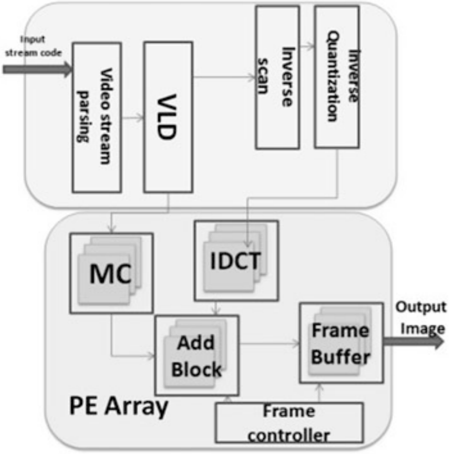


Fig. 1 MPEG2 decoding

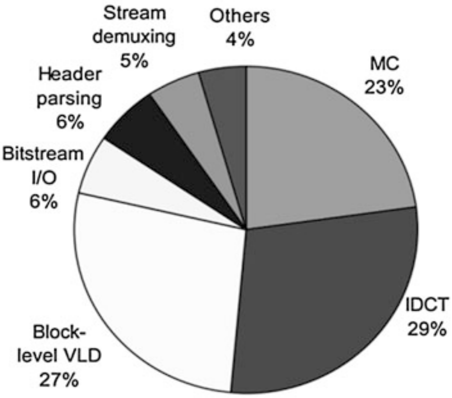


Fig. 2 MPEG2 decoding time on 3.06 GHz Pentium4

Architecture of REMUS and RPU

Architecture of REMUS

REMUS [3] is a reconfigurable Multi-media System, which is one kind of coarse-grained reconfigurable SOC. As shown below, REMUS contains RPU\u-PA\ARM7, Entropy Decoder (EnD), and some other modules. RPU contains four reconfigurable processing element array (PEA). The computing-intensive parts are processed on this part, which can achieve high efficiency and flexibility. UPA generates context for the RPU to realize different functions. The ARM7 is a RISC processor which is used for the application control. The EnD processes the entropy decoding such as Context Adaptive Variable-Length Decoding (CAVLD) and Context Adaptive binary Arithmetic Decoding (CABAD). The different parts are connected by AMBA bus, AHB bus is used to connect high speed modules, APB bus is used to connect low speed modules, dedicated high speed EMI bus is used to connect SSRAM with other modules (Fig. 3).

Architecture of RPU

As shown in Fig. 4, RPU consists of four processing element array (PEA), PEA controller, context interface, data interface, context memory, and data internal memory. The interface is used to connect RPU with outside bus. Memory is used to store context and data. The controller is used to control the progress of RPU. PEA is a powerful dynamic reconfigurable system; it consists of 8 × 8 processing elements (PE). Every PE can do kinds operation, like add, shift, multiply and so on. Context is used to control the function of PE. As shown below, PE consists ALU, MUX, and REG.

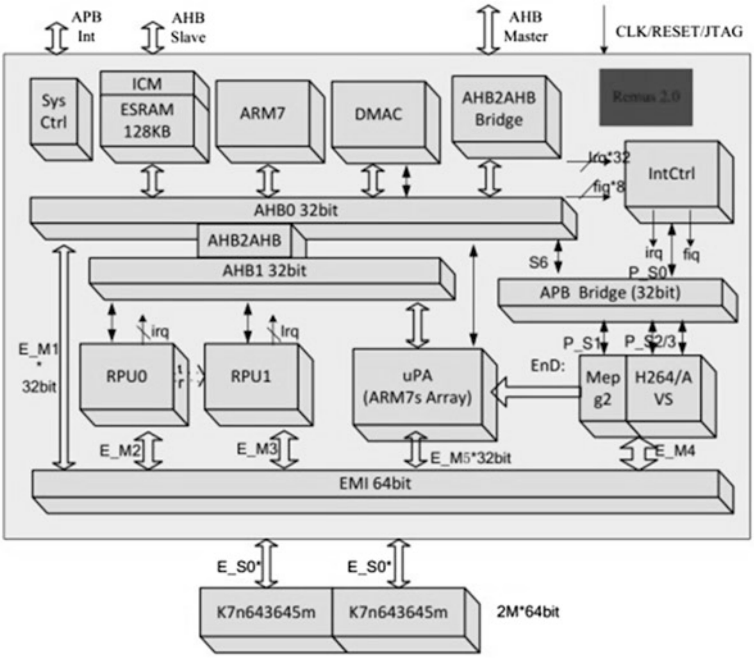


Fig. 3 Architecture of REMUS

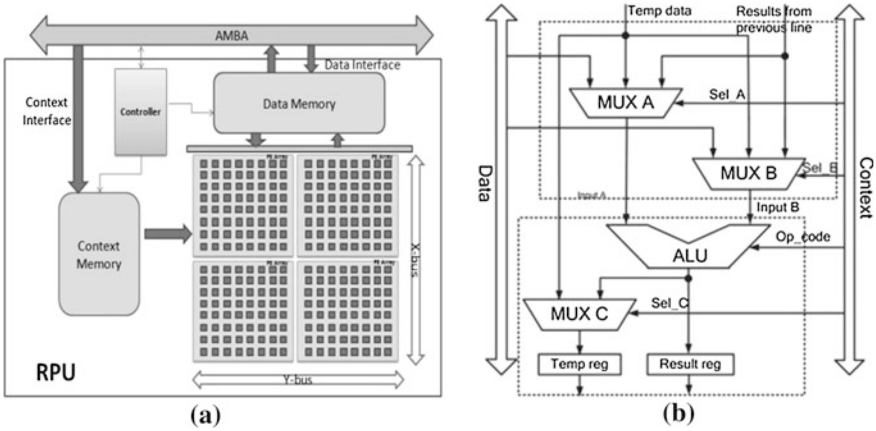


Fig. 4 **a** Architecture of RPU. **b** Architecture of PE

Implementation IDCT on REMUS

Task Partion

Realizing MPEG2 algorithm on REMUS should depend on the characteristic of this algorithm. According to the previous architecture of RPU, RPU contains four PE array. The computation ability of RPU is so strong that it fits for the algorithm which are low data dependency and data can be parallelized and pipelined. So mapping computing-intensive sub algorithm on RPU. Mapping control-intensive sub algorithm on EnD. ARM control the whole progress, context used for RPU is generated by uPA. For details, main ARM responds for the decoder algorithm above slice level (include slice level), setting DMA, setting EnD, frame buffer management, and communication with outside systems. EnD responds for parsing the code stream, inverse scan\inverse quantization and so on, the characteristic of this part is that data dependency is high, belong to control-intensive, it is hard to be parallelism. The uPA responds for parsing command from ARM and MB from EnD, generates contexts for RPU. RPU responds for algorithm of computing-intensive, contain IDCT\MC\AddBlock, the characters of this part are parallel and pipelined.

IDCT Analysis

Discrete cosine transform (DCT) is a Fourier-related transform similar to the discrete Fourier transform (DFT) [9]. IDCT is the inverse algorithm of DCT. It belongs to the orthogonal transformation coded system. IDCT is used to wipe off the spatial redundancy of image data. The procedure is that transform coded image intensity matrix (Time domain signal) to coefficient space (Frequency domain signal), then process the frequency domain signal. IDCT is widely used in image processing, such as JPEG, mpeg2, H.264 and etc. For MPEG2, macro block is separated into four 8 × 8 sub-blocks, for every sub-block does 2-D IDCT. MPEG2 takes 8 × 8 block 2-D IDCT. So we can get the formula below, f(x, y) (x, y = 0, 1…7) is the pixel of the image, F(u, v) (u, v = 0,1..7) is the correspond coefficient of DCT.

$$f(x,y) = \frac{1}{4}\sum_{u=0}^{7}\sum_{v=0}^{7} c(u)c(v)F(u,v)\cos\left(\frac{2x+1}{16}u\pi\right)\cos\left(\frac{2y+1}{16}v\pi\right)$$
$$x,y,u,v = 0,1,\ldots,7 \quad c(u) = c(v) = \begin{cases} \frac{1}{\sqrt{2}} & u = 0,\ v = 0 \\ 1 & others \end{cases} \tag{1}$$

The computation of IDCT is very large. For 8 × 8 block, 2-D IDCT needs 8,192 multiply and 3,584 add. It is hard to directly map 2-D IDCT on PE array.

This paper takes the method of matrix decomposition. 2-D IDCT decomposes to two 1-D IDCT. IDCT can be written in matrix form: $f = C^T FC$, f is the matrix expression of f(x, y), F is the matrix expression of F(u, v).

$$f = C^T FC \Rightarrow Y = C^T F^T, f = C^T Y^T \tag{2}$$

$C^T C = I_N$, C is 8 × 8 matrix composed of the cosine function, C^T is the transposition matrix of C. From above formula, 2-D IDCT separates into two steps, first is $Y = C^T F^T$, second is $f = C^T Y^T$. In this way 2-D IDCT changes into two 1-D IDCT, one is IDCT-ROW ($Y = C^T F^T$), the other is IDCT-COL ($f = C^T Y^T$). The formula of IDCT-ROW is below, IDCT-COL is the same principle:

$$f(x) = \frac{1}{2}\left[\sum_{u=0}^{7} C(u)F(u)\cos\frac{(2x+1)u\pi}{16}\right] \quad C(u) = \begin{cases} \dfrac{1}{\sqrt{2}}, (u = 0) \\ \quad 1, (u > 0) \end{cases}. \tag{3}$$

This paper adopts Chen-Wang [10] fast IDCT. Expressing above formula in matrix form. After some transformation, the formula presents in below, $ci = \cos(\frac{i \cdot \pi}{16}), \quad (1 \le i \le 7)$:

$$P = \begin{bmatrix} c4 & c2 & c4 & c6 \\ c4 & c6 & -c4 & -c2 \\ c4 & -c6 & -c4 & c2 \\ c4 & -c2 & c4 & -c6 \end{bmatrix} \cdot \begin{bmatrix} F0 \\ F2 \\ F4 \\ F6 \end{bmatrix} \tag{4}$$

$$Q = \begin{bmatrix} c1 & c3 & c5 & c7 \\ c3 & -c7 & -c1 & -c5 \\ c5 & -c1 & c7 & c3 \\ c7 & -c5 & c3 & -c1 \end{bmatrix} \cdot \begin{bmatrix} F1 \\ F3 \\ F5 \\ F7 \end{bmatrix} \tag{5}$$

$$\begin{bmatrix} f0 \\ f1 \\ f2 \\ f3 \end{bmatrix} = \frac{1}{2} \cdot P + \frac{1}{2} \cdot Q \tag{6}$$

$$\begin{bmatrix} f7 \\ f6 \\ f5 \\ f4 \end{bmatrix} = \frac{1}{2} \cdot P - \frac{1}{2} \cdot Q \tag{7}$$

Using Butterfly diagram to express calculation process, the Fig. 5 is shown:

Based on above analysis, IDCT algorithm is computing-intensive and the computations can be parallelized and pipelined.

IDCT Mapping

The Fig. 5 shows that IDCT algorithm is computing-intensive, and many computations can be parallelized and pipelined. These characteristics are very fit for processing on PE array. IDCT algorithm processes in 8 × 8 unit, this can implement on PEA. For 4:2:0 colour format, Y component exists 16 × 16, it can be decomposed to 8 × 8 base unit, then four PEA process 16 × 16. U\V component exists 8 × 8, these can be processed using 8 × 8 base unit. Therefore, IDCT of one macro block can be done by PEA. And the mapping is shown in (Fig. 6).

As previously introduced, 2-D IDCT can be decomposed to two 1-D IDCT, the algorithm of IDCT of MPEG2 works in three steps: IDCT-ROW, transposition, IDCT-COL. IDCT-ROW and IDCT-COL are the same principle, so the IDCT-ROW and IDCT-COL can be pipelined. Mapping 8 × 8 IDCT-ROW on the PE array. 8 × 8 IDCT is too complex to map the algorithm on one map, IDCT-ROW is decomposed into three maps. IDCT-COL uses the same three maps. The Fig. 7 maps the operation P (show in previous section, Eq. 4). In the Fig. 7 row[0]\row[2]\row[4]\row[6] are the input F0\F2\F4\F6, w2\w4\w6 are correspond to c2\c4\c6, but done scale. The Figs. 8, 9 maps the operation Q (show in previous section, Eq. 5). The Fig. 10 maps the operation of (P + Q)/(P − Q) (Eq. 6) and

Fig. 5 Chen-Wang fast IDCT

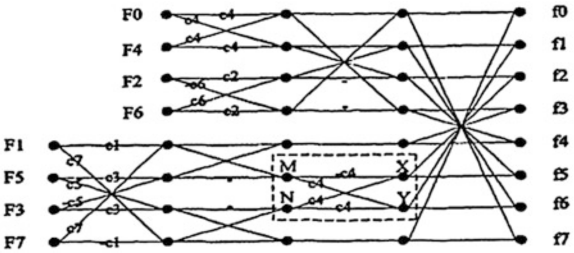
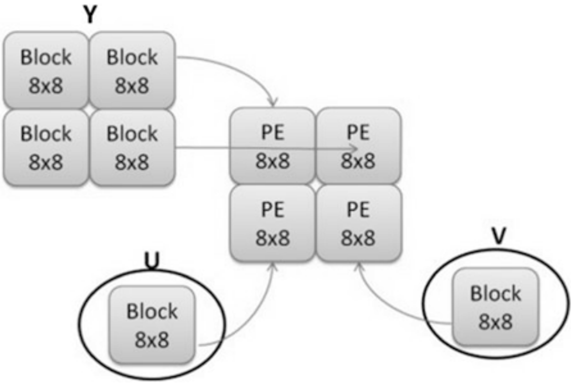


Fig. 6 Mapping 8 × 8 IDCT on PEA



transposition. The Figs. 7, 8, and 9 are mapping for one row. For 8 × 8 IDCT-ROW, 8 rows are needed, doing the operation in 8 loop, then doing maping 4, 8 × 8 IDCT-ROW is done, than continue to do 8 × 8 IDCT-COL in the same way.

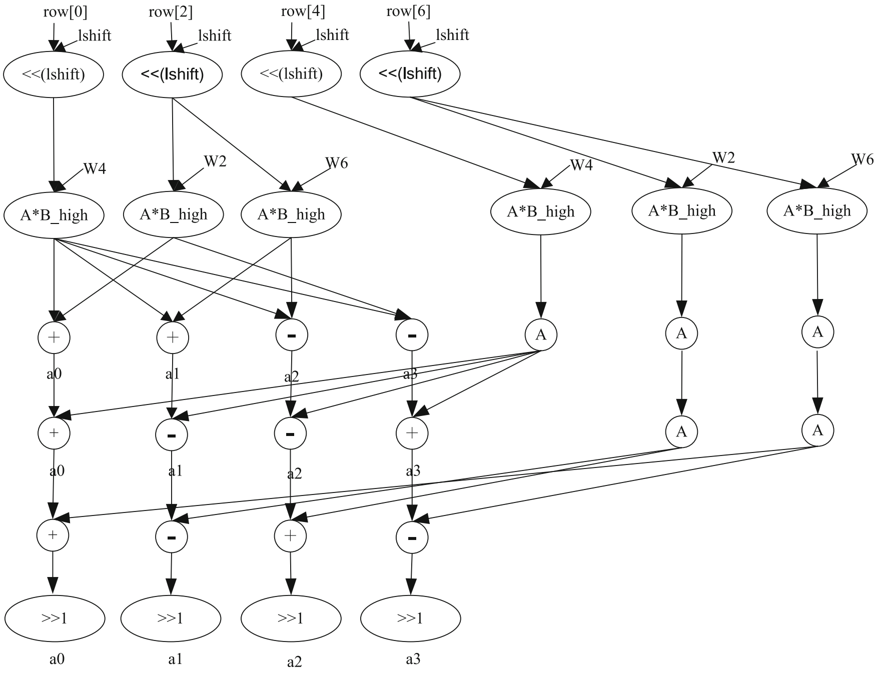


Fig. 7 IDCT mapping 1

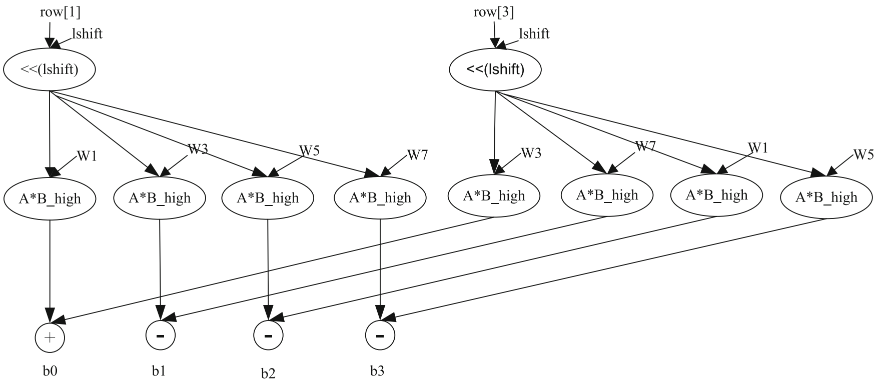


Fig. 8 IDCT mapping 2

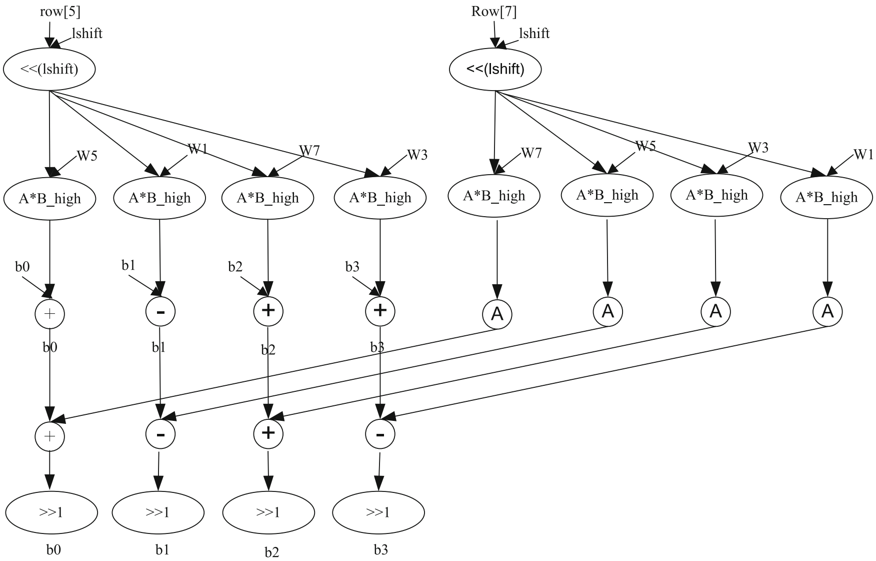


Fig. 9 IDCT mapping 3

Result

The simulation platform is based on REMUS. Implementation of the proposed mapping method for IDCT is verified with RTL simulation using Synopsys VCS. The frequency is set to 200 MHZ. Decades of 1080p MPEG2 code streams is run. Mapping computing-intensive sub algorithm on RPU, uPA generates context used for RPU. Mapping control-intensive sub algorithm on EnD, ARM control the whole progress. The result is shown below. The cycles needed for IDCT on REMUS is 693, just 36 % of XPP and just 24.7 % of ARM (Table 1).

To do more verification, test three 1080p code streams. MC and AddBlock belongs to computing-intensive are also mapping on CGRA. MC takes 589 cycles,

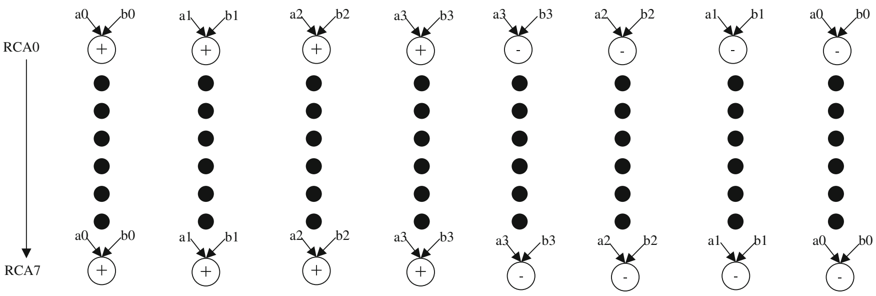


Fig. 10 IDCT mapping 4

Table 1 Cycles of IDCT

IDCT	REMUS	XPP [1]	ARM [11]
Cycles	693	1,700 + 192	2,796

Table 2 Performance of 1080p Decoding

Code streams	VanHelsing	benq2	Japan
Resolution	1,920 × 1,088	1,920 × 1,088	1,920 × 1,088
Performance (ms)	31.23479	32.16729	33.91178

and AddBlock takes 67 cycles. The overall cycles are reduced significantly compared to ARM. As shown in the figure below, the average of performance is around 32 fps. This performance meets the requirements of 1080p real-time decoding (Table 2).

Conclusion

For multimedia algorithm MPEG2, the sub-algorithms include computing-intensive and control-intensive. Mapping data-intensive on reconfigurable processing element array. This scheme can achieve both flexibility and high efficiency at the same time. IDCT is computing-intensive algorithm and it takes around 29 % of total time for MEPG2 decoding. Mapping it on PEA, parallelization of calculation and pipelined structure significantly reduces the execution cycle. The simulation result shows that 8 × 8 IDCT just needs 693 cycles on REMUS. The cycles needed is just 36 % of XPP, just 24.7 % of ARM. For the whole system, REMUS supports MPEG2 1080p @30 fps stream at 200 MHz.

Acknowledgments This work is supported in part by the China National High Technologies Research Program (No. 2012AA012701), the Tsinghua Information S&T National Lab Creative Team Project, the International S&T Cooperation Project of China grant (No. 2012DFA11170), the Tsinghua Indigenous Research Project (No. 20111080997), the Special Scientific Research Funds for Commonweal Section (No. 200903010), the Science and Technology Project of Jiangxi Province (No. 20112BBF60050) and the NNSF of China grant (No. 61274131).

References

1. XPP-III Processor Overview White Paper (2006)
2. Veredas F-J, Scheppler M et al (2005) Custom implementation of the coarse-grained reconfigurable ADRES architecture for multimedia purposes. In: International conference on field programmable logic and applications, 2005
3. Zhu M, Liu L, Yin S et al (2010) A reconfigurable multi-processor SoC for media applications. In: IEEE international symposium on circuits and systems, 2010

4. “MPEG-2 White Paper (2000)
5. Holliman M, YK Chen (2003) MPEG decoding workload characterization. In; Proceedings of workshop on computer architecture evaluation using commercial workloads 2003
6. Swamy R, Khorasani M, Liu Y, Elliott D, Bates S (2005) A fast pipelined implementation of a two-dimensional in verse discrete cosine transform. In: Conference on electrical and computer engineering 2005
7. Fang Bo et al (2005) Techniques for efficient DCT/IDCT implementation on generic GPU. In: IEEE international symposium on circuits and systems 2005
8. Winger LL Source adaptive software 2D iDCT with SIMD. In: IEEE international conference on acoustics, speech, and signal processing 2000
9. Wikipedia [Online]. Available: http://en.wikipedia.org/wiki/Discrete_cosine_transform
10. Rettberg A et al (2001) A fast asynchronous re-configurable architecture for multimedia applications. In: 14th symposium on integrated circuits and systems design 2001
11. Smit LT et al (2007) Implementation of a 2-D 8 × 8 IDCT On the Reconfigurable Montium Core”, International Conference on Field Programmable Logic and Applications, 2007

Green Master Based on MapReduce Cluster

Ming-Zhi Wu, Yu-Chang Lin, Wei-Tsong Lee, Yu-Sun Lin and Fong-Hao Liu

Abstract MapReduce is a kind of distributed computing system, and also many people use it nowadays. In this paper, the Green Master based on MapReduce is proposed to solve the problem between load balance and power saving. There are three mechanism proposed by this paper to improve the MapReduce system efficiency. First, a brand new architecture called Green Master is designed in the system. Second, Benchmark Score is added to each services in the cluster. In the last, an algorithm about how to distinguish the high score service and the low score service, and how to use them effectively.

Keywords MapReduce · Benchmark · Cloud network

M.-Z. Wu (✉) · Y.-C. Lin · W.-T. Lee
Department of Electrical Engineering, Tamkang University, Taipei City, Taiwan, Republic of China
e-mail: dean64188@hotmail.com

Y.-C. Lin
e-mail: bearlaker34@gmail.com

W.-T. Lee
e-mail: wtlee@mail.tku.edu.tw

Y.-S. Lin
Chung-Shan Institute of Science and Technology, Taoyuan, Taiwan, Republic of China
e-mail: amani.amani@mail.tbcnet.net

F.-H. Liu
National Defense University, Taipei City, Taiwan, Republic of China
e-mail: lfh123@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_63, © Springer Science+Business Media Dordrecht 2014

Introduction

The algorithm in this paper will be used to improve the system efficiency based on MapReduce [1] of Hadoop. Hadoop is a kind of open source software that develop from Google MapReduce, and it can will create a cluster that connects each services. The cluster is used to make more computing resources called computing pool, and it can be expanded more and more. In the end, we can decide what we want to get or how to execute the program through coding the Map Function and Reduce Function.

As usual, in order to make the maximum computing resources, the services must keep the high-speed state, but it also has a lot of unnecessary waste. For example, service performance usually are not the same to each other, some of them are very high, but some of them are very low. if we allocate the same amount of work to all service, it must cause a part of service will complete the work early, but it still have to wait other service that performance is poor, and the waiting time means resources wastes. We will talk about how to make the service off if the performance is too low that seriously affects the system performance.

Related Works

Master of MapReduce

Master of MapReduce Master Node is the most important node on MapReduce which cannot be replaced by other nodes. It includes map function, reduce function and mapreduce runtime system. Master node manages receiving command from user and assigning tasks to task trackers, and it stores status of task trackers in database. The status is verified in three different types: Idel, In-processing and completed. The memory address and size of processing data in HDFS (GFS in Google, HDFS in Hadoop) are notified to Master node, and assign map function and task tracker to complete the task (Fig. 1).

Benchmark

Benchmark [2], generally speaking, is a value about something's performance or ability and make comparison. However, a performance comparison of virtualization technology for the moment is not very common, VM Benchmark is a new type of test methods. It is discussed virtual environment build through virtualization and virtual machine management VM resources (hard discs, memory). We have adopted Virtual Machine system build, and we introduce the mechanism of the Benchmark to distinguish the VMs' performance.

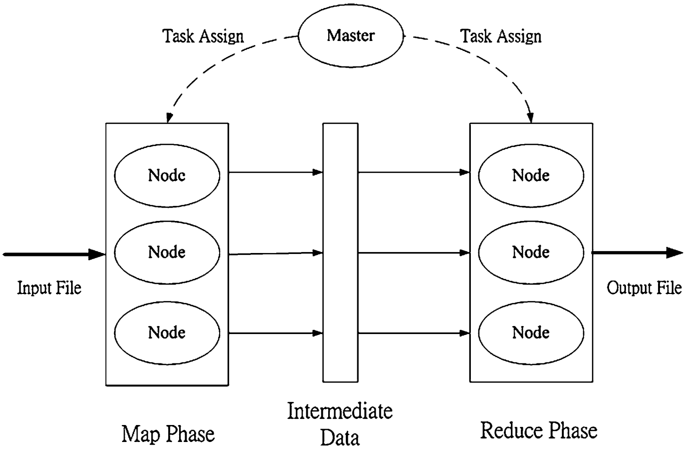


Fig. 1 MapReduce architecture

Implementation

In this section, the algorithm of Green Master will be explained how to implement. It includes Green Master System, Input File Index, Server Information, Queue, Record, Load Balance Optimization, Power Saving Algorithm, and Decision Algorithm. And we will discuss the detail at the following (Fig. 2).

Green Master System

The Green Master is a brand new architecture transformed from Hadoop's Master, and it can apply to each nodes that install the Hadoop. The brand new architecture called Green MapReduce System (GMS), and it can help users manage the node in the cluster to save the system consumption and service computing overhead. The Green Master does not change the Map Function and Reduce Function, it just changes the task allocation master according to server loading and server's Benchmark Score to achieve the goal about the energy saving.

It is not accepted that the system performance reduces caused by someone virtual machine low efficiency, especially in the Cloud Computing Network environment. It is not accepted that the system performance reduces caused by someone virtual machine low efficiency, especially in the Cloud Computing Network environment. In order to solve the above problems, Green Master is designed to delete the poor services and allocate the job distribution. Green Master is divides into eight blocks, and it includes Input File Index, Queue, Server Information, Record, Load Balance Optimization, Power Saving Algorithm and Decision Algorithm. Green Master has a strong adaptability to many systems, for

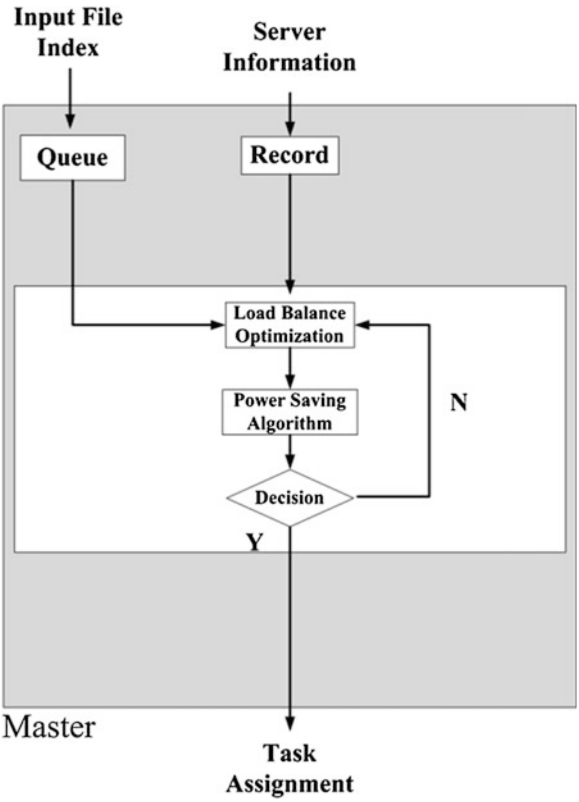


Fig. 2 Green master architecture

an instance, when we need a great amount of computing resources to calculate tasks, we can use Green Master to avoid energy wastes. For another instance, when the system equipment has a strong non-conformance, and the system can use the Benchmark Score in the Green Master to arrange the tasks allocation according to the services capability.

Server Information

The Server Information in the Green Master is to estimate the services' capability called Benchmark Score, and it will keep running and send the results to Green Master. In addition, whenever a new server join or quit the cluster, Benchmark Score will change. The range of the Benchmark Score is from zero to one hundred, and it is according to CPU computing performance, Memory read/write and Disk I/O rate to estimate the Benchmark Score. In other hand, the highest CPU response time, Memory read/write and Disk I/O is defined as 100 Benchmark Score. The definition of poorer virtual machines' Benchmark Score are based on the highest one.

$$BenchmarkScore = \frac{X_n}{Top_{VM}} \times 100\,\% \tag{3.1}$$

where the Top is the highest value of the virtual machine, and the x is the value of the virtual machine like CPU response time to be measured. Because of the CPU response time, Memory read/write and Disk I/O rate have to be considered in the formula, so we turn formulas evolution as follows:

$$Benchmark\,Score = \sum^{i} \frac{X_n}{Top_{VM}} \times W_i,\ W_i \in \{Measured\,Event\},\ i \in \{1, 2, \ldots, m\} \tag{3.2}$$

where the W_i is the event of Benchmark Score. In our case, the i of W_i is three, there are CPU response time, Memory read/write and Disk I/O rate respectively.

Recorder

Recorder is used for recording server information. Recorder refresh when it receives newer server information. A new recording table is established for information record when there is new node joins into the cluster. Servers update and refresh server information in recorder during the working time.

Load Balance Optimization

Load Balance [3] Optimization will allocate the work loading according to the information collecting from the above-mentioned blocks. The Benchmark Score is more higher, and the work loading is more; the Benchmark is lower, and the work loading is less. The job is allocated to VMs through Load Balance Optimization, and the formula is following:

$$Task\,Dsitribution\,Ratio = \frac{Local_{VM}}{\sum_{i=1}^{n} Score_i} \times 100\,\% \tag{3.3}$$

where Total Score is the sum of the VMs' Benchmark Score, and the Local Score is the VM's Benchmark what you want to estimate. In our experiment, we use six VM in the experiment environment and calculate the work loading ratio as following:

Power Saving Algorithm

In this paper, Power Saving Algorithm (PSA) [4, 5] will check the utilization of the server. In the first state, we allocate the work loading to VM according to the Benchmark Score, then the second state, we will determine the utilization of the VM. In Fig. 4, we can find that the huge difference of the work loading between Benchmark Score 100 and Benchmark Score 5, but they use almost same energy. This paper presents PSA to discuss how to get the balance between efficiency and energy management.

$$T_n = \sum_{i=1}^{n} a_i \times \frac{B_i}{\sum_{j=1}^{n} B_j} + \varepsilon \tag{3.4}$$

$$E_n = T_n \times N_{VM} \times P_{VM},\ N_{VM} \in \{1, 2, \ldots, n\} \tag{3.5}$$

$$V_n = \frac{E_{n+1} - E_n}{T_{n+1} - T_n},\ \mathrm{n} \geq 2 \tag{3.6}$$

where T_n is system computing time, and a_i is the system time which one virtual machine completed alone, and the B_i is the Benchmark Score of one virtual machine, and the ε is the error time. E_n is the energy (J) of virtual machine. P is the power (W) of virtual machine. V_n is the ratio of energy consumption.

Decision Algorithm

Decision Algorithm [6] in GMS is to judge the result which is from PSA reasonable or not. The formula is as following:

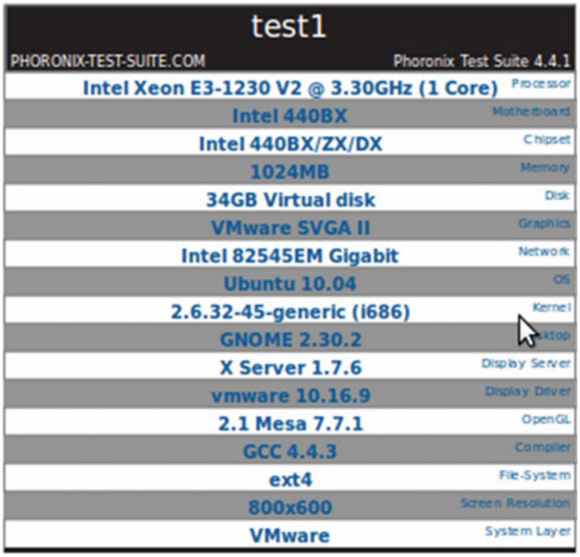
$$\alpha \leq \gamma \tag{3.7}$$

where α is the system consumption through PSA, and γ is without PSA. If γ is greater than α, then the system will back to Load Balance Optimization (Fig. 3).

Simulation Result

Figure 3 shows the highest performance virtual machine in the experiment environment of this paper.

Fig. 3 Experiment environment

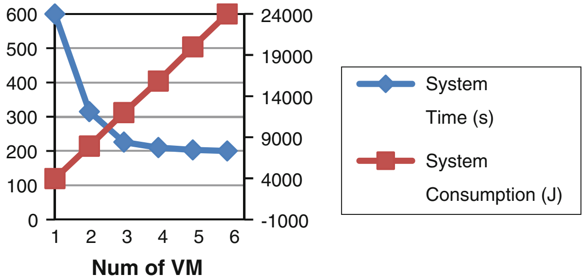


test1

PHORONIX-TEST-SUITE.COM Phoronix Test Suite 4.4.1

Intel Xeon E3-1230 V2 @ 3.30GHz (1 Core)	Processor
Intel 440BX	Motherboard
Intel 440BX/ZX/DX	Chipset
1024MB	Memory
34GB Virtual disk	Disk
VMware SVGA II	Graphics
Intel 82545EM Gigabit	Network
Ubuntu 10.04	OS
2.6.32-45-generic (i686)	Kernel
GNOME 2.30.2	ktop
X Server 1.7.6	Display Server
vmware 10.16.9	Display Driver
2.1 Mesa 7.7.1	OpenGL
GCC 4.4.3	Compiler
ext4	File-System
800x600	Screen Resolution
VMware	System Layer

Fig. 4 System time and consumption



The Relationship Between System Computing Time and System Consumption

In Fig. 4, we can find that the cross point between the system time and system consumption is between two VMs and three VMs. In fact, the number of VM of the best performance in our experiment is three VMs.

Comparison Between Original and Green Master

In Figs. 5, 6 and 7, we take several different sizes of test file in our experiment environment, we can clearly find the original system time is less than Green Master, but system consumption is almost twice larger than Green Master.

Fig. 5 System computing time

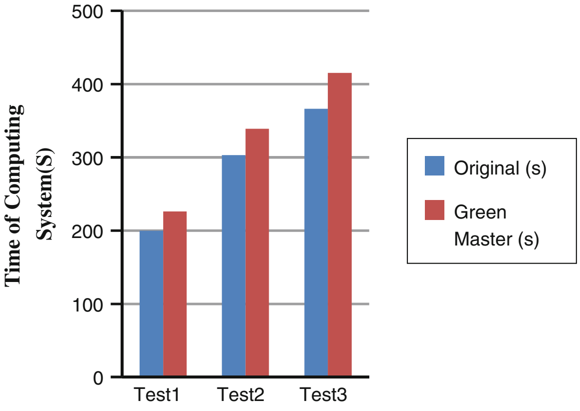


Fig. 6 Power consumption saving

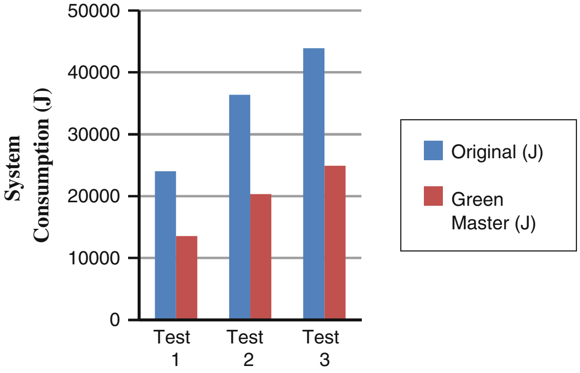
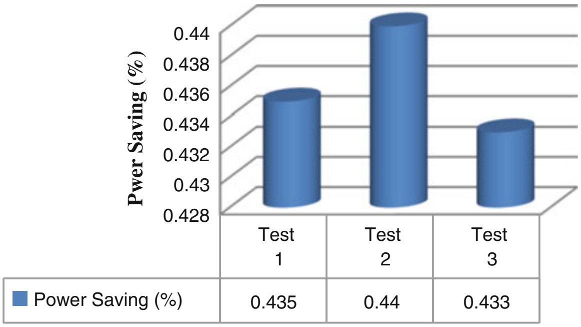


Fig. 7 Ratio of power saving



Conclusion

The idea of Green Master optimizes system power consumption by lower the performance slightly. In this paper, we provide a appropriate trade-off between power saving and performance loses, and improves energy conservation of the system.

References

1. Dean J, Ghemawat S (2004) MapReduce: simplified data processing on large clusters OSDO 2004
2. Xie X, Chen Q, Cao W, Yuan P, Jin H (2010) Benchmark object for virtual machines, 2010 second international workshop on education technology and computer science
3. Kolb L, Thor A, Rahm E (2012) Load balancing for mapreduce-based entity resolution, 2012 IEEE 28th international conference on data engineering
4. Zhu T, Shu C, Yu H (2011) Green scheduling: a scheduling policy for improving the energy efficiency of fair scheduler, 2011 12th international conference on parallel and distributed computing, applications and technologies
5. Sandhya SV, Sanjay HA, Netravathi SJ, Sowmyashree MV, Yogeshwari RN (2009) Fault–tolerant master-workers framework for mapreduce applications, 2009 international conference on advances in recent technologies in communication and computing
6. Liu C, Zhou S (2011) Local and global optimization of mapreduce program model, 2011 IEEE world congress on services
7. He Q, Li Z, Zhang X (2010) Study on cloud storage system based on distributed storage systems, computational and information sciences (ICCIS), 17–19 Dec 2010, pp 1332–1335
8. Bowers KD, Juels A, Oprea A (2009) HAIL: a high availability and integrity layer for cloud storage, computer and communications security (*CCS*), Nov 2009, pp 187–198
9. Luo M, Liu G (2010) Distributed log information processing with Map-Reduce: a case study from raw data to final models, information theory and information security (*ICITIS*), 17–19 Dec 2010, pp 1143–1146

Research on Graphic Digital Text Watermarking Research Framework

Jin Zhang, Xiaowei Liu, Xiaoli Gong, Rui Lu and Zhenlu Chen

Abstract This paper proposes a research framework for graphic digital text Watermarking algorithms and methods. We summarized graphic text watermarking after a brief review of digital text watermarking, and propose the research framework which consists of 8 levels, pixel, line, etymon, character, row, paragraph, page, and chapter. And we give the details of pixel level details and review most graphic text watermarking of this level as the example. At last we also figure out distribution of nowadays algorithms and discuss further research's possibility.

Keywords Text watermarking · Watermarking conceptual framework · Graphic digital text watermarking

Introduction

Digital Text watermarking technology is usually supposed to solve copyright protection and content authentication problems effectively [1]. For some reasons, it does not apply to practice systematically. However, protection measures of digital

J. Zhang (✉) · X. Liu · X. Gong
College of Information Technical Science, Nankai University, 300071 Tianjin, China
e-mail: Nkzhangjin@nankai.edu.cn

X. Liu
e-mail: Xiaoweiliu@nankai.edu.cn

X. Gong
e-mail: Gongxiaoli@nankai.edu.cn

R. Lu · Z. Chen
Tianjin Institue of Software Engineering, 300387 Tianjin, China
e-mail: ruilu@nankai.edu.cn

Z. Chen
e-mail: Chenzl@tjise.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_64, © Springer Science+Business Media Dordrecht 2014

text content can be seen everywhere. Those are similar to digital text watermarking, such as printing a logo on the background of text content, or adding additional and invisible information after the end of segments. In principle, these both belong to the category of digital text watermarking technology. Reference to digital watermarking [2], digital text watermarking is defined as follow:

Digital text watermarking technique, which embeds some sign information of the text into the content or format of digital text production without affecting originality's value or use, is an available protection technique to authenticate the ownership and the integrity of text content. It can't be perceived by people unless extracted by a dedicated detector or reader. Generally, watermark information was encoded with "0" and "1" respectively.

Generally, digital content is composed of words, symbols, graphics, images and other elements according to some certain rules. And digital text or document is the digital content packaged by format. Content is mainly used for ideographic expression and format is used to organize the content to display text content and significance. Except for visible appearance, Format also include invisible edit control part of the background, while similar with format, in addition to visible text, images and other elements, content also include the significance which is non-intuitive visualization and needs to understand and learn. The difference is, even if the format is not visible in the control section, it is also severed to visible format appearance and therefore it is indirectly visible; while the meaning of the text is implicitly featured by elements such as text, image and format, it belongs to the comprehension visible. Like shaped and righteousness in the linguistics, text can also be understood as two parts of the text appearance (shape) and text meaning (sense).

Therefore there are four kinds of digital text watermarking technologies: Document Format Text Watermarking [3], Image Text Watermarking [4], Feature Coding Watermarking [5] and Natural Language Watermarking based on Semantics [6].

Document Format Text Watermarking embeds the watermark by changing the text format slightly, such as line and word shifting. Image Text Watermarking treats the digital document as an image. For example, embedding the watermark through the conversion of Spatial-Temporal Domain and Transform Domain, like image watermarking regardless of the feature of text, and changing the pixel on the unobtrusive part of text. The feature coding is more complex by its special relevance, such as Chinese and Hindi. The Natural Language Watermarking embeds the watermark based on the imperceptibility of semantics. This kind of watermark is vision visible, but it does not affect the understanding of the text, thus it is difficult to perceive. The former three kinds of watermarking belong to the text graphical watermarking based on the text appearance. They embed the watermarks by changing the text appearance slightly, which do not have a large visual deformation. However, the last one belongs to the text semantic watermarking, which changes the text content slightly in the case of retaining the original meaning.

Graphic Text Watermarking Framework

Framework Description

Text graphical watermarking is classified further considering with the position and method of watermark embedded. Because of the restrictions of language familiarity, the new classification introduced is based on simplified Chinese, which is adjusted to other languages look like Chinese. As for English, Arabic and others, the classification needs to be modified partially according to the characteristics.

As Fig. 1 shows graphic watermarking is divided into eight levels by the embedding position: pixel, line, etymon, character, row, paragraph, page, and chapter. Every level is composed of a certain number of the upper level units according to a certain rule. Wherein, the pixel is the basic unit of the text graphics. Any text consists of a myriad of pixels. The chapter is the highest level of the text, which is the first sight of a digital document and is associated with a variety of text formats with the content inside. There are different embedding methods for each level, such as, changing the gray value (changing the self-specificity); replacing the font with another similar one (similar unit replacement); adjusting the gap between etymon of one word (adjusting the combination rule). Three typical characteristics of each level are showed from the above methods: self-characteristics, similarity and structural feature.

Overall, the self-specificity includes inherited, including and unique attributes. As for the inherited attribute, the attribute of the next higher level is inherited; as for the including attribute, the structural feature of one level is the internal structural feature of the higher level because of the layered structure; the unique attribute is comprised of its self-specificity. In general, the similarity of graphical text

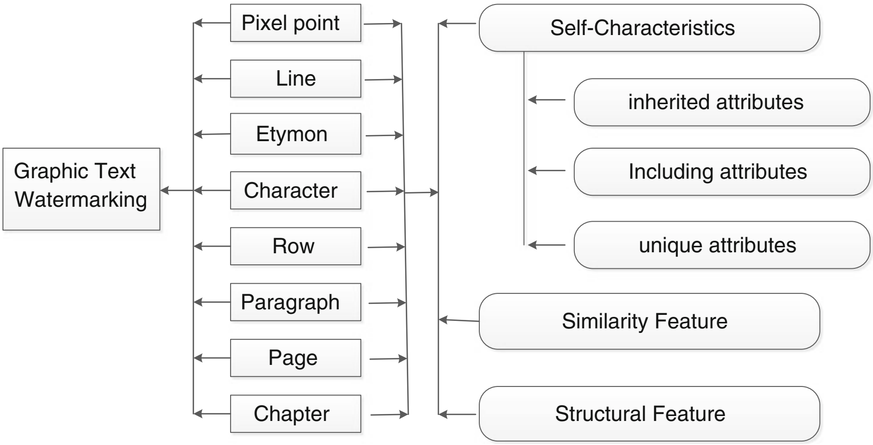


Fig. 1 This shows a graphic text watermarking framework figure consisting 8 embedding position levels

watermark is that an object is replaced by another similar one by all means on the premise of undetectable or imperceptible change. The common similarity replacement includes the similarity of fonts, languages, special symbols and error combinations. The structural feature is a structure which is constituted of different levels, and each level is constituted of a certain number of lower levels by certain rules. This feature is mainly reflected in two aspects: the number of rules. Apart from the pixel level, the amount and elements of other levels are fixed, as well as the theoretical rules (or the new unit will be created). We can only change the rules slightly, such as slight twist, tilt and stretching.

Framework Details

According to the definition of level, there are inclusion relations among the levels of graphical text watermark, that is to say, each chapter contains pages, and each page contains paragraphs, and so on. In other words, the pixel is the basic unit, which constitutes a line, then constitutes a etymon by some certain rules. The roots compose characters, which are put into a row. Many rows constitute a paragraph, and paragraphs build a page. Finally, all the pages are combined into a chapter with layout information, which is the common text we can see. All the levels are progressive and contained. As shown in Fig. 2.

Similarly, the inheritance between attributes and the transfer from characteristics to attributes embody the relationships among all the levels, that is, inheriting and inherited, transfer and inclusion. As the progressive inclusion relationship among levels, the characteristics and attributes is relevant closely, as well as that between the attributes. There are inclusion relationships among levels, which is the structural feature of each level. Therefore, all levels are linked closely together by the structural feature, and the structural feature is communicable, that is, the structural feature of one level can be the self-characteristic of the lower level. That is what called the transfer and inclusion relationship from characteristic to attribute, while some attributes are inheritable. To take the gray value as an example,

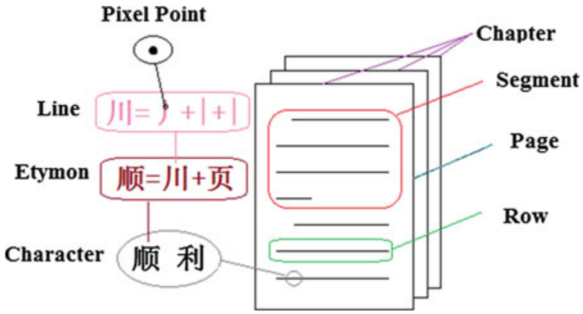


Fig. 2 This shows an example for graphic text watermarking framework embedding position

the gray value of all the layers are inherited from the pixel layer, which is inheriting and inherited relationship.

The inclusion and inheritance relations are different. The inclusion relation comes from the level definition. The higher level is formed by the structure of the lower level, which leads to a result that the internal attribute is the structural feature of the lower one. The point is combination method. However, the inheritance relation lays particular emphasis on the self-specificity itself. Some attributes can be inherited from the lower level, as the self-specificity of the higher one.

Pixel Point Level Study

Pixel Point Level Description

For the limitation of this paper, we only discuss the lowest and basic level-pixel point level, as shown in Fig. 3.

For the operation of the pixel gray value is divided into two aspects, one is to follow the image watermarking technology in the time-space domain and transform domain conversion algorithm; another more consideration text feature, combined with text characters, etymon and some of higher level characteristics or features, the use of certain rules relating to select the grayscale adjustment

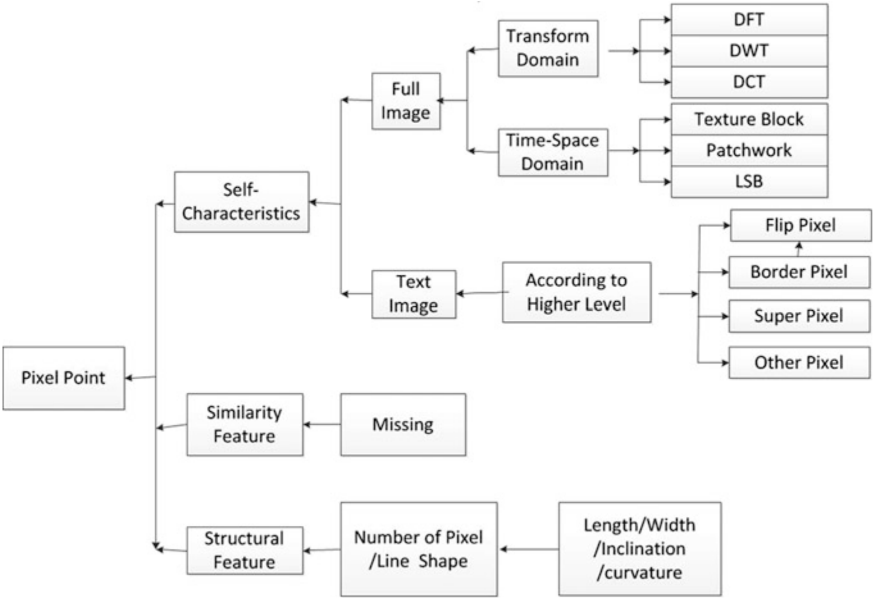


Fig. 3 This shows how pixel point level present under graphic text watermarking framework

imperceptible pixels flip. Time-space domain, transform domain watermark embedding technology from the image watermark, the current development is relatively mature, remarkable achievements in the study of image watermarking technique. Our work focuses on more text watermarking than image watermarking. The pixels constitute the basic unit of the text apart from the slightly itself level changes.

In addition, a different number of pixels according to a certain permutation or combination rules constitute the basic unit of the character—line, which forms the pixel level between the number (which determines the line length, width) and rules (which determines the inclination or curvature) structural characteristics.

Pixel Point Level Algorithms Review

Except inherited some of the traditional image watermarking algorithms, text graphical watermarking on pixel point level has also made a lot of results with considering text feature. Wu et al. [7, 8] introduced an algorithm that embedding the information by flipping the selected pixel called “flipped pixel”. The selected pixels were chose with a 3 × 3 window on the text block. So the eight neighbor pixels around the selected ones will be checked to keep connectivity and gliding property after flipped. There are some similar algorithms, for example, Alex in [9] use 3 × 3 window to check 5 × 5 neighboring points and minimum flip distance to ensure the quality of the text appearance; Mei in [10] proposed to use the eight communicating boundary in a connection region of character to embed watermarking information. Gou and Wu [11] proposed a concept of “super pixel”, which embedding the information through thickening or thinning a certain part of characters. It changes the structural feature of pixel point level.

As to modify the pixel values will cause distortion more or less. In order to enhance the imperceptibility of the embedded information, Anthony introduced an concept of Curvature Weighted Distance Difference, which picks flipped pixel by measuring the distortion after flip the pixel [12]. On this basis, Khen and Makur [13] take further study for improvement of coding efficiency.

Besides flipped pixel and super pixel, [14] proposed to add or delete a group of pixels in some characters, and the group of pixels relative to the character is central symmetry. In that paper, in order to reduce the loss of information by photocopying, it also introduced a packet-based synchronization method, thus greatly reducing the false detection rate.

The edge pixels are also used to hide information usually. Tirandaz et al. [15] embedded the secret watermark information into the edge pixels which will cause the minimum distortion, analyzes the edge pixels of character and defines edge to edge pixel distance. In addition, there are some algorithms with modifying the edge region of characters [16, 17]. Zhang et al. [18] modified the gray value of pixels with another rule. First, it selected words for information embedded with area of the words. Then, it checked the pixel in these words to get pixels whose

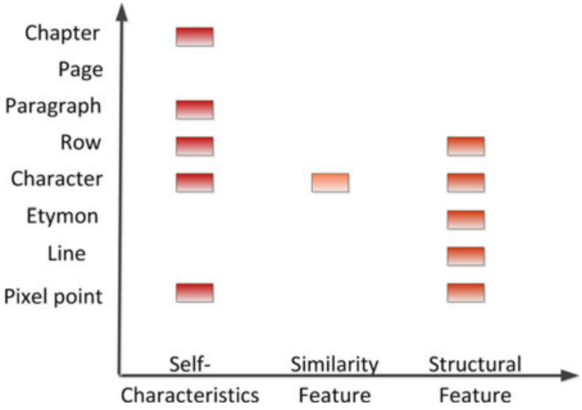


Fig. 4 This shows nowadays graphic digital text watermarking algorithms' distribution

gray value is closest to the average gray value of its eight neighboring pixels. Zhaoxing Yang proposed to adjust the stair edges of non-horizontal strokes and non-vertical strokes in characters in order to change the ratio of the sum of black pixels in the upper and lower halves of each character image for the purpose of embedding a watermark bit [19].

Finally, Qi in [20] proposed a method for binary text image digital watermarking algorithm. Under the assumption of multiplicative transformation model, this method flips boundary pixels of character image with a boundary point flip policy so that the watermarking embedded in the image is difficult to be modified traces. On this basis, [21] takes further study and makes the complexity of the characters as the standard to select embedded characters. Based on invariant after print or scan, it flips pixel with a quantization function to embed a number bits of information. It reduces visual distortion caused by the flip pixels effectively.

Conclusion

Under the framework in Graphic Text Watermarking Framework, we found nowadays graphic digital text watermarking algorithms' distribution map as Fig. 4 shows.

Most researchers is focus on self-characters and structural feature, while similarity is ignored. And the blank zone in Fig. 4 maybe the aspects for further research.

In this paper we propose a research framework for Graphic Digital Text Watermarking algorithms and methods. It consists of 8 levels, pixel, line, etymon, character, row, paragraph, page, and chapter; and 3 aspects, self-characteristics, similarity and structural feature. We hope it could help researchers find new research interests and ideas. In the future, we will try to build the relationship between this framework and application scene.

Acknowledgments This work is supported by a grant from the Research Fund for the Doctoral Program of Higher Education of China (Granted No.20110031120032), the Fundamental Research Funds for the Central Universities and the National Science Foundation of China (Grant No: 61103074).

References

1. Brassil JT, Low S, Maxemchuk NF (1999) Copyright protection for the electronic distribution of text documents. In: Proceedings of the IEEE. 87, IEEE Press, New York, pp 1181–1196
2. Cox I, Miller M, Bloom J et al (2007) Digital watermarking and steganography. Morgan Kaufmann Publishers In; 2nd revised edition
3. Maxemchuk NF (1994) Electronic document distribution. AT&T Tech J 6:73–80
4. Kim Y, Oh I (2004) Watermarking text document images using edge direction histograms. Pattern Recogn Lett 11:25
5. Shirali-Shahreza MH, Shirali-Shahreza M (2006) A new approach to persian/arabic text steganography, computer and information science. In: 5th IEEE/ACIS international conference, IEEE Press, New York, pp 310–315
6. Topkara U, Topkara M, Atallah MJ (2006) The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In: ACM multimedia and security conference, ACM Press, New York, Geneva
7. Wu M, Tang E, Liu B (2000) Data hiding in digital binary images. In: international conference on multimedia and expositions, IEEE Press, New York, pp 393–396
8. Wu M, Liu B (2004) Data hiding in binary image for authentication and annotation. Multimedia IEEE Trans 6(4):528–538
9. Yang H, Kot AC, Liu J (2005) Semi-fragile watermarking for text document images authentication. In: 2005 IEEE international symposium on circuits and systems, vol 4, IEEE Press, New York, pp 4002–4005
10. Mei Q, Wong EK, Memon ND (2001) Data hiding in binary text documents. In: 2001 SPIE 4314, IEEE Press, New York, pp 369–375
11. Gou H, Wu M (2007) Improving embedding payload in binary images with "Super-Pixels". In: 2007 IEEE international conference on image processing, vol 3, IEEE Press, New York, pp III-277–III-280
12. Ho ATS, Puhan NB, Makur A, Marziliano P, Guan YL (2004) Imperceptible data embedding in sharply-contrasted binary images. In: 2004 international conference on control, automation, robotics and vision, vol 2, IEEE Press, New York
13. Khen TV, Makur A (2006) A word based self-embedding scheme for document watermark. In: TENCON 2006. 2006 IEEE region 10 conference, IEEE Press, New York, pp 1–4
14. Varna AL, Rane S, Vetro A (2009) Data hiding in hard-copy text documents robust to print, scan and photocopy operations. In: 2009 IEEE international conference on acoustics, speech and signal processing, IEEE Press, New York, pp 1397–1400
15. Tirandaz H, Davarzani R, Monemizadeh M, Haddadnia J (2009) Invisible and high capacity data hiding in binary text images based on use of edge pixels. In: 2009 international conference on signal processing systems, IEEE Press, New York, pp 130–134
16. Yu X, Wang A (2009) Chain coding based data hiding in binary images. In: fifth international conference on intelligent information hiding and multimedia signal processing, IEEE Press, New York, pp 933–936
17. Zhang X, Liu F, Jiao L (2003) A new effective document water-marking technique based on outside edges. Syst Eng Electron 25(05):612–616
18. Zhang X, Liu F, Jiao L (2003) An effective document watermarking technique. J Commun 24(05):21–28

19. Zhao X, Sun J, Li L (2008) Watermarking of text images using character step edge adjustment. J Comput Appl 28(12):3175–3182
20. Qi W, Li X, Yang B, Cheng D (2008) Document watermark scheme for information tracking. J Commun 29(10):183–190
21. Guo C, Xu G, Niu X, Li Y (2011) High-capacity text watermarking resistive to print-scan process. J Appl Sci 29(02):140–146

A Predictive Method for Workload Forecasting in the Cloud Environment

Yao-Chung Chang, Ruay-Shiung Chang and Feng-Wei Chuang

Abstract Cloud computing provides powerful computing capabilities, and supplies users with a flexible pay mechanism, which makes the cloud more convenient. People are getting more and more usage of the cloud environment due to a steady increase of data. In order to improve the performance and energy saving of the cloud computing, the efficiency of resource allocation has become an important issue. In this study, a neural network model with learning algorithm is applied to predict the workload of the cloud server. The resource manager deployed on the cloud server provides the service of managing the jobs with a resource allocation algorithm. With this prediction mechanism, cloud service providers can forecast the following time workload of cloud servers in advance. The experimental results show that resources can be allocated efficiently and become load balanced by proposed mechanism. Therefore, the cloud server can avoid the problem of inadequate resources.

Keywords Cloud computing · Predictive workload · Neural network · Learning algorithm

Y.-C. Chang (✉)
Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan, Republic of China
e-mail: ycc@nttu.edu.tw

R.-S. Chang · F.-W. Chuang
Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan, Republic of China
e-mail: rschang@mail.ndhu.edu.tw

F.-W. Chuang
e-mail: m9921043@ems.ndhu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_65, © Springer Science+Business Media Dordrecht 2014

Introduction

Numbers of physical and virtual resources provide a powerful computing ability in the cloud computing environment. When a user requests to execute the particular application, the resource manager of cloud computing environment will create a new virtual machine, and assign jobs to the host which operating under servers. Virtual resources can be added or removed at any time in the cloud computing environment. This characteristic makes the cloud platform more flexible and efficient. However, this feature also brings some issues, such as resource allocation imbalance and energy consumption in equable. On the other hand, the workload of the cloud environment may change frequently in a short period due to the large number of jobs or when only a small amount of new jobs is received from users. Furthermore, when the cloud server receives a large number of new jobs, it also results in the same situation in a short period of time. This varying workload may result a situation where there are not enough resources to be allocated or cause a resource imbalance. Therefore, a prediction method for cloud environments to allow the cloud provider to avoid this situation is needed. In this study, a prediction method to forecasting the workload of cloud environments based on a neural network is proposed. The workload information from the cloud server was collected and calculated with the prediction method. The cloud provider can provide more efficiency management of the resource allocation by this method.

The rest of this study is organized as follows: section Related Works introduces an overview and related works of prediction methods. Section Proposed Method presents the proposed method and learning algorithm to forecast the workload of cloud environments. Data collection and implementation results of proposed method are sent out and analyzed in section Implementation. Finally, section Conclusions concludes the paper.

Related Works

Workload characterization and prediction have been studied for many years. Many algorithms and proposed models have a somewhat different focus. Some algorithms are focused on virtual machine placement, which is related to the workload of the virtual machine and virtual machine allocation. While other algorithms are focused on energy consumption which seek to allocate a maximum number of resources and attempt to ensure that all the Service Level Agreements (SLA) are satisfied. Finally, some algorithms focus on the future workload of the cloud server which is also this study's focus. The following section will introduce the different algorithms for workload prediction.

Time Delay Neural Network (TDNN)

The time delay neural network [1] is a feed forward neural network which uses time delay information combined with the input layer. At any given time n, the input of the network consists of the vector:

$$y(n) = \sum\nolimits_{j=1}^{m} w_j \phi\left(\sum\nolimits_{i=0}^{p} w_j(i) \times (n-1) + b_j\right) + b_0 \tag{1}$$

The output of the TDNN is the predicted next value in the time series which is computed as the function of the past values in the time series. To achieve this predictive method, the inputs need a large number of continuous data. It is not appropriate to predict the workload on the new server.

Regression Model

Regression analysis [2] is used for estimating the relationships between a dependent variable and one or more independent variables. Regression analysis helps to understand how the typical value of the dependent variable changes when each of the independent variables is varied, while the other independent variables are held fixed. A general form of the non-linear model is provided as:

$$y_i = a_0 + a_i x_i + a_i x_i^2 \cdots + a_i x_i^m + \delta_i \quad (i\ = 1, 2, \ldots, n) \tag{2}$$

where 'a' is constant and represents successive numeric quantities. δ represents the random prediction error in the model. The value of m depends on the degree of freedom of the model to be deduced.

Predictive Data Grouping and Placement

The method [3] is focused on the load balance problem on the server. The strategy is to group the data which have a similar workload, and then determine how data is placed in different servers according to capacity and the amount of category thresholds. The grouping method is to place related and popular contents on the same server. The placement method consolidates any pair of under loaded servers, whose aggregate workload remains under the maximum desired load and reassigns the dispatching probabilities.

Statistic Based Load Balance

The SLB [4] is to solve load imbalance and guide the host selection while the virtual machine starts. SLB uses historical performance data of virtual machines to estimate the resource requirement of each virtual machine. The virtual machine load forecast method is defined as the formula:

$$\mathrm{L_i}(\mathrm{VM_{cpu}}) = \mathrm{E}(\mathrm{l_i}(\mathrm{VM_{cpu}})) = \frac{\sum_{j=1}^{n} \mathrm{l_{ij}}(\mathrm{VM_{cpu}})}{\mathrm{n}} \tag{3}$$

where $\mathrm{L_i}$ is the predicted load at time *i* in one day, $\mathrm{l_i}$ is the load value at time *i*, $\mathrm{l_{ij}}$ is the load value at time *i* in the *j*th day.

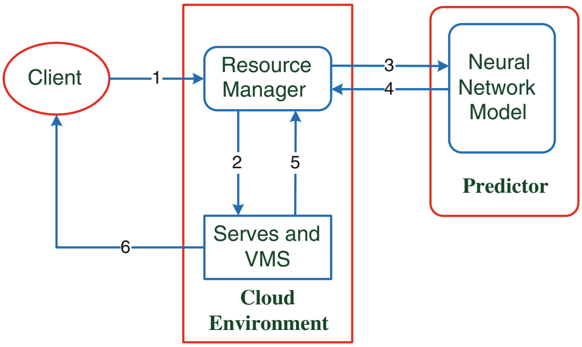
Proposed Method

The workload definition is the amount of processing slot that the computer has been given to do in the cloud environment. The workload consists of the amount of applications running in the computer and the number of users connected to or interacting with the applications. This study focuses on the amount of applications running in the computer. Besides, a recurrent neural network model is applied in the proposed system based on the prediction of total workload for the cloud severs. The workload data is collected from the NCHC Hadoop public experiment cluster.

The proposed system architecture consists of a client, resource master, predictor, servers and VMs which are shown in Fig. 1. The definition of resource manager is the service of managing the jobs with a resource allocation algorithm. The operation consists of the following steps:

- **Client submits a job**. The job in the cloud environment will split into many processes. The workload of the cloud environment is composed of these processes.

Fig. 1 System architecture



- **Choose available servers**. The resource manager will select the servers and create numbers of VMS based on the requirements from the job.
- **Send workload information to Predictor**. When the resource manager dispatches the processes to the servers and VMs, the predictor can get the workload information from the resource manager.
- **Send results to Resource Manager**. The predictor uses previous workload information to calculate the workload information for the following time. When the predictor has calculated the result of the next time's workload information, it will send the result to the resource manager. The result will help the resource manager to modify the resource allocation algorithm. The allocation algorithm will be suitable for the cloud environment.
- **Send resource information to resource manager**. When the servers or VMs have completed a job, it will send back the resource information and workload information.

The prediction of recurrent neural network model is illustrated in this work. The neural network consists of three layers with an input layer, one output layer and one hidden layer. In the input layer, the neural network receives the information of workloads from X_1 to X_a at various time steps of one complete sequence that constitutes the first epoch. At each time step the output is feedback to be employed as the input *k* for the next time step. At any given time *t*, the input of the neural network composed of the vector is shown as:

$$\mathrm{x(t) = [x(t), x(t\ -1), \ldots, x(t\ -}p\mathrm{)]^T} \tag{4}$$

where *p* is the number of selected delay line memory. The vector means the neural network can remember the numbers of workloads. For example, people usually can remember what they have already done recently. The vector is like the memory of people. For this vector, it works to select the number of *p* and decides how long the neural network can remember the workloads in the past.

For a neural network with one hidden layer, the output of a single neuron *j* in the hidden layer is given by the equation:

$$\mathrm{y_j(t)} = f\Big(\textstyle\sum_{\mathrm{i \in A \cup B}} \mathrm{w_{ji}(t\ -1)k_i(t-1) + b_j} \tag{5}$$

$$\mathrm{k_i(t)} = \begin{cases} \mathrm{x_i(t)} & \mathrm{if} \quad \mathrm{i \in A} \\ \mathrm{y_i(t)} & \mathrm{if} \quad \mathrm{i \in B} \end{cases} \tag{6}$$

where *f*() is the activation associated with the neuron *j*, $\mathrm{w_{ji}}$ is the weight, $\mathrm{b_j}$ is the bias.

In the Eqs. (8), if $i \in A$, it represents that $\mathrm{k_i(t)}$ is real input $\mathrm{x_i(t)}$; If $\mathrm{i \in B}$, it represents that $\mathrm{k_i(t)}$ is the feedback value from output $\mathrm{y_i(t)}$. It means a set of A is the workload information which is currently obtained from cloud servers; and a set of B is the past workload information. Therefore the output of a neural network with *m* neurons in the hidden layer is shown as

$$y(t) = \sum\nolimits_{j=1}^{m} w_j y_j(t) \tag{7}$$

From Eq. (7) and (9), the output vector of the neural network maps to the input vector is shown as:

$$y(t) = \sum\nolimits_{j=1}^{m} w_j f\left(\sum\nolimits_{i \in A \cup B} w_{ji}(t\ -1)k_i(t-1) + b_j\right) + b_0 \tag{8}$$

After the calculated function was found for the neural network, the learning algorithm is needed to define the activation function. Besides, the error function and error update must be defined. The error function *E*(*t*) is defined by selecting a output neuron $S_a(t)$ of the hidden state neurons which at time *t* is shown as:

$$\begin{cases} \text{state : on} & S_a(t) > 1-\beta \quad \text{if} \quad \text{accepted} \\ \text{state : off} & S_a(t) < \beta \quad \text{if} \quad \text{accepted} \end{cases} \tag{9}$$

where β is the tolerance of the response neuron.

There are two error situations which result from the definition, the first is the network fails to reject a negative string, i.e., $S_a(t) > \beta$; the second is the network fails to accept a positive string, i.e., $S_a(t) < 1 - \beta$. The error function is defined as:

$$E(t) = \frac{1}{2}\sum\nolimits_{a=1}^{a} (d_a(t) - S_a(t))^2 \tag{10}$$

where $d_a(t)$ is the target response value for the response neuron $S_a(t)$.

The steepest descent method [3] was used to calculate the weight update rule:

$$\Delta w_{kj} = -\alpha \frac{\partial E(t)}{\partial w_{kj}(t)} = \alpha(d_a(t) - S_a(t)) \frac{\partial S_a(t)}{\partial w_{kj}(t)} \tag{11}$$

where α is the learning rate.

Implementation

The experimental data is collected from the NCHC public environment. This environment is constructed from Hadoop and there are registered 3194 users with a total number of 160,447 jobs having already been run. Before the experiment, the number of neurons and delay times in the hidden layer for the proposed model is needed to find. From the results of experiment, the neural network had 10 neurons and 2 delay times in the hidden layer.

Figure 2 shows the simulation results. The learning effect of the algorithm is divided into three parts for analysis. The first part is the time before 600. It is noted that the experimental result does not seem better than the regression method and TDNN. This is because the proposed model does not have enough data to calculate the more accurate predicted results. The second part is the time between 600 and 1,600. This experimental result is better than previous data because the predicted

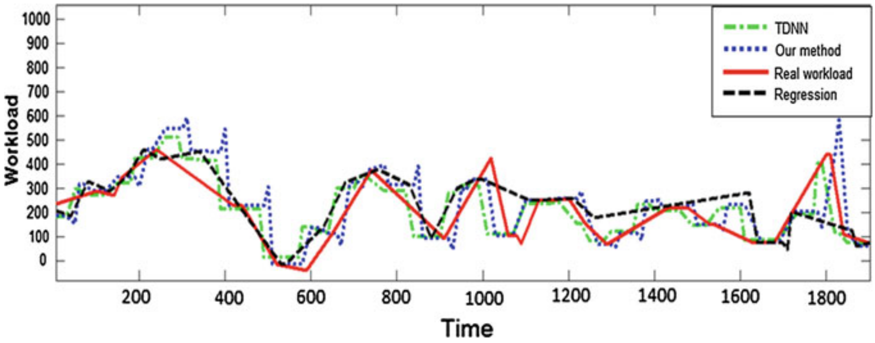


Fig. 2 The simulation result compare with regression and TDNN

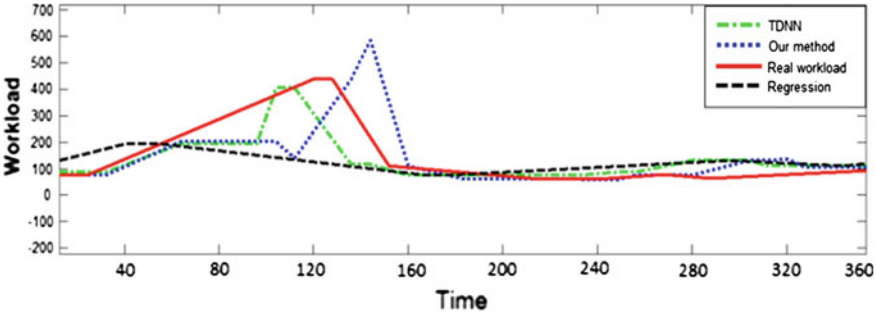


Fig. 3 The enlargement of simulation results compared with regression and TDNN

Table 1 Comparison between TDNN, regression and proposed method

Model	Mean squared error (MSE)
TDNN	557415.88979
Regression	7233256.47972
Proposed method	510293.67772

number is close to the real workload number. The last part is the time after 1,600. The proposed model is also better than the regression method. Even though there is some error, the model still represents a predicted line, which is close to the real line.

Figure 3 shows that the proposed algorithm and the regression method both have some errors in their predictions, but the learning effect of proposed method is clearly better than the regression method, which cannot show rapid changes in the data. The proposed method presents a trend similar to the real workload while the regression method does not. Therefore, the result of proposed method is more accurate than the regression one.

Mean Squared Error (MSE) [5] is used to calculate the predicted results. The calculation of the MSE is shown as below.

$$\mathrm{MSE} = \frac{1}{N}\sum\nolimits_{i=1}^{N} (p - t)^2 \qquad (12)$$

p = Predicted Workload Value, t = Total Workload Value, N = Number of data.

A comparison between TDNN, Regression and Proposed Method is given in Table 1. The result of proposed method is better than TDNN and regression method. Although the proposed method is only a little better than TDNN, TDNN requires a large amount of input data as the training set. Therefore, TDNN is not suitable to run in the new environment. The regression algorithm has high accuracy for prediction problems, but it is only suitable for long time analysis or stable system analysis. It is clearly to see this characteristic from the figures. If a time series over a month or a year are selected, the regression algorithm may not have more error than our algorithm. Consequently, the unit of time is an important factor for the prediction problem. The result shows that the proposed algorithm is suitable for predicting workloads over a short period of time, and the learning effect of proposed method clearly shows the trend of the real workload.

Conclusions

This study proposes the neural network model for predicting the workload of cloud environments. This work helps the cloud provider to avoid some unexpected situations such as a large number of resource requirements from users or applications. The experiment results show that the proposed model is able to predict the workload over a short period of time. Compared to the regression method, the proposed method is more accurate for workload prediction. Hence, the cloud provider can decide the resource allocation management more efficiently by applying the proposed method.

Acknowledgments The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially/partially supporting this research under Contract No. NSC101-2221-E-143-005-, NSC101-2221-E-259-003- and NSC101-2221-E-259-005-MY2.

References

1. Zhang Q, Cherkasova L et al (2007) A regression-based analytic model for dynamic resource provisioning of multi-tier applications. IEEE Int Conf Auton Comput (*ICAC*)
2. Tirando JM, Higuero D, Isaila F, Carretero J (2011) Predictive data grouping and placement for cloud-based elastic server infrastructures. IEEE/ACM international conference on cluster, cloud and grid computing, pp 285–294
3. Mell P, Grance T (2009) The NIST definition of cloud computing. Nat Inst Stand Technol 53(6):50
4. Zhang Z, Wang H, Xiao L, Ruan L (2011) A statistical based resource allocation scheme in cloud. Cloud and service computing (*CSC*), 2011 international conference, pp 266–273

5. Gallant S (1993) Neural network learning and expert systems. MIT Press, Cambridge
6. Peterson C, Södeberg B (1993) Artificial neural networks. Modern heuristic techniques for combinatorial problems. In: Reeves CR (ed) Advanced topics in computer science, Oxford Scientific Publications, New York, pp 197–242
7. Abramson D, Buyya R, Giddy J (2002) A computational economy for grid computing and its implementation in the Nimrod-G resource broker. Future Gener Comput Syst 18(8):1061–1074
8. Picht SW (1994) Steepest descent algorithms for neural network controllers and filters. IEEE Trans Neural Networks 198–212
9. Haykin S (2008) Neural networks and learning machines: a comprehensive foundation, 3rd ed. Prentice Hall
10. Apache Hadoop. http://hadoop.apache.org/
11. Borthakur D (2009) The Hadoop distributed file system: architecture and design. http://hadoop.apache.org/common/docs/current/hdfs-design.html

Implementation of Face Detection Using OpenCV for Internet Dressing Room

Li-Der Chou, Chien-Cheng Chen, Chun-Kai Kui, Der-Ching Chang, Tai-Yu Hsu, Bing-Ling Li and Yi-Ching Lee

Abstract Human face detection is an important technology that is positively developed by industry. Therefore, this paper proposes a dressing application that adopts a face detection technology for better human life. Furthermore, this paper detects human face by using the combination of OpenCV and Internet camera. When the proposed application detect a human face, the proposed application dresses target on a screen, such as putting one selected cloth on the correspond position of human. Finally, the proposed application has a dressing ability for customers who are shopping clothes on Internet, and the dressing ability improves the sales performance of Internet store.

Keywords OpenCV · Internet camera · Face detection

Introduction

Stay-at-home-economy is a huge consumption market. In November 2012, the statistics of Taiwan Institute for Information Industry (III) shows that the gross value of e-commerce in 2012 reach to NT 660.5 billion dollars have 17.4 % growing a year. The III also predicts the gross value of e-commerce will break-through NT 1 trillion dollars in 2015. People save the time for better usage, and geek "Otaku" like to stay home, even those people who lives in remote districts have a problem to shop in urban will use Internet shopping to buy daily living equipment. Clothes are the most common things for customers purchasing via Internet. Internet is very convenient to browse some popular clothes. Customers can quickly search their favorite one from a pile of clothes. However, purchasing

L.-D. Chou (✉) · C.-C. Chen · C.-K. Kui · D.-C. Chang · T.-Y. Hsu · B.-L. Li · Y.-C. Lee
Department of Computer Science and Information Engineering, National Central University, Jhongli, Taiwan
e-mail: cld@csie.ncu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_66, © Springer Science+Business Media Dordrecht 2014

clothes via Internet has one problem: customers cannot try clothes before buy them. Therefore, customers always feels disappoint when they receive their packs. Customers may concern about whether the purchased cloth is fitness or not. Moreover, cloth manufactures have different size model. One cloth is fit but the other is not, even these two clothes are the same size. According to these reasons, customers usually doubt about shopping on Internet.

This paper proposes a concept about the dressing system module which could be a part of a smart store for online shops. In an Internet dressing room, customers can browse popular clothes conveniently. They can not only see the model demonstrate the cloth, but also experience when they put clothes on. The proposed application makes some lazy customers can try new clothes comfortably and conveniently when they shopping on Internet. Furthermore, Internet dressing room do attract more popularity and reduce the cost of physical shop. Since the Internet dressing room, clothes can cost down, and Internet dressing room promotes the aspiration of shopping on Internet.

The proposed application is implemented by using an internet camera and face detection function in OpenCV. The application locates the human face by detection technology. After calculate the location of the face, the application will get the position of the human body. Finally, the application puts the chosen cloth on the body. The rest of this paper is organized as follows. Section Related Work describes related works about OpenCV. In section Implementation, we propose the proposed application. Section Conclusion concludes the paper.

Related Work

OpenCV is developed by Intel cooperation. Open represents Open Source and CV is Computer Vision [1]. OpenCV is an Open Source Image Process Library and it can be used to make image, video, matrix operations, statistics, graphic, data storing programed by C language. OpenCV can be used for image processing, computer vision, pattern recognition, computer graphics, game design, and so on. The most famous application of OpenCV is face detection, it can also use for integrate different picture format's matrix operation. Apply OpenCV in static picture (i.e. BMP, JPG, TIF, PNG), and image processing of dynamic Webcam. OpenCV provides an easy GUI interface. OpenCV can be integrated into Visual C++ and C++ Builder. OpenCV can support Windows and Linux. OpenCV has equipped powerful graphic and matrix operations [1]. However, OpenCV is not supported by Intel cooperation, so it cannot offer some latest algorithm. But it has been revision recently. In new vision, some new function, like Scale-invariant feature transform (SIFT) for computer vision, has been added.

Through the combination of Kinect and OpenCV, color image, 3D depth image, and voice signal can be catch in one time. Kinect equipped with 3 shots [2]. The middle shot is use for recognizes the identification and detects facial expression of the user. Kinect will turn the 3D depth image into skeleton tracing system.

This system can detect at most 6 people in one time. Including the trunk, arms, legs, even fingers are in tracing area. Microsoft also implement machine learning and large image database for understand users motivation.

Implementation

The system flow chart diagram is shown on Fig. 1. When users start the program, the application shows a cloth catalog provided by cloth vendors. After user choose a dress he/she like, the application judges whether the computer is equipped with camera or not. The application executes cvGrabFrame() function after judgment. The cvGrabFrame() function mainly grabs a image from the camera. Then, the application detects whether there is a human face in this image. If yes, this function will return the position of human face. Therefore, the proposed application can predict where the human body is. Finally, the chosen cloth picture is shown on the predicted position of body.

System Architecture

The proposed application provides user a cloth menu that is offered by cloth vendors. Users may choose some cloths they have will to buy. System confirms whether a camera is ready. If yes, the next procedure is using the ready camera to take an image and decides where the human face is. According the position of human face on an image, system deduces the position of body by comparing the face position with depth image. System automatically puts one chosen cloth picture on body and shows the composite portrait on a screen. Finally, system provides shopping advices to customer. The detail of system architecture is shown on Fig. 2.

Select the Cloth

This function provides a menu that lists all clothes of one store for customers. In Fig. 3a, customers can browse clothes and choose one cloth they like. After customers make a selection from the three cloths image on a screen, system go to next step. The application can increase products database based on salesman's whosh in the future.

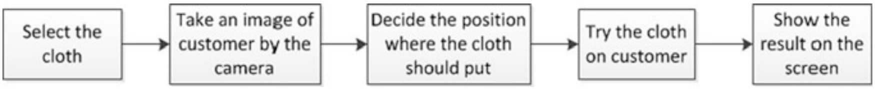
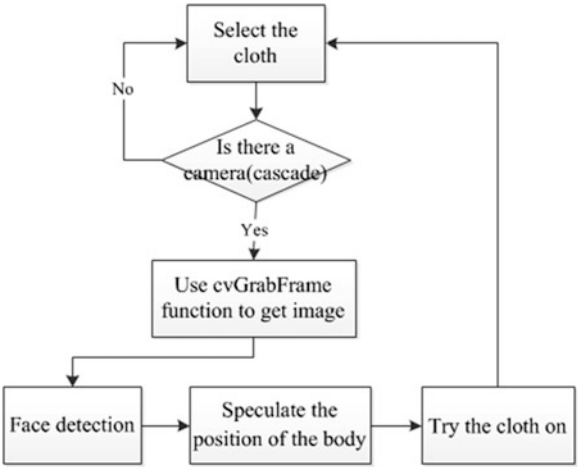

Select the cloth
Take an image of customer by the camera
Decide the position where the cloth should put
Try the cloth on customer
Show the result on the screen

Fig. 1 System flow chart diagram

Fig. 2 System architecture


Select the cloth
No
Is there a camera(cascade)
Yes
Use cvGrabFrame function to get image
Face detection
Speculate the position of the body
Try the cloth on

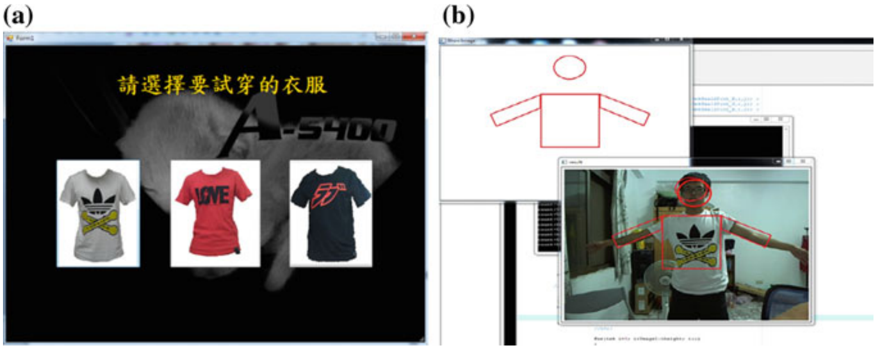


Fig. 3 **a** Select the cloth; **b** Face detection

Face Detection

This function detects human face, if users make a selection. A camera catches an image. The image is processed by AdaBoost Leaning with Haar-Like Features algorithm for face detection.

The first step of face detection is to define some Haar-Like features. Then, inputting some sample. For example, system detects the face features according to some face samples come from white men, black girl, and yellow boy. Then, this paper uses AdaBoost learning algorithm to find out some representative Haar-Like features. In this case, this function picks up some face features. Each feature represents a class. All features constitute a class array which called "strong classifier." Each class is used to analyze the input image whether contains a face or

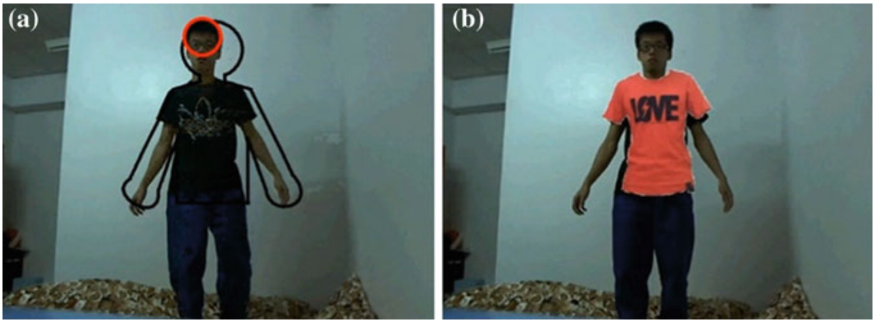


Fig. 4 **a** Speculate the position of the body; **b** Try the cloth on

not. If all classifiers return true to this image, then this image will be identified to be a face. In Fig. 3b user's face can be detected accurately.

Speculate the Body Position

This function begins to speculate the body position after the face detection function returns the coordination of human face. This function determines where the cloth should be put on, as shown in Fig. 4a. Through the image taken by the camera, this function recognizes the posture of the body. The posture will be compared with depth image and find out the most similar human body model. Finally, the application will get the real post of human and recognize every part of human body.

Try the Cloth On

This function decides the location of the cloth based on the function of speculate the body position. The application puts the chosen cloth on the position dynamically. Finally, accomplish the function of try the cloth on human.

Conclusion

This application detects face successfully by OpenCV and camera. After figure out the coordinate of a human face, the application will speculate where the body is and put the dress on the body position. Before this application, customers cannot try on the cloth unless they purchase on real store. According to this reason, a lot of customers don't like to buy the dress on the internet. By this application, customers' requirement could be satisfied easily. They can try on the cloth without span a lot of time to go to the physical store. Furthermore, this application could promote the developing of shopping on Internet. This application can not only increase the sales but also decrease the cost of real store.

In this thesis, the application merge camera, real-time face detection technology and body position speculate algorithm. This system makes a win-win situation to salesman and customers.

References

1. Intel Corporation, Open source computer vision library. Reference manual, Copyright © 1999–2001. Available. http://software.intel.com/en-us
2. http://www.techbang.com/posts/2936-get-to-know-how-it-works-kinect
3. Wren C, Azarbayejani A, Darrell T, Pentland A (1997) Pfinder: real-time tracking of the human body. IEEE Trans Pattern Anal Mach Intell 19(7):780–785
4. Lorsakul A, Suthakorn J Traffic sign recognition for intelligent vehicle/driver assistance system using neural network on OpenCV. International conference on ubiquitous robots and ambient intelligence
5. Haj MAI, Amato A, Roca X, Gonzàlez J (2007) Face detection in color images using primitive shape features. Computer Recognition Systems 2, vol 45
6. Haj MA et al (2009) Robust and efficient multipose face detection using skin color segmentation. Proceedings of the 4th Iberian conference on pattern recognition and image analysis
7. Chen L-F, Liao H-YM, Lin J-C, Han C–C (2001) Why recognition in a statistics-based face recognition system should be based on the pure face portion: a probabilistic decision-based proof. Pattern Recogn 34(5):1393–1403
8. Lapedriza A, Marin-Jimenez MJ, Vitria J (2006) Gender recognition in non-controlled environment. 18th IEEE international conference on pattern recognition, vol 3, pp 834–837
9. Hu M (1962) Visual pattern recognition by moment invariants. IRE Trans Inf Theory IT-8 179–187
10. Teague MR (1980) Image analysis via the general theory of moments. J Opt Soc Am 70:920–930
11. Khotanzad A, Hong YH (1990) Invariant image recognition by Zernike moments. IEEE Trans Pattern Anal Mach Intell 489–497
12. Ahonen T, Hadid A, Pietikäinen M (2004) Face recognition with local binary patterns. ECCV 2004

A Comparative Study on Routing Protocols in Underwater Sensor Networks

Jian Shen, Jin Wang, Jianwei Zhang and Shunfeng Wang

Abstract Underwater sensor networks (UWSNs) are a class of emerging networks that experience variable and high propagation delays and limited available bandwidth. Because of the different environment under the ocean, the whole protocol stack should be re-designed to fit for the surroundings. In this paper, we only focus on the routing protocols in the network layer. We survey the state-of-the-art routing protocols and give a comparison of them with respect to the important challenging issues in UWSNs. The pros and cons are discussed and compared for the routing protocols.

Keywords Underwater sensor networks (UWSNs) · Routing protocols · Range technology

Introduction

As the network communications technology developing, a new type of networks has appeared in the daily life which named underwater sensor networks (UWSNs). This research area has more attractiveness than ground-based networks because of its distinctive characteristics and the comprehensive applications. UWSNs are an

J. Shen (✉) · J. Wang
School of Computer and Software, Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Science and Technology, Nanjing 210044, China
e-mail: s_shenjian@126.com

J. Zhang
College of Mathematics and Statistics, Nanjing University of Information Science
and Technology, Nanjing 210044, China

S. Wang
College of Bin Jiang, Nanjing University of Information Science and Technology,
Nanjing 210044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_67, © Springer Science+Business Media Dordrecht 2014

occasionally connected network and experience frequent and long-duration partitions as well as long delay.

Moreover, UWSNs are very interesting in the ocean exploration applications and very important in military applications. In general, UWSNs are envisioned to enable applications for oceanographic data collection, pollution monitoring, offshore exploration, disaster prevention, assisted navigation and tactical surveillance applications. Multiple UUVs (unmanned underwater vehicles) and AUVs (autonomous underwater vehicles) equipped with underwater sensors, will also find application in exploration of natural undersea resources and gathering of scientific data in collaborative monitoring missions [1]. Sensors and vehicles under water manage and organize by themselves in an autonomous network which can adapt to the characteristics of the ocean environment in order to carry out a great variety of exploration and research missions.

The different environments under the ocean and such distinct features compared with the ground-based networks pose a number of technical challenges in designing the whole protocol stack [2]. In this paper, we only focus on the routing protocols in the network layer. We survey the state-of-the-art routing protocols and give a comparison of them with respect to the important challenging issues in UWSNs. The routing protocols are classified into three categories: proactive, reactive and geographical routing protocols. The pros and cons are discussed and compared for the routing protocols.

The rest of this paper is organized as follows. In the following section, the key properties of UWSNs are reviewed. Routing issues in UWSNs are presented in section Routing Issues in UWSNs. The existing routing protocols are summarized in terms of major features and characteristics in section Routing Protocols, and they are compared and discussed in section Comparison. Finally, the conclusions of this paper are covered in Conclusion.

Key Properties of UWSNs

There are some important key properties of UWSNs which have a great deal of discrepancy from the ground-based networks. They are briefly reviewed in this section.

Acoustic Wireless Communication

It is well-known that radio waves, optical waves and acoustic waves are three different kinds of information carriers. However, radio waves propagate at long distances through conductive sea water only at extra low frequencies. Meanwhile, optical waves do not suffer from the high attenuation but are affected by scattering. Therefore, acoustic wireless communication becomes the best choice in UWSNs.

Variable and High Propagation Delays

Variable and high propagation delays are the fundamental properties of UWSNs. The transmission under water experiences intermittent connectivity, and the transmit rate may be considerably low. Of course, the propagation delay is very high under water because of the underwater environment and asymmetric connection [3].

Limited Available Bandwidth

The available bandwidth is severely limited because of the underwater transmission. In general, the sensors in UWSNs are mobile and battery-operated with wireless connection and, thus, they have limited bandwidth.

Severely Impaired Channel

The multi-path transmission and the transmission fading severely affect the underwater channel. Acoustic waves do not pass through obstructions very well. In addition, even though the beam may be well focused at the transmitter, there is still some divergence during transmission [4].

High Bit Error Rates and Limited Battery Power

High bit error rates and temporary losses of connectivity can be experienced, due to the extreme characteristics of the underwater channel. Battery power is limited and usually batteries cannot be recharged, also because solar energy cannot be exploited.

Fouling and Corrosion

Underwater sensors are prone to failures because of fouling and corrosion. There are lots of corrosive chemical elements under the ocean such as sulfur, nitrogen, chlorine, bromine and so on. Hence, the sensors are easily destroyed or broken down by oxidative decomposition.

Routing Issues in UWSNs

The peculiar properties of UWSNs inevitably raise a number of interesting issues which are summarized in this section [5].

Routing Objective

The most important routing objective in UWSNs is to minimize the transmission delay. To minimize the communication signaling overhead and establish energy efficient paths are also important routing objectives. Because of the complex environment under water, as we known, the communication is asymmetric.

Reliability

For the reliable delivery of data in UWSNs, any routing protocol should have some acknowledge, which can ensure successful and stable delivery of data. For example, when a message correctly reaches to a destination, some acknowledge messages should be sent back from destination to source for later use.

Energy

Because of the severe environment underwater, it is impossible to change the sensor nodes frequently. It is also difficult to connect the underwater sensor nodes to the power station. Hence, the nodes in UWSNs are usually lack of energy. During the message routing lots of energy should be consumed for sending, receiving and storing messages as well as performing computation. Therefore, the energy-efficient design of routing protocols is of importance.

Security

Security is always an important issue not only in the UWSNs but also in all the traditional networks. The cryptographic techniques may be beneficial for secure end-to-end routing. The security in routing protocols is still an open issue to be studied.

Routing Protocols

In general, the existing routing protocols are usually divided into three categories, namely proactive, reactive and geographical routing protocols. Proactive routing protocols are traditional distributed shortest-path protocols, which maintain routes between every host pair at all time. Based on periodic updating, the routing overhead is very high by utilizing proactive routing protocols. Reactive routing protocols determine route if and when needed, in which source initiates route discovery. Geographical routing protocols establish source–destination path by the localization information, in which the sender uses a location service to determine the position of the destination. And the routing process at each node is based on the destination's location within the packet header and the location of the forwarding node's neighbors.

Proactive Routing Protocols

The proactive routing protocols attempt to minimize the message latency induced by route discovery, by maintaining up-to-date routing information at all times from each node to every other node. This is obtained by broadcasting control packets that contain routing table information. These protocols provoke a large signaling overhead to establish routes for the first time and each time the network topology is modified because of mobility or node failures, since updated topology information has to be propagated to all the nodes in the network. This way, each node is able to establish a path to any other node in the network, which may not be needed in UWSNs. For this reason, proactive protocols maybe not suitable for underwater networks.

Destination-Sequenced Distance-Vector Protocol

DSDV is a proactive hop-by-hop distance vector routing protocol. Every host maintains a routing table for all the possible destinations and the number of hops to each destination. Meanwhile, each host broadcasts routing updates periodically in order to achieve the newest and the most accurate routing table [6].

Wireless Routing Protocol

In wireless routing protocol, sensor nodes keep each other informed of all link changes through the use of update messages [7]. Obviously, WRP is a table-driven protocol with the goal of maintaining routing information among all sensor nodes.

Reactive Routing Protocols

In reactive routing protocols, a node initiates a route discovery process only when a route to a destination is required. Once a route has been established, it is maintained by a route maintenance procedure until it is no longer desired. These protocols are more suitable for dynamic environments but incur a higher latency and still require source-initiated flooding of control packets to establish paths. Reactive protocols are deemed to be unsuitable for UWSNs as they cause a high latency in the establishment of paths, which may be even amplified underwater by the slow propagation of acoustic signals. Furthermore, links are likely to be asymmetrical, due to bottom characteristics and variability in sound speed channel. Hence, protocols that rely on symmetrical links, such as most of the reactive protocols, are unsuited for the underwater environment.

Dynamic Source Routing Protocol

In brief, source routing means the source node is in charge of the whole transmitting and determines the path based on the topology of the network before the message gets into the node. The source must know all the intermediate nodes to be traversed from the source to a destination.

Ad Hoc On-demand Distance Vector Routing Protocol

AODV is an improvement on DSDV because it minimizes the number of the required broadcasts by creating routes on demand basis [8]. It carries out the route discovery by using on-demand mechanism and maintains from DSR.

Temporarily Ordered Routing Algorithm

TOAR is a highly adaptive loop-free distributed routing algorithm. It minimizes reaction due to topological changes while exhibits multipath routing capability [7].

Geographic Routing Protocols

These protocols establish source–destination path by the localization information. Each node selects its next hop based on the position of its neighbors and of the destination node. In fact, fine-grained localization usually requires strict synchronization among nodes, which is difficult to achieve underwater due to the variable propagation delay. Virtual circuit routing techniques can be considered in

UWSNs. In these techniques, paths are established a priori between each source and sink, and each packet follows the same path. Localization schemes are the most important issues in geographic routing protocols [9].

Routing Protocols with Respect to Infrastructure-Based Localization Schemes

Infrastructure-based (anchor-based) localization systems are similar to the GPS scheme. The distance to multiple anchor nodes is computed by using the propagation time of the sound signals between the sensor or the AUV and the anchors.

Routing Protocols Based on Distributed Positioning Schemes

In distributed positioning schemes, nodes are able to communicate only with their one-hop neighbors and compute the distances to their one-hop neighbors by making RSSI or ToA measurements.

Routing Protocols with Respect to Mobile Beacon-Based Schemes

In this scheme, a mobile beacon traverses the sensor network while broadcasting beacon packets, which contain the location coordinates of the beacon. RSSI measurements of the received beacon packets are used for ranging purposes.

Routing Protocols with Respect to Hop Count-Based Schemes

In brief, the anchor nodes are placed at the corners or along the boundaries of a square grid. In order to estimate the distance to the landmark by approximately knowing one hop distance, each node maintains and updates a table in term of hop number.

Routing Protocols Based on Centroid Scheme

In this scheme, anchor nodes are placed to form a rectangular mesh. The location of the node is then estimated to be the centroid of the anchor nodes that it can receive beacon packets.

Table 1 Comparison of the three different kinds of routing protocols

	Flexibility	Route acquisition	Resource usage	Multipath support	Flood for route discovery	Latency	Overhead	Routing table	Effectiveness
DSDV	Bad	Compute a priori	High	No	No	Short	High	Yes	Bad
WRP	Bad	Compute a priori	High	No	No	Short	High	Yes	Bad
DSR	Normal	On-demand	Normal	Not explicitly	Yes	Long	Normal	Yes	Bad
AODV	Normal	On-demand	Normal	Not directly	Yes	Long	Normal	Yes	Bad
TORA	Normal	On-demand	Normal	Yes	Yes	Long	Normal	Yes	Bad
Geographic routing protocols	Good	Compute a priori	Low	No	No	Normal	Low	No	Good

Routing Protocols with Respect to Area-Based Localization Scheme

In very large and dense wireless sensor networks, a coarse estimate of the sensors' locations may suffice for most applications. ALS [10] and APIT [11] approximate the area in which a node is located, rather than the exact location.

Comparison

We compare the state-of-the-art routing protocols proposed so far in the literature. First, we evaluate the three kinds of routing protocols such as proactive routing protocols, reactive routing protocols and geographic routing protocols, in terms of various characteristics including important performance metrics. Flexibility, route acquisition, resource usage, multipath support, flooding for route recovery, latency, overhead, routing vector/table, and effectiveness are studied in the comparative analysis. Table 1 summarizes the comparison results.

From the comparison table, some conclusive comments can be inferred: The geographic routing protocols are most suitable for underwater communication. Of the geographic routing protocols, ALS can be primarily chosen thanks to its many outstanding features although it has not accurate information.

Conclusion

Routing in UWSNs is a new area of research, with a limited but rapidly growing set of research results. The routing protocols have the common objective of trying to increase the delivery ratio while decreasing the resource consumption and latency. In this paper, we have presented a comparative survey of various routing techniques in UWSNs. Meanwhile, we have studied about some localization schemes on geographic routing protocols. The advantages and disadvantages of the routing protocols have been discussed with comparison results as well. Although many routing protocols have been studied so far, there are still many challenges to be solved. For example, we should make the routing protocol more scalable for a large network. Loop freedom, energy conservation, and efficient resource usage should be also addressed. Our future work is to design a robust routing protocol with good localization schemes for harsh operational environments.

Acknowledgments This work was supported by the research fund from Nanjing University of Information Science and Technology under Grant No. S8113003001, National Science Foundation of China under Grant No. 61272421.

References

1. Akyildiz F, Pompili D, Melodia T (2004) Underwater acoustic sensor networks: research challenges
2. Rahman RH, Benson C, Frater M (2012) Routing protocols for underwater ad hoc networks. OCEANS, 2012—Yeosu, pp 1–7
3. Proakis JG, Rice JA, Sozer EM, Stojanovic M (2003) Shallow water acoustic networks. Encyclopedia of Telecommunications
4. Tanenbaum AS (2003) Computer networks. Fourth edition, pp 104
5. Yang X, Ong KG, Dreschel WR, Zeng K, Mungle CS, Grimes CA (2002) Design of a wireless sensor network for long-term, in situ monitoring of an aqueous environment. Sensors, pp 455–472
6. Perkins CE, Bhagwat P (1994) Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. ACM SIGCOMM Commun Rev
7. Cordeiro CM, Agrawal DP (2006) Ad Hoc and sensor networks, pp 23–24
8. Perkins C, Belding-Royer E, Das S (2003) Ad hoc on-demand distance vector (AODV) routing. Internet RFCs
9. Chandrasekhar V, Seah WKG (2006) Localization in underwater sensor networks—survey and challenges. In: International conference on mobile computing and networking, pp 33–40
10. Chandrasekhar V, Seah WKG (2006) Area localization scheme for underwater sensor networks. In: the IEEE OCEANS Asia Pacific Conference
11. He T (2003) Range-free localization schemes for large scale sensor networks. In: 9th ACM international conference on mobile computing and networking (Mobicom2003)

A Novel Verifiably Encrypted Signature from Weil Pairing

Jian Shen, Jin Wang, Yuhui Zheng, Jianwei Zhang and Shunfeng Wang

Abstract A novel efficient verifiably encrypted signature (VES) which makes use of the Weil pairing is presented in this paper. VES can be used in optimistic contract signing protocols to enable fair exchange. Compared with the previous schemes, our proposed VES scheme is more efficient.

Keywords Verifiably encrypted signature · Weil pairing · Fair exchange

Introduction

The verifiably encrypted signature (VES) was first proposed by Asokan et al. [1], which can keep well the fairness of the trading. The realization of VES relies on the trusted third party which needs not to join the exchange protocol in the online mode. Verifiably encrypted signatures are used in optimistic contract signing protocols [1, 2] to enable fair exchange. For example, by using VES, a user Alice can give Bob a signature on a message encrypted by a third party's public key. Meanwhile Bob can verify the encrypted signature after he receives it but cannot deduce any information of the ordinary signature. In 2003, Boneh et al. [3] first proposed a practical verifiably encrypted signature scheme based on bilinear maps.

J. Shen (✉) · J. Wang · Y. Zheng
School of Computer and Software, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, China
e-mail: s_shenjian@126.com

J. Zhang
School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China

S. Wang
College of Bin Jiang, Nanjing University of Information Science and Technology, Nanjing 210044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_68, © Springer Science+Business Media Dordrecht 2014

Recently, in [4], Ming proposed a security model of verifiably encrypted signature schemes. In his security model, a trusted third party (TTP) called adjudicator is used which generates a pair of public key and private key. The public key serves as a public parameter of the system and the corresponding private key kept secretly by the adjudicator is used to resolve the possible dispute between two trading parties.

In this paper, we develop a more efficient verifiably encrypted signature from the Weil pairing. Our scheme requires less computation complexity compared with the previous schemes.

Preliminaries

In this section, we briefly introduce the Weil pairing which is necessary for description of our signature scheme.

Weil pairing: Let p be a prime number such that $p = 12q - 1$ for some prime number q and E a super singular elliptic curve defined by the Weierstrass equation $y^2 = x^3 + 1$ over F_P. The set of rational points $E(F_p) = \{(x, y) \in F_p \times F_p : (x, y) \in E\}$ forms a cyclic group of order $p + 1$. Furthermore, because $p + 1 = 12q$ for some prime number q, the set of points of order q in $E(F_p)$ form a cyclic subgroup, denoted as G_1. Let $\mathcal{G}$ be the generator of G_1. Let G_2 be the subgroup of F_{p^2} containing all elements of order q. The modified Weil pairing [5] is a map:

$$\widehat{e} : G_1 \times G_1 \rightarrow G_2,$$

which has the following properties:

- Bilinear: For any $\mathcal{P}, \mathcal{Q} \in G_1$ and a, $b \in Z$, we have $\widehat{e}(aP, bQ) = \widehat{e}(P, Q)^{ab}$.
- Non-degenerate: if $\mathcal{G}$ is a generator of G_1, then $\widehat{e}(\mathcal{G}, \mathcal{G}) \in F_{p^2}^*$ is a generator of G_2.
- Computable: Given $\mathcal{P}, \mathcal{Q} \in G_1$, there is an efficient method to compute $\widehat{e}(P, Q) \in G_2$.

Efficient Verifiably Encrypted Signature from Weil Pairing

There are three entities in our signature scheme: user, verifier and adjudicator. Our signature scheme runs in seven algorithms: *KeyGen*, *Sign*, *Verify*, *AdjKeyGen*, *VESSign*, *VESVerify*, and *Adjudication*. The algorithms are described as follows:

- *KeyGen*: A user randomly picks $x \in Z_p^*$ as its private key. The user's public key is computed as: $x \cdot P = P_{pub} = (x_p, y_p)$.

- *Sign*: Given the private key x of the user, the message $m \in (0,1)^l$, the hash function H and the public key $P_{pub} = (x_p, y_p)$, the use signs a signature σ on m: $\sigma = \frac{H(m)}{x+x_p} \cdot P$
- *Verify*: Given the public key $P_{pub} = (x_p, y_p)$, the message m and the signature σ, the verifier verifies whether $\widehat{e}\left(\sigma, \frac{P_{pub}+x_p \cdot P}{H(m)}\right) = \widehat{e}(P, P)$.
- *AdjKeyGen*: The adjudicator randomly picks $y \in Z_p^*$ as its private key, and computes the corresponding public key as $P_{pub}{}' = y \cdot P$.
- *VESSign*: Given the user's private key x, the message m and the adjudicator's public key $P_{pub}{}'$, the user computes the verifiably encrypted signature σ_{VES} as $\sigma_{VES} = \frac{H(m)}{x+x_p} \cdot P_{pub}{}'$.
- *VESVerify*: Given the verifiably encrypted signature σ_{VES}, the message m, the public key of user $P_{pub} = (x_p, y_p)$, and the public key of adjudicator $P_{pub}{}'$, the verifier verifies whether $\widehat{e}\left(\sigma_{VES}, \frac{P_{pub}+x_p \cdot P}{H(m)}\right) = \widehat{e}\left(P_{pub}{}', P\right)$ can hold.
- *Adjudication*: Given the verifiably encrypted signature σ_{VES}, the message m, and the private key of adjudicator, the adjudicator extracts the original signature on the message m as: $\sigma = \frac{\sigma_{VES}}{y}$.

Validity Analysis

Validity requires that verifiably encrypted signatures are able to be successfully verified as ordinary signature and adjudicated verifiably encrypted signatures are also able to be successfully verified as ordinary signature. This means *VESVerify*(*m*, *VESSign*(*m*)) and *Verify*(*m*, *Adjudication*(*VESSign*(*m*))) hold for all *m* and for all properly generated key pairs and adjudicator key pairs.

For a verifiably encrypted signature σ_{VES} on a message m, the validity is easily proven as follows:

$$\begin{aligned}
\widehat{e}\left(\sigma_{VES}, \frac{P_{pub} + x_p \cdot P}{H(m)}\right) &= \widehat{e}\left(\frac{H(m)}{x + x_p} \cdot P_{pub}{}', \frac{P_{pub} + x_p \cdot P}{H(m)}\right) \\
&= \widehat{e}\left(\frac{H(m)}{x + x_p} \cdot P_{pub}{}', \frac{x \cdot P + x_p \cdot P}{H(m)}\right) \\
&= \widehat{e}\left(\frac{H(m)}{x + x_p} \cdot P_{pub}{}', \frac{x + x_p}{H(m)} \cdot P\right) \\
&= \widehat{e}\left(P_{pub}{}', P\right)
\end{aligned}$$

Hence, *VESVerify*(*m*, *VESSign*(*m*)) $= 1$ holds. On the other hand,

Table 1 Performance comparison

	Boneh's scheme	Ming's scheme	Our scheme
Size	320 bits	320 bits	320 bits
VESSign	3 M	3 M	1 M
VESVerify	3e	3 M + 1e	1 M + 1e
Adjudication	1 M	1 M	1 M

$$\begin{aligned}\widehat{e}\left(\sigma, \frac{P_{pub} + x_p \cdot P}{H(m)}\right) &= \widehat{e}\left(\frac{H(m)}{x + x_p} \cdot P, \frac{P_{pub} + x_p \cdot P}{H(m)}\right) \\ &= \widehat{e}\left(\frac{H(m)}{x + x_p} \cdot P, \frac{x \cdot P + x_p \cdot P}{H(m)}\right) \\ &= \widehat{e}\left(\frac{H(m)}{x + x_p} \cdot P, \frac{x + x_p}{H(m)} \cdot P\right) \\ &= \widehat{e}(P, P)\end{aligned}$$

Therefore, *Verify*(m, *Adjudication*(*VESSign*(m))) $= 1$ holds.

Performance Analysis

We compare our proposed signature scheme with previous schemes [3, 4] in Table 1. Note here that M means a scalar point multiplication and e indicates a pairing computation. In the proposed scheme, the size of the VES is 320 bits, which has the same security level compared with 1,024-bit RSA signature. As described above, there are seven algorithms in our scheme including *KeyGen*, *Sign*, *Verify*, *AdjKeyGen*, *VESSign*, *VESVerify*, and *Adjudication*. The costs of the first three algorithms in our scheme are almost same as that in the previous schemes. Hence, we focus only on comparing the cost of *VESSign*, *VESVerify*, and *Adjudication*. It is worth noting that, in our scheme, we can pre-compute $\widehat{e}\left(P_{pub}', P\right)$ in *VESVerify* phase. Therefore, there is only one pairing computation. From Table 1, we can conclude that our scheme is more efficient than Boneh's and Ming's schemes.

Conclusion

Verifiably encrypted signature is very important and useful cryptographic primitives. We proposed a new verifiably encrypted signature scheme based on bilinear pairings. We show that our scheme is more efficient than previous schemes.

Acknowledgments This work was supported by the research fund from Nanjing University of Information Science and Technology under Grant No. S8113003001 and Natural Science Foundation of Jiangsu Province under Grant No. BK2011825.

References

1. Asokan N, Shoup V, Waidner M (2000) Optimistic fair exchange of digital signature. IEEE J. Selected Areas in Comm. 18(4):593–610
2. Bao F, Deng R, Mao W (1998) Efficient and practical fair exchange protocols with offline TTP. In: IEEE symposium on security and privacy, pp 77–85
3. Boneh D, Gentry C, Lynn B, Shacham H (2003) Aggregate and verifiably encrypted signatures from bilinear maps. In: Eurocrypt'03, LNCS 2656, pp 416–432
4. Ming Y, Wang Y (2009) An efficient verifiably encrypted signature scheme without random oracle. Int J Network Security 8(2):125–130
5. Boneh D, Franklin M (2001) Identity-based encryption from the Weil pairing. In: Crypto'01, Santa Barbara, CA, USA, pp 213–229

Location-Aware Routing Protocol for Underwater Sensor Networks

Jian Shen, Jin Wang, Jingwei Wang, Jianwei Zhang and Shunfeng Wang

Abstract As the network communications technology developing, a new type of networks has appeared in the daily life which is named underwater sensor networks (UWSNs). Routing protocols in UWSNs should ensure the reliability of message transmission, not just decrease the delay. In this paper, we propose a novel routing protocol named Location-Aware Routing Protocol (LARP) for UWSNs, where the location information of nodes are used to help the transmission of the message. Simulation results show that the proposed LARP outperforms the existing routing protocols in terms of packet delivery ratio and normalized routing overhead.

Keywords Underwater sensor networks (UWSNs) · Location-aware · Anchor node

Introduction

Underwater sensor networks (UWSNs) are a class of emerging networks that experience variable and high propagation delays and limited available bandwidth. Compared with ground-based networks, UWSNs has more attractiveness due to its

J. Shen (✉) · J. Wang · J. Wang
School of Computer and Software, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, China
e-mail: s_shenjian@126.com

J. Zhang
School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China

S. Wang
College of Bin Jiang, Nanjing University of Information Science and Technology, Nanjing 210044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_69, © Springer Science+Business Media Dordrecht 2014

distinctive characteristics and the comprehensive applications. UWSNs are very interesting in the ocean exploration applications and very important in military applications, such as oceanographic data collection, pollution monitoring, offshore exploration, disaster prevention, assisted navigation and tactical surveillance applications [1]. In addition, Multiple unmanned underwater vehicles (UUVs) and autonomous underwater vehicles (AUVs) equipped with underwater sensors will also find application in exploration of natural undersea resources and gathering of scientific data in collaborative monitoring missions [2].

UWSNs have great potential and contain enormous values in economic and social field [3]. Sensors and vehicles under water manage and organize by themselves in an autonomous network which can adapt to the characteristics of the ocean environment in order to carry out a great variety of explore and research missions [4]. Because of the different environment under the ocean, the routing protocol should be re-designed to fit for the surroundings. However, the different environments under the ocean and such distinct features compared with the ground-based networks pose a number of technical challenges in designing the routing protocol [5]. In this paper, we propose a novel routing protocol named Location-Aware Routing Protocol (LARP) for UWSNs, where the location information of nodes are used to help the message transmission. Resort to a range-finding technique called received signal strength indicator (RSSI) [6], a node can easily obtain its location information [7]. Simulation results show that the presented LARP outperforms the existing routing protocols in terms of packet delivery ratio and normalized routing overhead.

The rest of this paper is organized as follows. In the following section, related works on routing protocols in UWSNs are briefly discussed. A Novel Location-Aware Routing Protocol for UWSNs is described in detail in section A Novel Location-Aware Routing Protocol for UWSNs. Simulations and results are presented in section Simulations and Results. Finally, the conclusions of this paper are covered in section Conclusions.

Related Works

Compared with ground-based networks, UWSNs has the following key properties: (1) acoustic wireless communication, (2) variable and high propagation Delays, (3) limited available bandwidth, (4) severely impaired channel, (5) high bit error rates and limited battery power, (6) fouling and corrosion. Some researchers have made a lot of effort in designing new protocols in this area. In general, the routing protocols in UWSNs are classified into three categories: proactive, reactive and geographical routing protocols.

Proactive Routing Protocols

The proactive routing protocols attempt to minimize the message latency induced by route discovery. This is obtained by broadcasting control packets that contain routing table information. These protocols provoke a large signaling overhead to establish routes for the first time. The representative of this category is destination-sequenced distance-vector (DSDV) [8] protocol.

Reactive Routing Protocols

In reactive routing protocols, a node initiates a route discovery process only when a route to a destination is required. Once a route has been established, it is maintained by a route maintenance procedure until it is no longer desired. These protocols are more suitable for dynamic environments. The representative of this category is ad hoc on-demand distance vector routing (AODV) [9] protocol.

Geographic Routing Protocols

These protocols establish source–destination path by the localization information. Each node selects its next hop based on the position of its neighbors and of the destination node. Localization schemes are the most important issues in geographic routing protocols [10]. The representative of this category is area localization scheme (ALS) [11] for UWSNs.

A Novel Location-Aware Routing Protocol for UWSNs

In this section, LARP is described in detail. We utilize anchor nodes to estimate the location information of general nodes. We determine the best next hop to relay the message by location information. In our protocol, anchor nodes equipped with GPS traverse the sensor network and broadcast beacon packets, which contain the location coordinates. RSSI measurements of the received beacon packets are used for ranging purposes. General nodes estimate the location information by cooperating with at least three anchor nodes. Every node stores its own location information. When an anchor node is situated in the transmission range of a certain general node, the information of this general node can be stored in this anchor node.

The proposed routing protocol has two steps. At the beginning of routing, the location information of destination node should be obtained by the source first. Suppose that there is a message transmitted from the source (S) to the destination

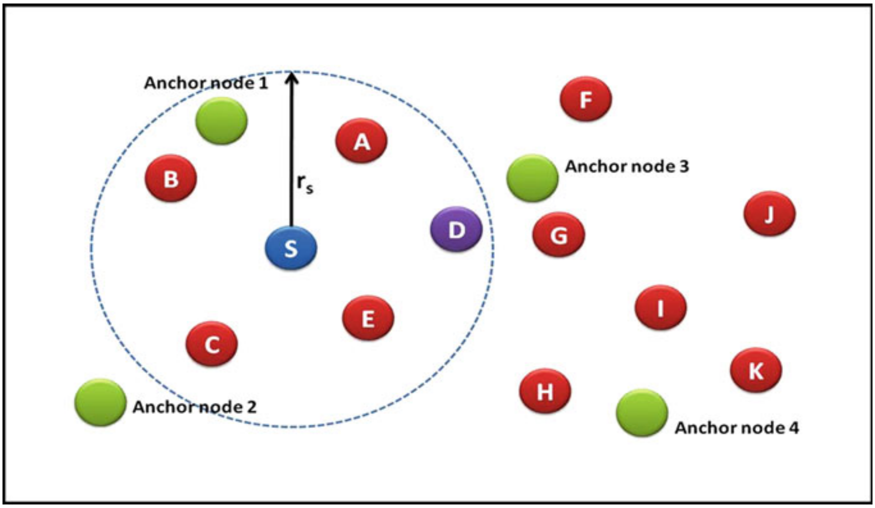


Fig. 1 Node S broadcasts “destination location” request

(D). If there is an anchor node in the transmission range of node S, S can request the anchor node to find the destination’s location. Otherwise, node S will wait until an anchor node appears in its transmission range. After this anchor node broadcasts the ID of the destination node, all the other anchor nodes will check their lists to find the destination node. If one anchor node finds the destination node, the source can obtain the information about it.

After node S getting the information of the destination D, the second step is determining the next hop for this transmission. In the beginning, node S broadcasts “destination location” request. As shown in Fig. 1. If node D is in the transmission range of S, then D replies to S before S directly transmitting the message to D. Otherwise, no node replies to S, and node S broadcasts the “moving direction” request. All the information of nodes in the transmission range of S is collected by S through directly communicating with these nodes. As shown in Fig. 2, node A’s moving direction is same as the message’s transmission direction, so node A replies to S and the message is delivered to A immediately. That is to say, A becomes the best next hop. Note that the moving direction information can be easily calculated by the location information at different times. If two nodes have the same moving direction as the message’s transmission direction, then the node with higher speed can only become the next hop. If all the moving directions of nodes in the transmission range of S are different from the message’s transmission direction, then no node replies to S. Therefore, node S will wait.

Finally, the best next hop is decided. The message is delivered and stored in this intermediate node, which continues to determine the next hop until the message successfully arriving at the destination node.

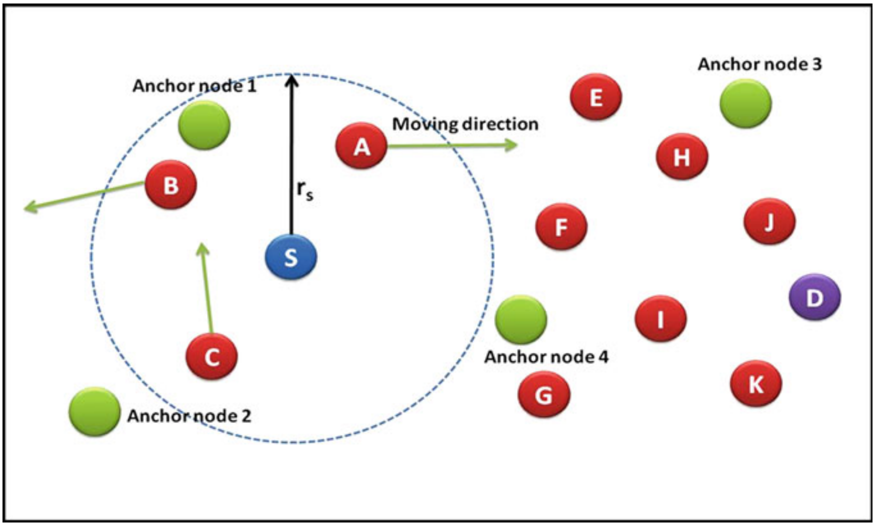


Fig. 2 Node A replies to node S

Simulations and Results

Simulation Environment

We implemented LARP by using the ns-2 simulator [12]. IEEE 802.11 [13] Medium Access Control (MAC) protocol is implemented. We model 50 nodes (including 10 % anchor nodes) in a square area 1,000 × 1,000 m during the simulation time 1,000 s. Each node moves with a speed uniformly distributed between 0 and 5 m/s. The radio transmission range is assumed to be 250 m and a two ray ground propagation channel is assumed. Most other parameters use ns-2 defaults. Three performance metrics of packet delivery ratio, delivery delay and normalized routing overhead are compared. For measuring the three metrics, two simulation factors of the pause time and the transmission rate are varied in a meaningful range.

Results and Discussion

We present a comparative simulation analysis of LARP with DSDV, AODV and ALS. We first analyze the packet delivery ratio. As shown in Fig. 3, the packet delivery ratio gradually increases as the pause time increasing. The curve of LARP shows that the packet delivery ratio persistently increases as the pause time raising. Figure 4 describes the change of packet delivery ratio as the transmission rate

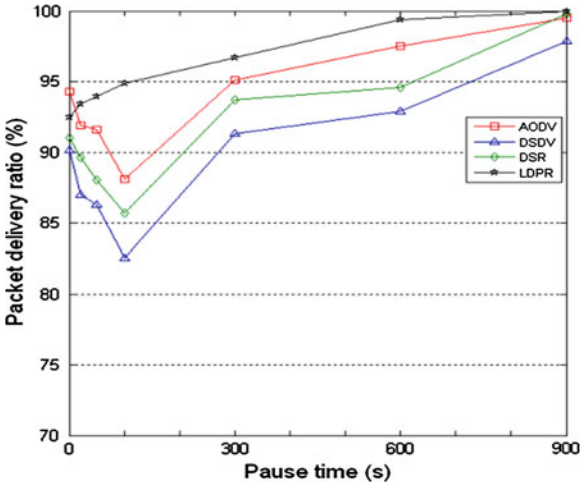


Fig. 3 Packet delivery ratio versus pause time

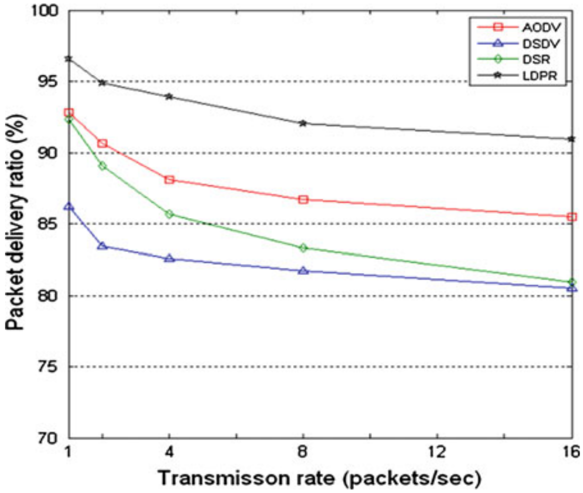


Fig. 4 Packet delivery ratio versus transmission rate

increasing. The packet delivery ratio reduces as the transmission rate raising. Here, LARP performs the highest packet delivery ratio under various transmission rates.

Another critical aspect we investigated is the normalized routing overhead. Fig. 5 shows that the routing overhead decreases with the pause time increasing. In particular, the change of LARP's curve is small. Note that when the pause time is more than 300 s, the routing overhead of LARP is a little higher than AODV and ALS. That's because the simulation environment trends to be static and the mobility of node declines. Figure 6 presents that the routing overhead increases as the transmission rate increasing. In addition, seen from the shape of the LARP's curve, the routing overhead of LARP always maintains a low level.

Fig. 5 Normalized routing overhead versus pause

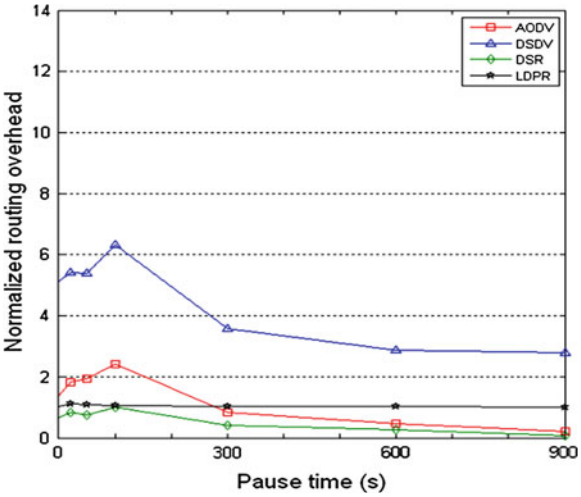
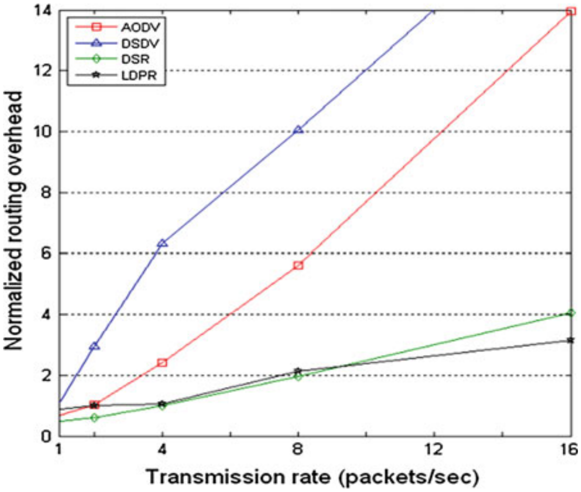


Fig. 6 Normalized rout-ing overhead versus trans-mission rate



It is still of interest to consider the end-to-end delay. Figures 7 and 8 show that the packet delivery delay of LARP is longer than other three routing protocols. In LARP, nodes are required enough time to obtain the information of location. The feature of packet delivery delay in LARP determines that LARP can be only implemented in the environment which focuses on the validity and integrity of the message rather than delivery delay.

Fig. 7 End-to-End delay versus pause time

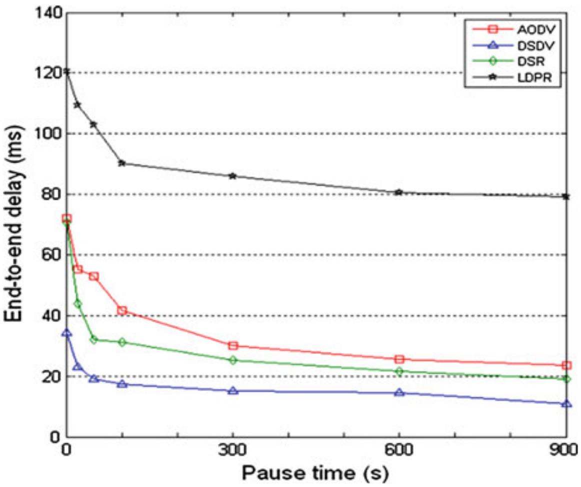
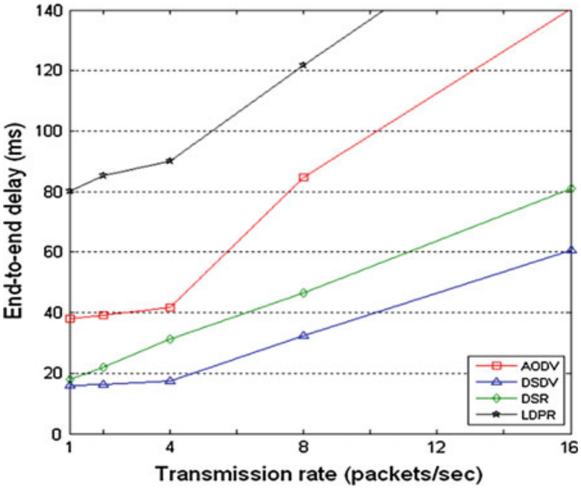


Fig. 8 End-to-End delay versus transmission rate



Conclusions

In this paper, we have proposed a routing protocol named LARP for UWSNs, which utilizes the location information of nodes to transmit a message. Resort to a range-finding technique RSSI, a node can easily obtain its location information. The simulation experiments have shown that LARP is able to ensure the reliability of message transmission. It is worth noting that LARP can be implemented in the environment which focuses on the reliability and validity of message transmission rather than delivery delay.

Acknowledgments This work was supported by the research fund from Nanjing University of Information Science and Technology under Grant No. S8113003001 and Natural Science Foundation of Jiangsu Province under Grant No. BK2011825.

References

1. Akyildiz F, Pompili D, Melodia T (2004) Underwater acoustic sensor networks: research challenges
2. Tanenbaum, A.S.: Computer networks. Fouth edition, pp.104 (2003)
3. Proakis JG, Rice JA, Sozer EM, Stojanovic M (2003) Shallow water acoustic networks. Encyclopedia of Telecommunications
4. Rahman, R.H., Benson, C., Frater, M.: Routing protocols for underwater ad hoc networks. OCEANS, 2012 - Yeosu, pp. 1-7 (2012)
5. Yang X, Ong KG, Dreschel WR, Zeng K, Mungle CS, Grimes CA (2002) Design of a wireless sensor network for long-term, in situ monitoring of an aqueous environment. Sensors, pp 455–472
6. He, T.: Range-free Localization Schemes for Large Scale Sensor Networks. Proceedings of the 9th ACM International Conference on Mobile Computing and Networking (Mobicom2003) (2003)
7. Cordeiro CM, Agrawal DP (2006) Ad Hoc and sensor networks, pp 23–24
8. Perkins CE, Bhagwat P (1994) Highly dynamic Destination-Sequenced Distance-Vector routing (DSDV) for mobile computers. ACM SIGCOMM Commun Rev
9. Perkins C, Belding-Royer E, Das S (2003) Ad hoc On-Demand Distance Vector (AODV) routing. Internet RFCs
10. Chandrasekhar V, Seah WKG (2006) Localization in underwater sensor networks: survey and challenges. In: International conference on mobile computing and networking, pp 33–40
11. Chandrasekhar V, Seah WKG (2006) Area Localization Scheme for Underwater Sensor Networks. Proceedings of the IEEE OCEANS Asia Pacific Conference (2006)
12. CMU Monarch Project. The CMU Monarch Project's wireless and mobility extensions to ns. ftp.monarch.cs.cmu.edu/pub/monarch/wireless-sim/ns-cmu.ps (1999)
13. IEEE Computer Society. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications (1997)

Efficient Key Management Scheme for SCADA System

Jian Shen, Jin Wang, Yongjun Ren, Jianwei Zhang and Shunfeng Wang

Abstract Currently Supervisory Control And Data Acquisition (SCADA) system intends to be connected to the open operating environment. Thus, protecting SCADA systems from malicious attacks is getting more and more attention. A key management scheme is essential for secure SCADA communications. In this paper, we propose an efficient key management scheme for SCADA systems with good security properties and performance.

Keywords Supervisory control and data acquisition (SCADA) · Key management · Secure SCADA communications

Introduction

In order to deliver critical services, such as water, sewerage and electricity distribution, nations are increasingly depends on Supervisory Control And Data Acquisition (SCADA) systems. As the change of the operating environment in SCADA system from close to open, the risk of SCADA incidents occurring is increasing. Nowadays, SCADA system has been exposed to a wide range of network security problems. If the SCADA system is damaged from the attacks,

J. Shen (✉) · J. Wang · Y. Ren
School of Computer and Software, Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Science and Technology, Nanjing 210044, China
e-mail: s_shenjian@126.com

J. Zhang
School of Mathematics and Statistics, Nanjing University of Information Science
and Technology, Nanjing 210044, China

S. Wang
College of Bin Jiang, Nanjing University of Information Science and Technology,
Nanjing 210044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_70, © Springer Science+Business Media Dordrecht 2014

this system can have a widespread negative effect to society. One critical security requirement for SCADA systems is that communication channels need to be secured. Secure keys need to be established before cryptographic techniques can be used to secure communications.

Note that un-encrypted data communication via networks is vulnerable to several types of attacks. Therefore, secure data communication between each device is required to secure the SCADA system. Secure key management is essential for data encryption. In this paper, we focus on the key management scheme for SCADA systems and propose an efficient key management scheme (EKMS) with good security properties. Compared with the previous schemes, the presented key management scheme is more efficient in terms of the communication cost. Our scheme is based on a symmetric balanced incomplete block design (SBIBD), which can provide the authentication service and resist different key attacks. The structure of SBIBD makes the computation of a common conference key for each remote terminal unit (RTU) quite convenient.

The rest of this paper is organized as follows: In the following section, related work is briefly introduced. The proposed key management scheme is described in detail in section Efficient Key Management Scheme for SCADA System. Security analysis and performance analysis of our scheme are presented in section Security Analysis and Performance Analysis. Finally, the conclusions of this paper are covered in section Conclusions.

Related Work

A SCADA system consists of three types of equipment communicating with each other: (1) human–machine interface (HMI) that operators interact with; (2) master terminal unit (MTU) that provides supervisory control of an RTU; and (3) the remote terminal unit (RTU) that interacts with the physical environment. In this paper, the term node will be used to refer to any entity in the system. The structure of SCADA systems is based on master–slave structure, which is shown in Fig. 1. The structure of a SCADA system will normally include one central MTU, which communicates with a hierarchy of other nodes, including Sub-MTU and RTUs. Master stations and sub-master stations, are computers with resources at least as plentiful as a modern desktop computer.

Recently, SKE [1] was proposed by Sandia, where the MTU has to encrypt data with each key of the RTUs individually to broadcast a message. After that, SKMA [2] was proposed, where two types of keys must be managed by an MTU or RTUs. The long-term node-key distribution center (KDC) key is shared between the KDC and a node. The other key is the long-term node–node key shared between two nodes. Later, ASKMA [3] proposed a key-management scheme suitable for secure SCADA communication using a logical key hierarchy to support broadcast communication and multicast communication, but it may be less efficient.

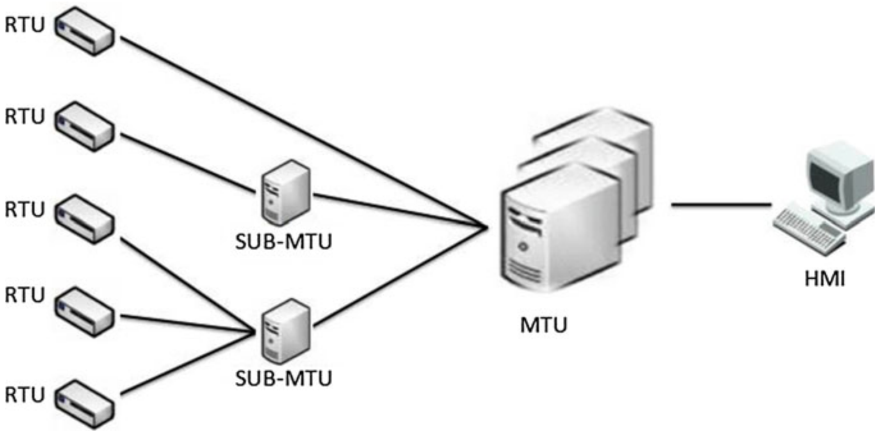


Fig. 1 SCADA system architecture

Due to the constrains of low-rate data transmission and real-time processing in different operational environment, satisfying the security requirements of confidentiality, integrity and availability in a SCADA system is really a challenging issue. In this paper, resort to a symmetric balanced incomplete block design (SBIBD), we design a novel key management scheme for SCADA systems with good security properties and performance.

Efficient Key Management Scheme for SCADA System

In this section, we propose an efficient key management scheme for SCADA system. By our scheme, the communication among Sub-MTUs can be secure and efficient, so can the communication among RTUs. The process of key management among RTUs is described as follows. Note that the process of key management among Sub-MTUs is similar to that of RTUs.

Each RTU registers to the Sub-MTU and gets their private key. After that, every RTU can process the key agreement to compute the common conference key. First of all, the Sub-MTU chooses two prime order group G_1 and G_2 and a modified Weil pairing map $\widehat{e}$ defined in [4]. Next, the Sub-MTU selects two one-way hash functions $H : \{0,1\}^* \to G_1$ and $h : \{0,1\}^* \to Z_q^*$ where H maps its arbitrary length to a nonzero point of G_1 while h maps its input with arbitrary length to a nonzero integer. At last, the Sub-MTU picks a random integer $s \in Z_q^*$ as its private key, computes its public key $P_{pub} = s\mathcal{G}$, and publishes $\left(p, q, G_1, G_2, \mathcal{G}, \widehat{e}, P_{pub}, H, h\right)$, but keeps s secret. Each RTU U_i's identity is $ID_i \in (0,1)^*$. The Sub-MTU computes U_i's public key $Q_i = H(ID_i)$ and then U_i's private key $S_i = sQ_i$ which is issued to U_i via a secure channel.

Fig. 2 (7 × 7) incidence matrix corresponding to the (7, 4, 2)-design

$$\mathcal{L} = (\ell_{ij}) = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The common conference key among RTUs is calculated by employing SBIBD, where the number of blocks is the same as that of participants. We choose a (7, 4, 2)-design. Let a finite set X $= \{1, 2, 3, 4, 5, 6, 7\}$, then $B_1 = \{1, 2, 4, 7\}$, $B_2 = \{1, 2, 3, 5\}$, $B_3 = \{2, 3, 4, 6\}$, $B_4 = \{3, 4, 5, 7\}$, $B_5 = \{1, 4, 5, 6\}$, $B_6 = \{2, 5, 6, 7\}$, $B_7 = \{1, 3, 6, 7\}$. Accordingly, a (7 × 7) incidence matrix L is depicted in Fig. 2. The rows and columns of the matrix correspond to the blocks and the elements, respectively. The entry l_{ij} in the ith row and the jth column of L is a 1 if the block i contains the element j and is a 0 otherwise.

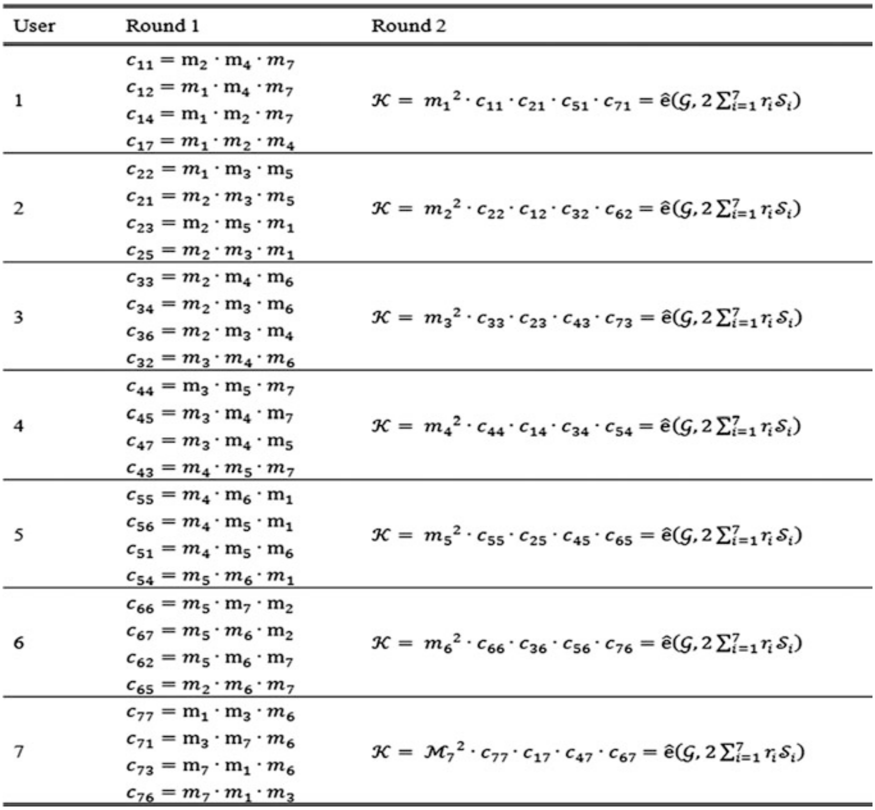
For computing the common conference key among RTUs, two rounds are required in our scheme.

1. Each RTU U_i selects a random number r_i as secret key by itself for every session and then calculates $m_i = \widehat{e}(\mathcal{G}, r_i S_i)$. Simultaneously, U_i calculates $T_i = r_i Q_i$. Let $D_i = \{m_i, T_i\}$. RTU i receives message D_j from RTU j in case $l_{ij} = 1$ and $j \neq i$, namely $j \in B_i - \{i\}$. m_i is used for generating conference key while T_i is used for authentication. We now describe the key agreement process from the viewpoint of RTU 1. U_1 receives D_2, D_4, D_7 from U_2, U_4, U_7 and then makes

$$\begin{aligned} c_{11} &= m_2 \cdot m_4 \cdot m_7 = \widehat{e}(\mathcal{G}, r_2 S_2 + r_4 S_4 + r_7 S_7), \\ c_{12} &= m_1 \cdot m_4 \cdot m_7 = \widehat{e}(\mathcal{G}, r_1 S_1 + r_4 S_4 + r_7 S_7), \\ c_{14} &= m_1 \cdot m_2 \cdot m_7 = \widehat{e}(\mathcal{G}, r_1 S_1 + r_2 S_2 + r_7 S_7), \\ c_{17} &= m_1 \cdot m_2 \cdot m_4 = \widehat{e}(\mathcal{G}, r_1 S_1 + r_2 S_2 + r_4 S_4), \\ W_{12} &= T_1 + T_4 + T_7, \\ W_{14} &= T_1 + T_2 + T_7, \\ W_{17} &= T_1 + T_2 + T_4, \end{aligned}$$

where $c_{ij} = \prod_{x \in B_i - \{j\}} m_x$ and $W_{ij} = \sum_{x \in B_i - \{j\} \text{ and } j \neq i} T_x$. In the viewpoint of RTU 1, we have that $c_{1j} = \prod_{x \in B_1 - \{j\}} m_x$ and $W_{1j} = \sum_{x \in B_1 - \{j\} \text{ and } j \neq 1} T_x$. Simultaneously, other RTUs do the same process.

2. Let $E_{ji} = \{c_{ji}, W_{ji}\}$. RTU i receives E_{ji} from RTU j in case $l_{ji} = 1, j \neq i$. Here, similar to that in round 1, c_{ji} is used for generating conference key while W_{ji} is used for authentication. Particularly, U_1 receives E_{j1} from RTU j, if $l_{j1} = 1$,



User	Round 1	Round 2
1	$c_{11} = \mathrm{m}_2 \cdot \mathrm{m}_4 \cdot m_7$ $c_{12} = m_1 \cdot \mathrm{m}_4 \cdot m_7$ $c_{14} = \mathrm{m}_1 \cdot \mathrm{m}_2 \cdot m_7$ $c_{17} = m_1 \cdot m_2 \cdot m_4$	$\mathcal{K} = m_1{}^2 \cdot c_{11} \cdot c_{21} \cdot c_{51} \cdot c_{71} = \hat{e}(\mathcal{G}, 2\sum_{i=1}^{7} r_i \mathcal{S}_i)$
2	$c_{22} = m_1 \cdot \mathrm{m}_3 \cdot \mathrm{m}_5$ $c_{21} = m_2 \cdot m_3 \cdot m_5$ $c_{23} = \mathrm{m}_2 \cdot \mathrm{m}_5 \cdot m_1$ $c_{25} = m_2 \cdot m_3 \cdot m_1$	$\mathcal{K} = m_2{}^2 \cdot c_{22} \cdot c_{12} \cdot c_{32} \cdot c_{62} = \hat{e}(\mathcal{G}, 2\sum_{i=1}^{7} r_i \mathcal{S}_i)$
3	$c_{33} = m_2 \cdot \mathrm{m}_4 \cdot \mathrm{m}_6$ $c_{34} = m_2 \cdot \mathrm{m}_3 \cdot \mathrm{m}_6$ $c_{36} = m_2 \cdot \mathrm{m}_3 \cdot \mathrm{m}_4$ $c_{32} = m_3 \cdot m_4 \cdot m_6$	$\mathcal{K} = m_3{}^2 \cdot c_{33} \cdot c_{23} \cdot c_{43} \cdot c_{73} = \hat{e}(\mathcal{G}, 2\sum_{i=1}^{7} r_i \mathcal{S}_i)$
4	$c_{44} = \mathrm{m}_3 \cdot \mathrm{m}_5 \cdot m_7$ $c_{45} = m_3 \cdot \mathrm{m}_4 \cdot \mathrm{m}_7$ $c_{47} = m_3 \cdot \mathrm{m}_4 \cdot \mathrm{m}_5$ $c_{43} = m_4 \cdot m_5 \cdot m_7$	$\mathcal{K} = m_4{}^2 \cdot c_{44} \cdot c_{14} \cdot c_{34} \cdot c_{54} = \hat{e}(\mathcal{G}, 2\sum_{i=1}^{7} r_i \mathcal{S}_i)$
5	$c_{55} = m_4 \cdot \mathrm{m}_6 \cdot \mathrm{m}_1$ $c_{56} = m_4 \cdot \mathrm{m}_5 \cdot \mathrm{m}_1$ $c_{51} = m_4 \cdot \mathrm{m}_5 \cdot \mathrm{m}_6$ $c_{54} = m_5 \cdot m_6 \cdot m_1$	$\mathcal{K} = m_5{}^2 \cdot c_{55} \cdot c_{25} \cdot c_{45} \cdot c_{65} = \hat{e}(\mathcal{G}, 2\sum_{i=1}^{7} r_i \mathcal{S}_i)$
6	$c_{66} = m_5 \cdot \mathrm{m}_7 \cdot \mathrm{m}_2$ $c_{67} = m_5 \cdot m_6 \cdot \mathrm{m}_2$ $c_{62} = m_5 \cdot \mathrm{m}_6 \cdot \mathrm{m}_7$ $c_{65} = m_2 \cdot m_6 \cdot m_7$	$\mathcal{K} = m_6{}^2 \cdot c_{66} \cdot c_{36} \cdot c_{56} \cdot c_{76} = \hat{e}(\mathcal{G}, 2\sum_{i=1}^{7} r_i \mathcal{S}_i)$
7	$c_{77} = \mathrm{m}_1 \cdot \mathrm{m}_3 \cdot m_6$ $c_{71} = \mathrm{m}_3 \cdot \mathrm{m}_7 \cdot m_6$ $c_{73} = \mathrm{m}_7 \cdot \mathrm{m}_1 \cdot m_6$ $c_{76} = m_7 \cdot m_1 \cdot m_3$	$\mathcal{K} = \mathcal{M}_7{}^2 \cdot c_{77} \cdot c_{17} \cdot c_{47} \cdot c_{67} = \hat{e}(\mathcal{G}, 2\sum_{i=1}^{7} r_i \mathcal{S}_i)$

Fig. 3 Generating a common key

$j \neq 1$. Therefore, U_1 receives E_{21}, E_{51}, E_{71} from U_2, U_5, U_7 and derives c_{21}, c_{51}, c_{71}. Then the common conference key K is calculated as $K = m_1 \times c_{11} \times c_{21} \times c_{51} \times c_{71} = \widehat{e}\left(\mathcal{G}, 2\sum_{i=1}^{7} r_i S_i\right)$, where $c_{21} = \widehat{e}(\mathcal{G}, r_2S_2 + r_3S_3 + r_5S_5)$, $c_{51} = \widehat{e}(\mathcal{G}, r_4S_4 + r_5S_5 + r_6S_6)$, and $c_{71} = \widehat{e}(\mathcal{G}, r_3S_3 + r_7S_7 + r_6S_6)$.

Then, following our scheme, the process for calculating the common conference key among all the RTUs is shown in Fig. 3.

In our scheme, we take advantage of RTUs' identity information for authentication.

1. Let $D_i = \{m_i, T_i\}$, RTU i receives D_j from RTU j in case $l_{ij} = 1$ and $j \neq i$. We now describe the authentication process from the viewpoint of RTU 1. U_1 receives D_2, D_4, D_7 from U_2, U_4, U_7 and makes

$$\begin{aligned}
\widehat{e}\left(P_{pub}, T_2\right) &= \widehat{e}(s\mathcal{G}, r_2 Q_2) = \widehat{e}(\mathcal{G}, r_2 s Q_2) = m_2,\\
\widehat{e}\left(P_{pub}, T_4\right) &= \widehat{e}(s\mathcal{G}, r_4 Q_4) = \widehat{e}(\mathcal{G}, r_4 s Q_4) = m_4,\\
\widehat{e}\left(P_{pub}, T_7\right) &= \widehat{e}(s\mathcal{G}, r_7 Q_7) = \widehat{e}(\mathcal{G}, r_7 s Q_7) = m_7,
\end{aligned}$$

Hence, U_1 can authenticate the entity of U_2, U_4, U_7 only if $\widehat{e}\left(P_{pub}, T_2\right) = m_2$, $\widehat{e}\left(P_{pub}, T_4\right) = m_4$, and $\widehat{e}\left(P_{pub}, T_7\right) = m_7$, respectively. Generally speaking, if $\widehat{e}\left(P_{pub}, T_i\right) = m_i$, then U_j can authenticate counterpart's entity.

2. Let $E_{ji} = \{c_{ji}, W_{ji}\}$ and $W_{ji} = \sum_{x \in B_j - \{i\}\,\text{and}\, j \neq i} T_x$. RTU i receives E_{ji} from RTU j in case $l_{ji} = 1$, $j \neq i$. Particularly, in the viewpoint of RTU 1, $W_{j1} = \sum_{x \in B_j - \{1\}\,\text{and}\, j \neq 1} T_x$ and $E_{j1} = \{c_{j1}, W_{j1}\}$. U_1 receives E_{21}, E_{51}, E_{71} from U_2, U_5, U_7, then derives W_{21}, W_{51}, W_{71} and calculates

$$\begin{aligned}
\widehat{e}(P_{pub}, W_{21}) &= \widehat{e}(s\mathcal{G}, T_2 + T_3 + T_5) = m_2 \cdot m_3 \cdot m_5 = c_{21},\\
\widehat{e}(P_{pub}, W_{51}) &= \widehat{e}(s\mathcal{G}, T_4 + T_5 + T_6) = m_4 \cdot m_5 \cdot m_6 = c_{51},\\
\widehat{e}(P_{pub}, W_{71}) &= \widehat{e}(s\mathcal{G}, T_3 + T_7 + T_6) = m_3 \cdot m_7 \cdot m_6 = c_{71},
\end{aligned}$$

Therefore, the RTU of U_2, U_5, U_7 can pass the authentication by U_1 only if $\widehat{e}(P_{pub}, W_{21}) = c_{21}$, $\widehat{e}(P_{pub}, W_{51}) = c_{51}$, $\widehat{e}(P_{pub}, W_{71}) = c_{71}$, respectively. Broadly speaking, if $\widehat{e}(P_{pub}, W_{ji}) = c_{ji}$, then U_i can authenticate counterpart's entity.

Security Analysis and Performance Analysis

A passive adversary tries to learn information about the conference key by eavesdropping on the broadcast channel. We show that an eavesdropper cannot get any information about the secret key r_i of U_i due to Weil Diffie-Hellman (WDH) problem [5] in $(G_1, G_2, \widehat{e})$ and discrete algorithm problem (DLP) in elliptic curves. In active attack, an adversary not only just records the data, but also can alter, inject, intercept and replay messages. Our protocol can be able to provide the authentication service by sending a special message T_i and W_{ji} in first round and second round, respectively. Our scheme has the security properties of known session key security, perfect forward secrecy, key-compromise impersonation resistance and no key control.

The communication cost of previous schemes are all $O(n^2)$, while the communication cost of our scheme is only $O(n\sqrt{n})$even though the communication round is 2.

Conclusions

SCADA system is a significantly important system that plays a very important role in national infrastructure, such as electric grids and water supplies. However, SCADA system is becoming increasingly vulnerable to adversarial manipulation due to the extreme operational environment. In this paper, we present a novel key management scheme for SCADA systems with good performance and security properties. We believe that our scheme must be promising in the secure communication in SCADA system in the future.

Acknowledgments This work was supported by the research fund from Nanjing University of Information Science and Technology under Grant No. S8113003001, National Science Foundation of China under Grant No. 61272421.

References

1. Beaver C, Gallup D, Neumann W, Torgerson M (2002) Key Management for SCADA. Available: http://www.sandia.org/ scada/documnets/013252.pdf
2. Colin RD, Boyd C, Manuel J, Nieto G (2006) SKMAA key management architecture for SCADA systems. In: 4th Australasian information security workshop, pp 138–192
3. Choi D, Kim H, Won D, Kim S (2009) Advanced key management architecture for secure SCADA communications. IEEE Trans Power Del 24(3):1154–1163
4. Boneh D, Franklin M (2001) Identity-based encryption from weil pairing. In: Advances in Cryptology-CRYPTO01, Lecture Notes in Computer Science, vol 2139, pp 213–229
5. Kim Y, Perrig A, Tsudik G (2004) Group key agreement efficient in communication. IEEE Trans Comput 53(7):905–921

Part VIII
Smart System

Dynamic Migration Technology Platform in the Cloud Computer Forensics Applied Research

Lijuan Yang and Shuping Yang

Abstract This paper presents a new cloud computing environment, computer forensics, and migration of a virtual machine image file, the virtual machine images from cloud computing platform migration to the local forensics environment analysis, the cloud computing platform, electronic evidence and the performance and application of experimental analysis comparing the two migration strategy.

Keywords Dynamic migration · Cloud computing · Computer forensics · Virtual machine

Introduction

Computer Forensics Analysis of the computer technology, computer crime analysis to identify criminals and computer evidence, computer intrusion and criminal evidence acquisition, preservation, analysis and production, and proceedings accordingly, is an invasion the computer system for scanning and crack invasion reconstruction process. Businesses and individual users can achieve the free sharing of information by a massive repository of cloud computing, forensic technology effectively find the necessary means to establish whether violations, but the traditional paper-based forensics longer meet the cloud computing service model, forensic work to become an important issue in the cloud computing environment.

Use of data migration technology to cloud computing environment for forensic work possible, by Forensics modeling of cloud computing, cloud computing platform as the system into multiple virtual body, running virtual machine instance

L. Yang (✉) · S. Yang
School of Computer Engineering, Shenzhen PolyTechnic, Shenzhen 518055, China
e-mail: yylljj928@szpt.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_71, © Springer Science+Business Media Dordrecht 2014

as the object of forensic analysis, virtual machine instances in the virtualization software layer security, load the virtual machine image file in the localized system for forensic analysis, the use of a separate division of the temporary image file partition as the exchange of information between the image file and localization system places the virtual machine image file can be loaded correctly, the scene forensics work under the cloud computing environment [1].

Computer Forensic Model of the Cloud Platform

Cloud computing platform provider of computing resources for effective resource allocation and access control, and provide end users with a variety of cloud services, virtualization technology of distributed computer resources are managed and integrated into the virtual host resources available to the user cloud platform virtualization features enable users do not have to care about the location of the real host, maintenance and fault tolerance, to help the user to get out from the computer hardware and software resources management burden.

Computer Forensics by function can be defined as five levels, namely the discovery layer, evidence of the fixed layer, abstraction layer of evidence, evidence analysis layer and evidence presentation layer, the division of the five levels together make up the hierarchical model of computer forensics. The forensic model is divided into basic service layer, the mirror migration evidence for layer, the evidentiary regulators, as shown in Fig. 1.

Basic service layer corresponds to the layer of the discovery of evidence. This layer is the place of cloud computing services, a variety of cloud computing services through the basic service layer are classified, organized into mutual contact multiple cloud computing services sequence, the sequence of these services constitute a virtual machine instance. Distinction between cloud computing services, forensics center of gravity shifted to the virtual machine instance, while avoiding the complexity of the network and the dispersion of the cloud structure to address the relevance of the evidence [2].

The mirror migration evidence for layer corresponding evidence abstraction layer is a major component of cloud computing forensic model, the software layer divided according to the mutual isolation of virtual machine instances, the use of the virtual machine instances interfere with each other, clearly forensics object and forensics range.

Forensics evidence regulators correspond to the fixed layer, scheduling and monitoring the mirror evidence abstraction layer every migration operation, and record the migration process to ensure that the migration process of the Strategies and execution control and to generate regulatory process reported Reports, to ensure the reliability of forensic data and legal integrity.

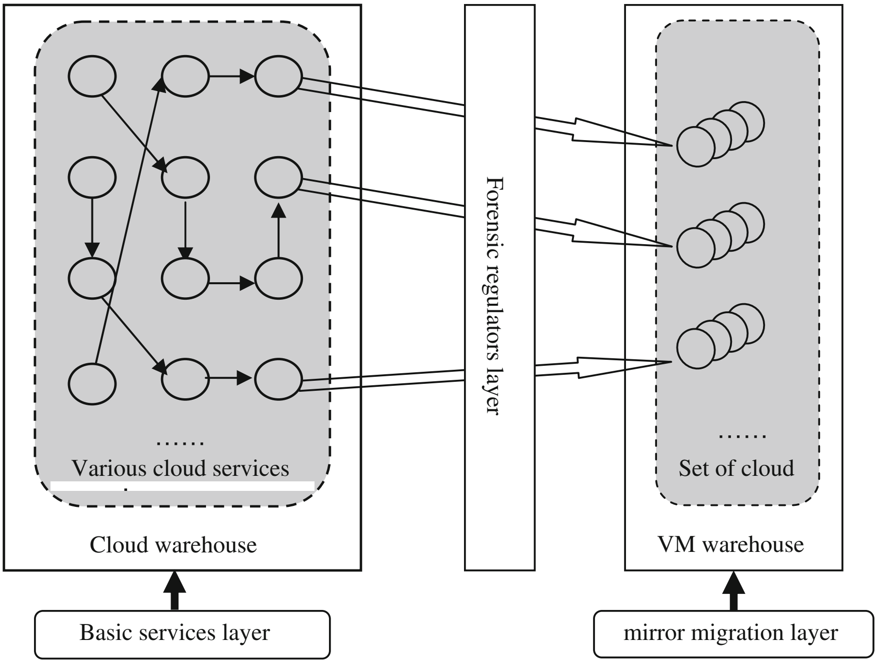


Fig. 1 Cloud platform computer forensics model framework

Dynamic Migration Technology

In the cloud computing platform built model of computer forensics, migration hardware configuration and system migration, live migration technology, the use of the operating system. Migration process needs to be saved and remodeling of the operating system, including the process of identification, information reconstruction of the file system, memory migration, network connection information monitoring, as well as many aspects of the concrete realization of registry information reconstruction. And through combined and feedback mechanisms, design data migration system architecture, data migration process will not affect the entire cloud computing platform service request. The different effects of different migration strategies of response systems, these migration strategies can be adjusted according to the working status of the migration [3].

Hardware Abstraction Layer Virtual Machine Technology

Hardware Abstraction Layer virtual machine technology is the use of an intermediate layer to isolate code running environment, virtual machine monitor layer to run on the operating system provides a variety of customer hardware response, mapping the physical device. The virtual machine monitor layer is able to create a copy of the system isolated from each other, each copy of the system shared physical device are isolated from each other, are not adversely affected, can play to the highest utilization of the physical device, the virtual machine copy of the system is concerned, this computer forensics analysis object of concern [4].

Preservation and Reconstruction of Virtual Machine Migration Process

Can achieve data migration of virtual machine instances in a virtual machine monitor layer, the virtual machine instance is saved as an image file, to facilitate the latter part of the environmental restoration and forensic analysis. The migrated data comprises a variety of sensitive data [5]:

- The page table entries and directory entries from a writable mode switch to read-only mode;
- Virtual machine's thread priority of the control module, switch from high-level to low-level, lower privileges;
- Virtualization software layer to migrate virtual machine scheduling data;
- Virtualization software layer calls to the underlying hardware of the virtual machine you want to migrate.

The entire virtual machine migration process under the state protection of sensitive data, the need to have four key modules, Is the process ID information module, memory mapping information module, the network connection module, file system information Module, in order to complete preservation of these data, to facilitate the full migration of the virtual machine instance, these four related The key part of the remodeling and save information the entire virtualization software layer module structure shown in Fig. 2 [6].

The Two Modes of Conversion of the Migration Process

Cloud computing platform virtualization software layer module add, the definition of a unified virtual communication interface, can be achieved in two modes of action, namely the ordinary cloud computing service model and cloud services migration patterns [7].

In normal mode, the cloud computing platform to provide normal cloud computing services, response to the application of the service request. Migration patterns, the virtualization software layer as multiple modules to ensure the consistency of the state of the operating system to load and record the current process state of the virtual machine system, the memory mapping, network connection and

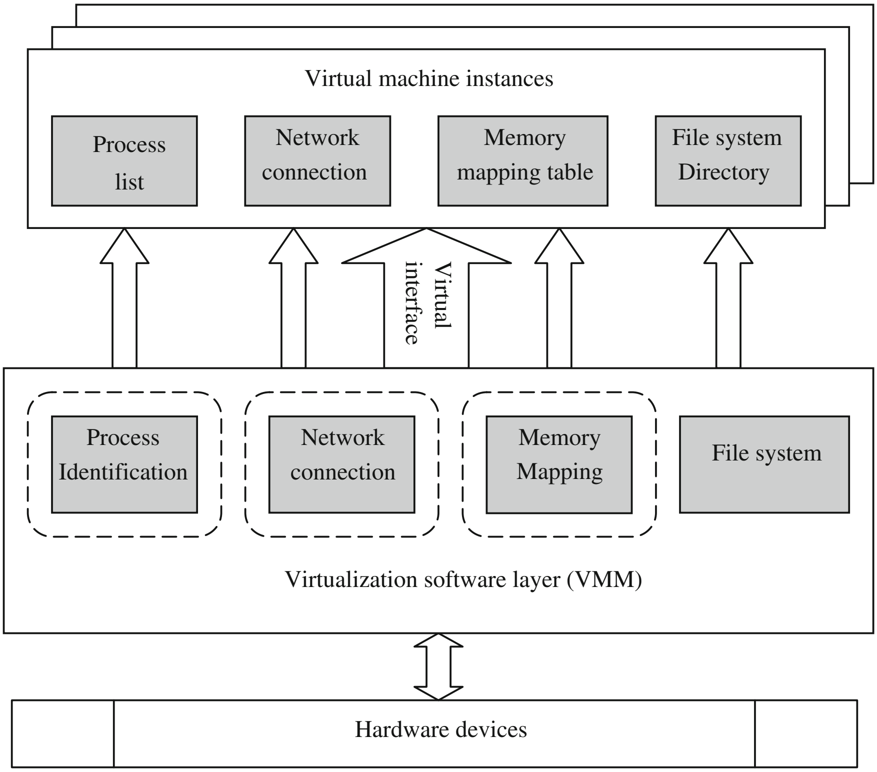


Fig. 2 Virtualization software layer module functional block

file system modifications change, Preparing Virtual Machines the mirrored data migration [8]. Data the migration scenarios migration task adaptive scheduling algorithm flow as shown in Fig. 3, the specific implementation process [9]:

Pl: The workload threshold value read from the migration task queue the migration task requests, set up to monitor the sampling period T and storage components $W_{\max}$, Counter K = 1, Take the initial migration rate $Migrate\_rate(0) = 0$;

P2: According to the migration strategy *MIGRATE_POLICY*, if the migration strategy is $PRIOR\_LEVEL[0]$, confers on the $Migrate\_rate(k)$ value is *MAX_RATE*, no rate limiting, turn P4, cycle $(k-1)*T$ to $K*T$. The time during which a load feedback $W_f = Feedback\_load(t)$, calculate the actual load and the load threshold difference $E(k) = p \times W_{\max} - W_f$, among $P(0 < P < 1)$ is reserve coefficient.

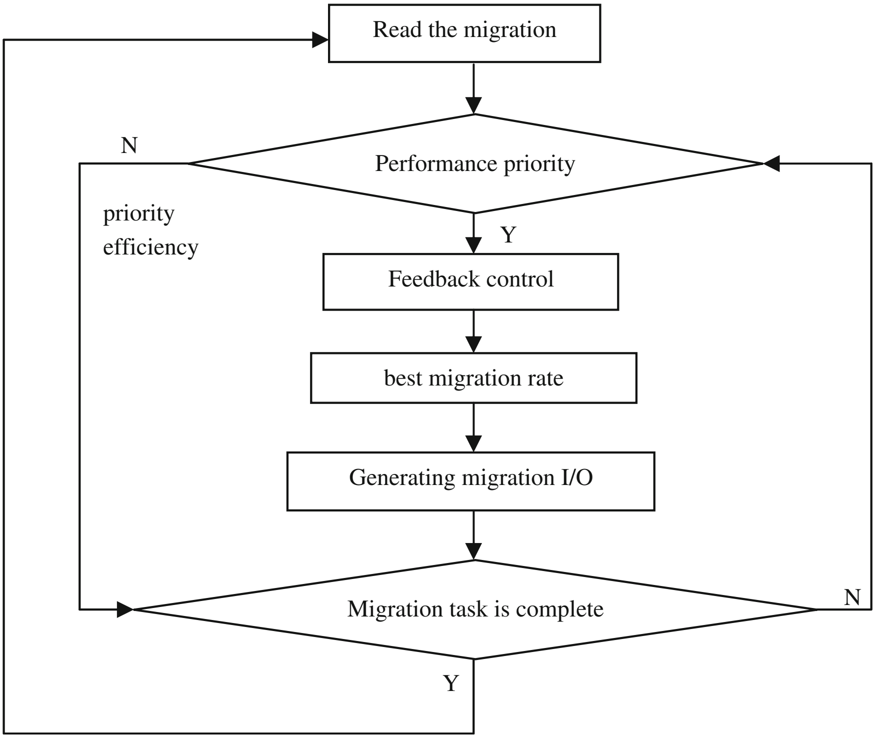


Fig. 3 Migration scheduling algorithm flowchart

P3: The difference obtained by the previous step, calculate the optimum value of the rate of migration of the next cycle:: $Migrate\_rate(k) = Migrate\_rate(k-1) + a \times E(k)$, K is the adjustment coefficient.

P4: $Migrate\_rate(k)$ send data migration generated migrate I/O, this procedure stored in the member receiving the sampling of the monitoring module, and based on the member state to modify or maintain the current task migration strategy.

P5: If this migration task is completed, turn P1, otherwise the counter K value plus 1, turn P2.

Experiment and Analysis

The Linux platform using the Xen open source virtualization tool to build a cloud computing platform, configure the virtualization monitoring layer, run the virtual machine instances, complete cloud computing platform features, and then use the testing instrument for the Linux platform, by analyzing the dynamic migration process virtual converted to the migration from the service-state mode of monitoring layer state mode, the system state transition time and performance comparison, draw mode changes the impact of cloud computing services platform, followed by the use of different data migration strategy, different migration strategy process in the amount of time the performance impact of cloud computing platform, the conclusion that the mirror migration method of performance evaluation [10, 11].

1. Performance test

Virtualization software layer needs to correspond to a virtual machine system for a recording and reconstruction, when the platform mode conversion from the normal mode to the migration mode takes considerable platform resources, and therefore significantly extend the run time of several test procedures, fork program as it relates to the large number of page table updates, adds to the reconstruction of platform information resource overhead.

2. Time Test

The global file system module using virtualization software layer traversal completed to measure the time course of the mode converter. Switch between normal mode and migration patterns of time longer than the switching time between migration mode and normal mode, migration mode and normal mode conversion services in the Xen test environment under quite small.

3. Data migration process virtual machine performance impact

The data migration process, the performance of the entire cloud computing platform is quite large, because the migration takes up bandwidth, resulting in customer service response is too slow.

4. The performance test results of different migration strategies

The efficiency priority migration strategy for the 1/0 response still have a considerable impact, due to the migration process need to ensure that the normal response of cloud computing services in a complete image file migration process is complete. Takes a large number of 1/0 channel.

The performance priority migration strategy, the migration process duration becomes longer, from the beginning lasted until 3005, than giving priority to efficiency strategy takes a long time, but did not bring the platform I/O response time Much impact.

Different from the experimental data can be concluded that, during the migration process can use different migration strategies, the use of giving priority to efficiency or service priority principle to divide, making migration Cheng Duiyun computing services impact as small as possible, but also in the cloud computing services when the system is idle, the migration of virtual machine images.

References

1. Clark C, Fraser K, Hand S et al (2005) Live migration of virtual machines. In: Proceedings of the 2nd international conference on networked systems design and implementation, Berkeley, CA, USA
2. Wood T (2007) Blaek-box and Gray-box strategies for virtual machine migration. In: Proceedings of the 4th international conference on networked systems design and implementation, IEEE Press
3. LaVelle C, Konrad A (2007) FriendlyRoboCoPy: A GUI to RoboCoPy for computer forensic investigators. Digital Invest 4:16–23
4. Arnold J (2008) Ksplice: An automatic system for rebootless kernel security updates. Ph.D. thesis, Massachusetts Institute of Technology
5. Zhou G, Cao Q, Mai Y (2011) Forensic analysis using migration in cloud computing environment. In: 2nd International conference on intelligent transportation systems and intelligent computing, Suzhou, China
6. Fasheng, based on resource management end-system traffic shaping algorithm (2007). Comput Eng Appl (18):117–119
7. Grossman RL, Yunhong Gu (2009) On the varieties of clouds for data intensive computing. IEEE Data Eng Bull 32(1):44–50
8. Tu WW, Zhang J, Zhang X (2006) Grade allocation token parameter of traffic shaping algorithm. Comput Appl (9):2175–2177
9. Ladan-Mozes E, Shavit N (2008) An optimistic approach to lock-free FIFO queues. Distrib Comput 20(5):323–341
10. Yadav AK, Tomar R, Kumar D, Gupta H (2012) Security and privacy concerns in cloud computing. Comput Sci Softw Eng 2(5)
11. Valenzuela JL (2004) A hierarchical token bucket algorithm to enhance QoS in IEEE 802.11: proposal, implementation and evaluation. In: IEEE 60th vehicular technology conference, 2004. VTC2004-Fall

Research on Parameters of Affinity Propagation Clustering

Bin Gui and Xiaoping Yang

Abstract The affinity propagation clustering is a new clustering algorithm. The volatility is introduced to measure the degree of the numerical oscillations. The research focuses on two main parameters of affinity propagation: preference and damping factor, and considers their relation with the numerical oscillating and volatility, and we find that the volatility can be reduced by increasing the damping factor or preference, which provides the basis for eliminating the numerical oscillating.

Keywords Affinity propagation · Damping factor · Preference · Volatility

Introduction

Affinity propagation (AP) is a new clustering algorithm published in Science magazine and proposed by Frey and Dueck [1]. Its advantage is that it finds clusters with much lower error than other methods, and it does so in less than one-hundredth the amount of time [1].Although results on small data sets (900 ≤points) demonstrate that vertex substitution heuristic (VSH) is competitive with AP, VSH is prohibitively slow for moderate-to-large problems, whereas AP is much faster and could achieve lower error [2].

There are some scholars who have done some improvements on AP algorithm. Literature [3, 4] proposed a new message transmission algorithm soft constraint

B. Gui (✉)
School of Information, Remin University of China, Beijing, China
e-mail: guibin_163@163.com

X. Yang
School of Computer Science and Technology, Huaiyin Normal University, Huaian, China
e-mail: yang@ruc.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_72, © Springer Science+Business Media Dordrecht 2014

condition who allowed a single cluster exemplar did not represented itself in the search and optimization process of the cluster exemplar. Kaijun Wang proposed an algorithm that could adjust adaptively the damping factors and preferences to eliminate oscillations and find the optimal clustering result [5]. Xiao Yu proposed a semi-supervised clustering method based on affinity propagation, it can produce good clustering results for the datasets with complex cluster structures by using the prior known labeled data or pairwise constraints to adjust the similarity matrix [6]. Although there are some research for AP, no one does overall research on AP.

Affinity Propagation Clustering

AP works based on similarities between pairs of data points (Euclidean distance), each similarity is set to negative squared error:

For points x_i and $x_k, s(i,k) = -||x_i - x_k||^2$.Take the negative in order to facilitate the calculation, the greater the value, the higher similarity. These similarity can be symmetric, i.e.,$s(i,k) = s(i,k)$; they can also be asymmetric. i.e.,$s(i,k) \neq s(i,k)$. The similarity between the N data points is composed of N × N similarity matrix S.Rather than requiring that the number of clusters be prespecified, AP takes as input a real number $s(k,k)$ for each data point k so that data points with larger values of $s(k,k)$ are more likely to be chosen as exemplars. These values are referred to as “preferences”. The number of identified exemplars (number of clusters) is influenced by the values of the input preferences. If a prior, all data points are equally suitable as exemplars, the preferences should be set to a common value this value can be varied to produce different numbers of clusters. The shared value could be the median of the input similarities (resulting in a moderate number of clusters) or their minimum (resulting in a small number of clusters). There are two kinds of message exchanged between data points: responsibility and availability. The “responsibility” r(i, k), sent from data point i to candidate exemplar point k, reflects the accumulate evidence for how well-suited point k is to serve as the exemplar for point i, taking into account other potential exemplars for point i. The “availability” a(i, k),sent from candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar, taking into account the support from other points that point k should be an exemplar, and the support should be positive.

AP searches for clusters through an iterative process. The iteration may be terminated after a fixed number of iterations, after the changes in the messages fall below a threshold, or after the local decisions stay constant for some number of iterations. As affinity propagation is an iterative method, a damping factor $\lambda(0 \leq \lambda < 1)$ is introduced to reserve some information about the old values of messages to avoid numerical oscillations that arise in some circumstances. Each message is set λ times its value of the previous iteration plus $1 - \lambda$ times its prescribed updated value. The process of affinity propagation clustering is described as follow:

INPUT: data points $/x_i, i = 1, 2, \ldots, n$
OUTPUT: K cluster centers

1. Initialization: set $r(x_i, x_j) = 0$ and $a(x_i, x_j) = 0$ for all x_i and x_j.Compute similarity to generate the similarity matrix.

$$p(k) = s(k,k) = \underset{i \neq k}{median}\,(s(i,k)) \tag{1}$$

2. Responsibility updates:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\,(a(i,k') + s(i,k')) \tag{2}$$

$$\mathrm{r(k,k) = p(k)\ -\ \max_{j \neq k}\{\ a(k,j)\ +\ s(k,j)\}} \tag{3}$$

3. Availability updates:

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k))\} \tag{4}$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)) \tag{5}$$

4. Iterate (2) and (3) until a fixed number of iterations, or the changes in the messages fall below a threshold, or the local decisions stay constant for some number of iterations.
5. if $/\, r(k,k) + a(k,k) > 0$, then data point k is the center.
6. Data points are assigned to the corresponding cluster center End.

The above algorithm would produce numerical oscillations. The numerical oscillation is the number of cluster centers generated in the iterative process continue to fluctuate. AP adjusts the damping factors to eliminate oscillations, each message is set to $lam\,(0 \leq lam < 1)$ times its value from the previous iteration plus $1 - lam$ times its prescribed updated value. The formula used for this method is the following:

$$R_i = (1 - lam) * R_i + lam * R_{i-1} \tag{6}$$

$$A_i = (1 - lam) * A_i + lam * A_{i-1} \quad (7)$$

According to Eqs. (6) and (7), we can see that R_i and A_{i} are influenced by the lam in each iteration. When lam is small, R_{i} and A_{i} are very different from $R_{\mathrm{i}-1}$ and $A_{\mathrm{i}-1}$, and when lam is large, R_{i} and A_{i} are very close to $R_{\mathrm{i}-1}$ and $A_{\mathrm{i}-1}$.

Research on Affinity Propagation Parameters

In this section, we will study the AP parameters based on experiment result. The experimental data is randomly generated in order to make the conclusions more catholicity. The experimental platform is Matlab 2009a. The hardware environment is Intel CPU PD T4500 2.3G, main memory is 2G.

AP algorithm will produce numerical oscillation in the process of iteration. it is usually too hard for AP to converge. So, it is very important to study to eliminate the numerical oscillation. In order to better reflect the degree of oscillation in the clustering process, we introduced a volatility index. Volatility is an index to measure the fluctuation degree of the underlying assets price, and it is usually expressed by standard deviation. The essence of numerical oscillation is the clustering number varying with time, Therefore, it is appropriate to use volatility to measure the degree of it.

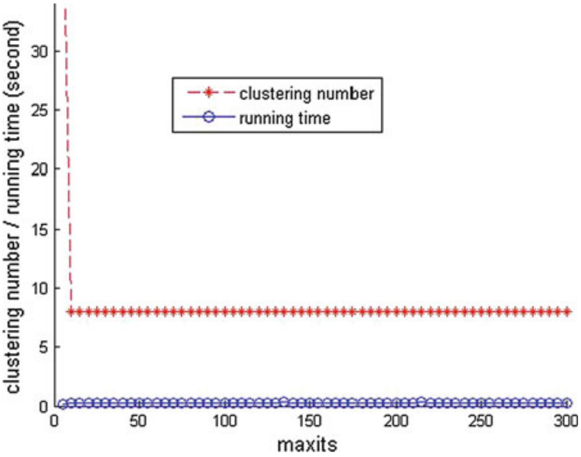
Although AP need not require the number of clusters be prespecified, the final clustering result can not be known exactly. We usually wish it to produce wanted clustering number in practice. It is important to study the relation between clustering number and related parameters in forecasting the AP the clustering number. When the similarity matrix is determined, there are some parameters like preference, damping factor, maxits which could be adjusted in the run-time of AP Algorithm

Maxits Parameter

Maxits means maximum number of iterations for AP. Like convits, We are also more concerned about its relations to the running time and the clustering number. Set N = 100, lam = 0.5, convicts = 5, p = median (s). The relations are shown in Fig. 1.

It can be seen from Fig. 1 that the running time and the clustering number was little influenced by the maxits. Therefore, in order to ensure adequate iteration for AP,we recommend the value of maxits may be set higher.

Fig. 1 Maxits's relations to the running time and the clustering number

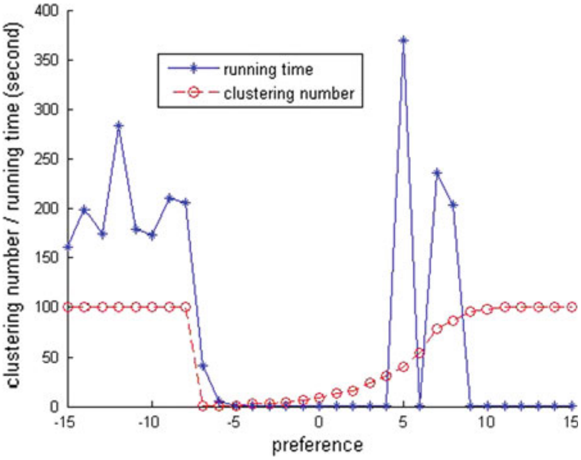


Preference Parameter

As the above points out, preference is a very important parameter for AP. It determines the clustering number, it also exercises an crucial influence over AP convergence rate. So, it is necessary to study its relations to the numerical oscillation, running time, and the clustering number in details. Set N = 100, lam = 0.5, convicts = 50, maxits = 50000. The relations between preference and the running time, and the relations between preference and the clustering number are shown in Fig. 2.

The horizontal axis shows preference varies from median(s)*2^15 to median(s)*2^ − 15. In order to make the conclusions more universal, convicts and maxits are set relatively larger values according to above conclusion. It can be seen

Fig. 2 Preference's relations to the running time and the clustering number



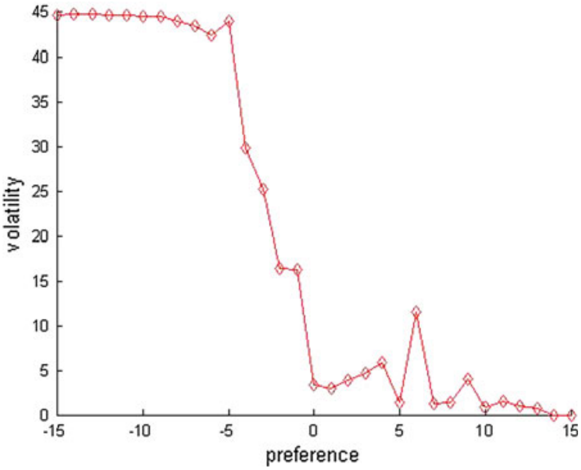


Fig. 3 The relation between preference and volatility

from Fig. 3 that the clustering number varies from 1 to N when the preference is in a certain range; When p = median (s), the clustering number is slightly less than $\sqrt{N}$; It will be difficult to converge and more clustering number, if the preference value too high or too low; Increase the algorithm scale and repeated runs on the certain dataset, it can still come to these conclusions. According to the above conclusions, we can get the idea for Re–K-AP, re-runs AP until the specified clustering number is generated. Of course, there are other ways for K-AP, Zhang proposed an idea for K-AP by introducing a constraint in the process of message passing, was more effective than k-medoids w.r.t. the distortion minimization and higher clustering purity [7]. Our experiments shows that K-AP preserves the clustering quality as Re–K-AP in terms of the distortion, and it has negligible increase of computational cost compared to Re–K-AP.

The relation between preference and volatility is shown in Fig. 3.

The horizontal axis means preference varies from median(s)*2^−15 to median(s)*2^15. It can be seen from Fig. 3 that the volatility declines with increasing the value of preference on the whole. Therefore, the degree of numerical oscillation can be reduced by enlarging the value of preference. Combined with Fig. 3, We find that it does not mean AP must converge when volatility declines. Literature [5] pointed out that numerical oscillation could be escaped by decreasing the preference. According to our conclusions, it may be right under certain circumstance, but it is more like wrong, because numerical oscillation will increase. Through a lot of studies and experiments, it shows that better experiment result can be gained when p is in the range: median(s)* 2^−5 ~ median(s)* 2^5.

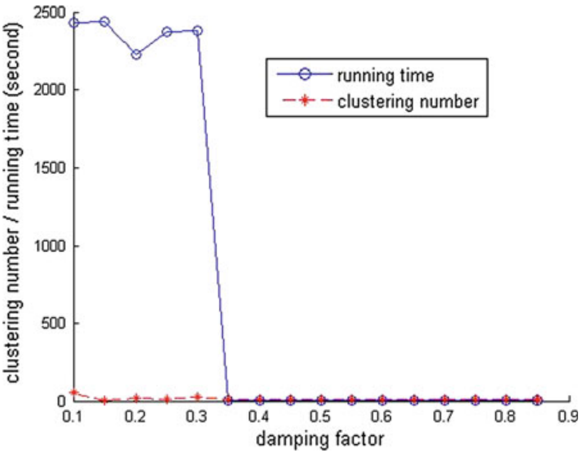


Fig. 4 Damp factor' relations to the running time and the clustering number

Damping Factor Parameter

Like preference, damping factor is also an vital parameter. According to the above conclusions, the damping factor can be used to eliminate the numerical oscillation. We are also more concerned about its relations to numerical oscillation, the running time, and the clustering number. Set N = 100, convicts = 50, maxits = 50,000, p = median(s).The relation between the damping factor and running time,and the relation between damping factor and clustering number are shown in Fig. 4.

It can be seen from Fig. 4 that the clustering number is relatively steady when the damping factor increases from 0.3 to 0.85, and the running time is also steady

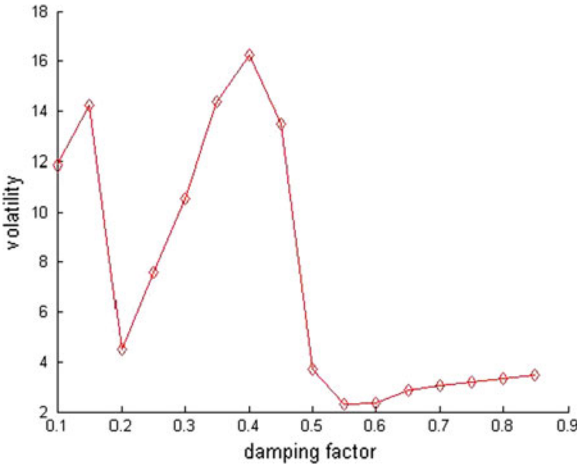


Fig. 5 The relation between damping factor and volatility

when the damping factor is greater then 0.35. Therefore, damping factor has a little effect on the clustering number and running time.

The relation between damping factor and volatility is shown in Fig. 5. It can be seen from Fig. 5 that the volatility declines with enlarging the value of damping factor on the whole. Therefore, the degree of numerical oscillation can be reduced by enlarging the value of damping factor. Through a lot of studies and experiments, the damping factor should be set a large value from the beginning and subject to/$0.5 \leq \lambda < 1$.

Conclusions

The main parameters of AP are discussed in this paper. Because of the complex using circumstance, our conclusions is only for reference. Of course, It may further develop the studies on some parameters, for example, how to generate the specified clustering number quickly according to the relation between preference and clustering number? how to combine and balance the preference and damping factor to get the best result? and so on. All of these will be researched in the future.

Acknowledgments. This research was supported by the grants from the Natural Science Foundation of China (No. 71271209); Huaiyin Normal University Youth Talents Support Project (NO. 11HSQNZ18).

References

1. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315(5814):972–976
2. Frey BJ, Dueck D (2008) Response to comment on “clustering by passing messages between data points”. Science 319(5864):726
3. Leone M, Sumedha S, Weigt M (2007) Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics 23(20):2708–2715
4. Sumedha ML, Weigt M (2008) Unsupervised and semi-supervised clustering by message passing: Soft-constrain affinity propagation. Eur Phys J B 66:125–135
5. Wang K, Zhang J, Li D, Zhang X, Guo T (2007) Adaptive affinity propagation clustering. J Acta Automatica Sinica, 33(12): 1242–1246, (In Chinese)
6. Yu X, Yu J (2008) Semi-supervised clustering based on affinity propagation algorithm. J Software, 19(11):2803–2813, (In Chinese)
7. Zhang X, Wang W, Nørvåg K, Sebag M (2010) K-AP: generating specified K clusters by efficient affinity propagation. ICDM 2010: 1187–1192

Ant Colony Algorithm and its Application in QoS Routing with Multiple Constraints

H. E. Huilin and Y. I. Fazhen

Abstract In the modern communication network, QoS routing optimization problem acts as one of the most important types of discrete optimization problem which could normally be solved by heuristic algorithm. And Ant Colony Algorithm as a new heuristic optimization algorithm shows good performance in solving complex optimization problems. In this paper, Ant Colony Algorithm is presented to solve the QoS unicast routing problem under the constraints of bandwidth and delay, using the mechanism that ants are able to find the optimal path through pheromone and joining the heuristic strategy. The simulation results show that this algorithm can quickly find the routing that meets the constraints of time delay and bandwidth with minimum cost and minimum time delay.

Keywords Ant colony algorithm · QoS routing · Cost · Delay · Dijkstra algorithm

H. E. Huilin (✉) · Y. I. Fazhen
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: 10271037@bjtu.edu.cn

Y. I. Fazhen
e-mail: fzhyi@bjtu.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_73, © Springer Science+Business Media Dordrecht 2014

The Research and Exploration of the Development Trend of Cloud Computing

Shuping Yang and Lijuan Yang

Abstract Cloud computing is praised as the third-time IT revolution following the transformation of the personal computer, and the Internet change, its future trend of development will greatly impact on the existing IT concept that will change the business models and and people's way of life and work. This paper do the summary analysis and research for present global cloud computing, put forward the development measures and the problem to solve about China's cloud computing.

Keywords Cloud computing · Moble internet · SME

Introduction

In the era of cloud computing, companies no longer need to self-built data centers, since the group IT team to maintain and manage the system, enterprises need IT services via the Internet by specialized companies, like running water or electricity.

According to the prediction of international authoritative institutions, in 2014 global cloud computing output will reach 1,700 billion dollars [1]. BRICS cloud computing investment and research in 2012 has been far behind the United States, Canada and other developed countries, at present, the development of cloud computing has entered the stage of the campaign of, in China, If we do not increase investment and research of cloud computing, we are be again behind the developed countries in Europe and the United States after the Internet.

S. Yang (✉) · L. Yang (✉)
School of Computer Engineering, Shenzhen Polytechnic, Shenzhen 518055, China
e-mail: ysping@szpt.edu.cn

L. Yang
e-mail: Ylljj928@szpt.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_74, © Springer Science+Business Media Dordrecht 2014

People have a certain inertia and dependence using technologies or products, everyone has the inertia you want to re-capture the market is not easy.

What is Cloud Computing?

The concept of cloud computing is Google first proposed in 2006, 2007 IBM technical white paper also mentioned cloud computing, since the industry began an extensive discussion and research for cloud computing [2]. A time, on overwhelming the concept of cloud computing, so as SaaS, IaaS, PaaS, and so all of a sudden people into the "cloud", "fog".

In fact, Google and Baidu's search engine is the earliest of cloud computing, search engine to the majority of individuals and enterprises to provide information query function, the sharing of information resources, is cloud computing; Sina's microblogging, Tencent's micro-letters, etc., for the users to provide a platform for communication, sharing of platforms and services, it is cloud computing, Jinshan network fast disk, etc., are cloud computing technology. This is what we see around the touch of cloud computing. In fact, we are already living in the "cloud".

So, what is cloud computing? Although the definition given we are not the same, but the core idea is the same. Simply put, cloud computing is to optimize computing resources on the Internet integration, unified management and scheduling, constitute a powerful computing pool to provide services to the user on demand.

The cloud computing system architecture is shown in Fig. 1.

Type of Cloud Computing

Cloud computing application range is divided into public cloud, private and hybrid clouds [3].

- The public cloud is to provide services for the majority of consumers or businesses via the Internet, it is a new business model using cloud services to provide services to end users.
- Private cloud is the enterprise cloud technology to build enterprise network applications within the enterprise information platform, more emphasis on efficiency and cost advantages of cloud applications within the enterprise, as well as to the internal IT management.
- Hybrid cloud companies face different applications use to hire an external public cloud service or self-built cloud platform.

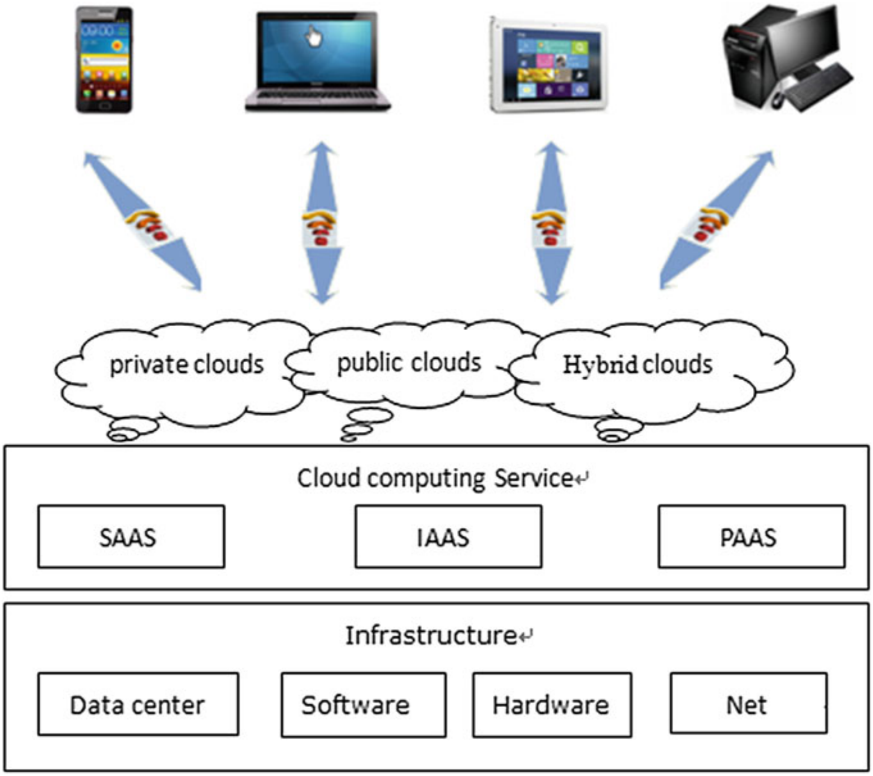


Fig. 1 Cloud computing system architecture

Services and Key Technology of Cloud Computing

Application structure in terms of, in order from the underlying hardware to the upper application, cloud computing offers three levels of service [3].

- IaaS:Infrastructure as a Service, infrastructure cloud.
- Paas:Platform as a Service,platform cloud.
- SaaS:software as a service,application cloud.

Involved in cloud computing technology [3]: Data Storage Technology, Data Management Technology, programming model, cloud security etc.

Features of Cloud Computing

Cloud computing is to continuously improve the processing capability of the "cloud", to reduce the burden of the user terminal handling, and finally to the user terminal simplified to a simple input and output devices, and can enjoy a "cloud" of the powerful computing capabilities demand. Cloud computing has the following characteristics [3]:

- scalability, flexibility and on-demand services and billing. Cloud services with on demand resource allocation and charges.
- extremely cheap. Is the internal private clouds or external public cloud, due to the centralization of the data center scale, economies of scale led to the rapid decline in costs and expenses.
- The reliability and versatility. Cloud computing can provide more reliability than the local computer. At the same time, cloud computing has the versatility with a "cloud" can support different applications run.
- large-scale. Google "cloud" has more than 100 million servers, declared objective is the development of more than 10 million units, Amazon, IBM, Microsoft and other "cloud" also hundreds of thousands of enterprise private cloud scale is generally within a few hundreds to thousands of servers. Larger centralized data center industry trends.
- high efficiency. The industry estimates the cloud computing implementation of the current IT resource utilization can be increased by 5–7 times.

How Cloud Computing from "Clouds" into "Spring Rain"?

Now, cloud computing technology has been steadily through the early "most difficult" stage of development, many industries also fully aware of the benefits of cloud computing, more and more companies have adopted or integrated some cloud computing technology to the business, which means that they benefit a lot from the cloud.

The saying goes, "your spring such as oil", cloud computing from the "cloud" becomes priceless "spring rain", it has the inevitability of technology and market.

Mobile Internet to Promote Cloud Computing Priceless

Cloud Computing Center of the investment of resources and energy consumption and waste is amazing, the initial proposed not to be optimistic. However, the widespread use of smart phones, tablet PCs cloud computing leaps and bounds, it can be said mobile Internet cloud computing priceless. The rapid expansion of the user base so that cloud computing offers free platform to become possible, open platform Baidu cloud, computing, storage, and use of network resources is no longer charges are freely available to the majority of small and medium-sized enterprises and application developers to attract massive subscribers and to generate a great deal of network traffic, then flow returns to achieve profitable platform.

Market research firm IDC survey forecast, the 2013 iphones and ipads market will grow by 20 %. Among them, the screen is smaller than an 8-inch mini ipads accounted for 60 % of the tablet market. Mobile manufacturers will break the present status, or the rise or decline [4].

SaaS Applications to Flourish

At present, SaaS is the most mature, the most famous and most widely used cloud computing model. In the past few years, has paid off handsomely for many companies to bring the rapidly expanding enterprise, such as Salesforce, Amazon and other. Today, cloud computing service objects are small and medium-sized enterprises from around the world gradually extended to the government, research, education, medical and other fields, and large enterprises.

2013 SME SaaS market booming. According to Gartner research, in 2012 the global SaaS applications market has reached $12 billion, in 2015 to more than 20 billion U.S. dollars. Although the main consumers of SaaS applications or Europe and the United States, Japan and South Korea, Australia and New Zealand and other developed countries and regions, but the growth of the emerging markets, including China, should not be underestimated. IDC survey shows that in 2013 global IT spending will be up to $2.1 trillion [5], the golden age of cloud computing upcoming.

Cloud Computing Trends

- mobile cloud will be more brilliant

Mobile Internet to a large customer base, the flow its development offers great potential. Commissioned a slightly more complex due to the restrictions on the capacity of smart phones, cloud computing has some limitations, advantages and disadvantages, and commissioned the task is relatively simple, relatively easy to implement and manage SME service providers and high enthusiasm, to facilitate the formation of a huge market. In addition, the development of smart phone hardware is also important part of the mobile Internet, I believe that in the near future, smart phones will be more capacity.

- small-scale industry cloud colorful

We will provide a small cloud computing platform cloud services in specialized industry applications and services known as the small-scale industry. Small industry cloud computing platforms to specialized areas, professional services, fast and efficient and has broad prospects.

- cloud security issues highlighted

Cloud computing as a system service center under attack is inevitable, the consequences are disastrous. Increased protection measures to improve the efficiency of security is bound to affect the security of cloud even affect the development of cloud computing, governments must attach great importance, sincere cooperation, security technology is important, but the fight against cybercrime must be ruthlessly expected to cloud security technology in there will be a $10 billion market in the next five years [6].

- cloud integration imperative

Cloud computing is still in its infancy, the relevant technical standards, cloud agreements are still inadequate perfect, cloud computing integration, large IT companies will acquire smaller cloud computing companies, systems and norm-setting is imperative. IDC believes that in the next 20 months, the cost of acquisition of SaaS will be more than 25 billion U.S. dollars [4].

- Data rapid market growth

IDC said that the big data market will be an annual growth rate of 40 %. 2012 big data market size of approximately $ 5 billion, will double in 2013, will reach $ 53 billion in 2017. This will cloud computing bring considerable profit margins [4].

China's Cloud Computing How to Achieve the Bend to Overtake?

China's cloud computing research and investment is relatively backward, you want to bend to overtake the need to rely on innovation. As the saying goes "non-profits can not afford early, the development of new technologies ultimately rely on the enterprise. If there is no profit, the firm is not dry.

- cloud computing has the advantage. At present, China's number of Internet users is about more than 300 million mobile phone users reached more than 600 million, the market potential is huge, which means the companies to develop sufficient power.
- American experience detours, open platform for Internet sharing, and energy conservation can be more optimized solution.
- construction of smart city for cloud computing to provide a broad market prospect. In 2012, the mobile Internet, cloud computing, Internet of Things rapid development, further promote economic and social informatization process. Than 320 cities across the country to invest 30 billion yuan in building smart city [7].

- China's three major telecom operators can be greater as. China's huge market potential for the three major telecom operators to provide a broader prospects and kinetic energy, telecommunications operators in the traditional line services can provide cloud computing services platform.
- Baidu, Tencent, Huawei, ZTE, wave and other large domestic IT companies have been established in the field of cloud computing cloud computing service platform, will be to lead and promote the development of China's cloud computing.

Conclusion

Cloud computing has moved from concept to the market and a huge market driven by the interests of governments and companies are competing invested heavily. China should develop countermeasures as soon as possible, otherwise it will again in the wave of cloud computing in a backward position, the domestic Baidu, Tencent, Huawei, Inspur and other large enterprises should seize the opportunity to turn overtaking, the three major telecom operators should also be to actively participate in the competition of cloud computing.

References

1. Gao W, He B, Li J (2012) The development and analysis of the global cloud computing, telecommunications network technology 2(2), 31–34
2. Zhou S, Su C (2012) Cloud computing technology research. Sci Technol Inf 24 (4)
3. Zeng X (2012) The concept of cloud computing applications, wireless internet technology, (2), 32–33
4. Predict 2013 nine technology trend of mobile will break the status quo. http://www.csdn.net/article/2012-12-04/2812492
5. Across the tide SaaS. http://www.programmer.com.cn/15835/
6. Network security overhead climbed to $10 billion after five years. http://net.chinabyte.com/230/11774230.shtml
7. 320 of city of our country investment wisdom city in 2012. http://www.cnscn.com.cn/policy/show-htm-itemid-538.html

The Research of Intelligent Storage System Based on UHF RFID

Yong Lu, Zhao Wu and Zeng-Mo Gao

Abstract Warehousing is an important link of the enterprise operation. Usually, traditional management of warehousing has many defects and low efficiency. We need to change its management mode for improving the efficiency of logistics. Currently, most of warehousing and logistics take barcode as an intelligent way on management. In this way, the intelligence of the system has greatly improved than before, but it still need a lot of manpower and material resources into warehousing and logistics. This thesis designed an automatic identification system which bases on UHF-RFID and embedded control technology. The system can identify goods wirelessly, and realize automated management of warehousing goods.

Keywords RFID · Warehousing system

Background

RFID, as a new generation of automated data collection tool, has been widely used in many fields and gradually began to be used in warehouse management. Compared with the barcode which has been widely used in warehousing, RFID has higher efficiency, longer read distance and large capacity for storage. With economic development and the advancement of technology, the application of RFID in warehousing and logistics will become more and more common.

Y. Lu (✉) · Z. Wu · Z.-M. Gao
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing, China
e-mail: 11120003@bjtu.edu.cn

Y. Lu
e-mail: ylu@bjtu.edu.cn

Z. Wu
e-mail: 10120017@bjtu.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_75, © Springer Science+Business Media Dordrecht 2014

Introduction of RFID

The Structure of Passive RFID

Passive RFID system generally consists of two parts, electronic labels and readers. Usually some specific format information stores in the electronic label. In actual applications, electronic label attached to the surface or inside of the object to be identified. When the object with electronic label falls into the area where the reader can connect, the reader automatically read out the information contained in the electronic label with non-contact manner.

In the field of UHF RFID [1], the mainstream protocol is ISO/IEC 18000-6C and EPC G1C2 [2]. In both protocols, the communication between readers to label is based on the PIE encoding. All communications should be started with previously synchronized code or frame synchronization. The reader takes DSB-ASK, SSB-ASK or PR-ASK modulation scheme for communication. The label should be able to demodulate the above-mentioned three types of modulation. The encoding of data signals are mainly use the FM0 baseband coding and Miller subcarrier modulation.

Algorithm for Anti-Collision in RFID

During the multi-target communication, multiple readers or multiple labels take up a communication channel to send data at the same time. As a result, collision and the interference between devices will occur. In this condition, communication absolutely fails. Taking the above situation into consideration, an algorithm for anti-collision is of great significance for a warehousing system based on UHF RFID.

The algorithm for anti-collision based on ALOHA is suitable for our system because it has been widely used in many fields and easy to implement [3]. ALOHA can be divided into pure ALOHA, slotted ALOHA, frame slotted ALOHA and dynamic frame slotted ALOHA.

Combining the rules of the EPC Gen2 protocol label, this design make a definite improvement on the traditional frame slot ALOHA anti-collision algorithm. The improved algorithm similar to the dynamic frame slotted ALOHA algorithm. But the concept of a frame is replaced by the identification cycle which is the time between two Query commend. In this algorithm, the number of slots in each identification cycle can adaptively change According to the label identification availability. In addition, every identification cycle can stops at any time based on the adaptive and then start another identification cycle. It replaces the mechanism that the number of slots changes after the identification cycle end. This algorithm has good performance on throughput and stability and is applicable to the warehousing system based on RFID.

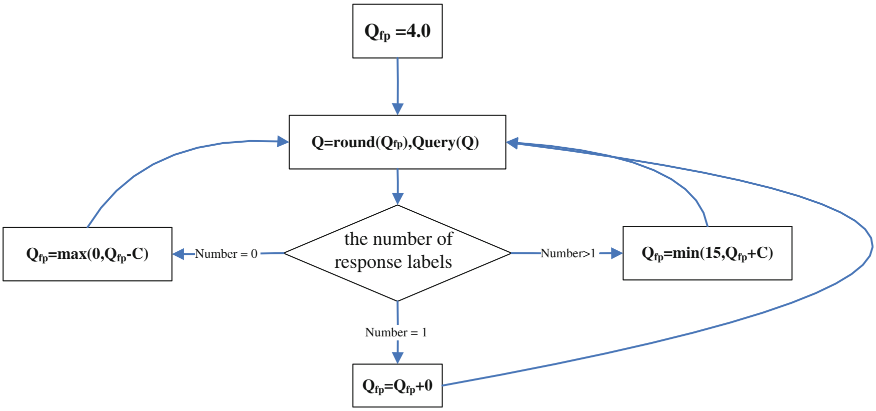

Q_{fp} =4.0
Q=round(Q_{fp}),Query(Q)
the number of response labels
Number = 0
Q_{fp}=max(0,Q_{fp}-C)
Number>1
Q_{fp}=min(15,Q_{fp}+C)
Number = 1
Q_{fp}=Q_{fp}+0

Fig. 1 Anti-collision a algorithm

In this algorithm reader sent EPC protocol command Query [4] to define the number of time slots contained in each piece of identification cycle. Query command sends a parameter Q to the label in RF field. After receiving the Q value, the slot counter in respective label will generate a random number within the range of 0–2Q-1 and then whenever the reader sends QueryRep command to the label, random number of the label is decremented by one automatically. To the random number down to 0, the label will return the RN16 handle and the identification is successful. The collision appears when more than one label's random number reaches to 0. The unidentified labels enter the sleep state and wait for getting a new random number in next identification cycle. Based on the number of unidentified labels, Q value can be made adaptive. This thesis set 4 as the initial value of Q and the algorithm processes shown in the Fig. 1.

After simulated in MATLAB, the algorithm shows a good performance. When the number of labels is under 30, the throughput rate increases from 0.27 to 0.35 with the increase in the number of labels; when the number of labels is more than 30 and less than 500, the throughput rate is floating around 0.35.

Program Design

Hardware Structure

The design goal of intelligent warehousing recognition system RFID reader [5] is to accurately identify the goods affixed with EPC labels through the working table at a certain speed and transmit the label information to the host computer. According to the label content, host computer perform specific statistical functions

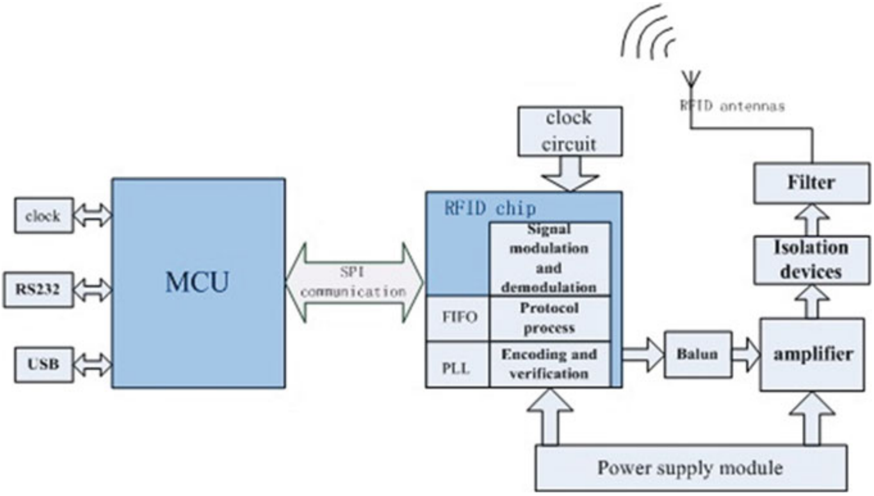


Fig. 2 Hardware system

and update the database. The following figure shows the structure of the hardware of the RFID reader.

The function of the RFID reader:

Launch command to search for labels; Detected electronic labels and generate RF continuous carrier; Provide energy to the labels to communicate, modulation, demodulation; read, write and encrypt the label information.

The hardware part includes a RFID chip, controller, signal isolation circuit, power supply circuit, a voltage -controlled oscillator and clock circuit. Figure 2 shows the structure.

In this design, control part and RF part coexist in one circuit board. FL2440 ARM9 development kit is selected to be control part of the reader. The control part and RF part is connected via SPI communication.

The design choices AS3992 as the RFID chip. AS3992 provide a downlink communication rate up to 640 kbps and add DRM process.

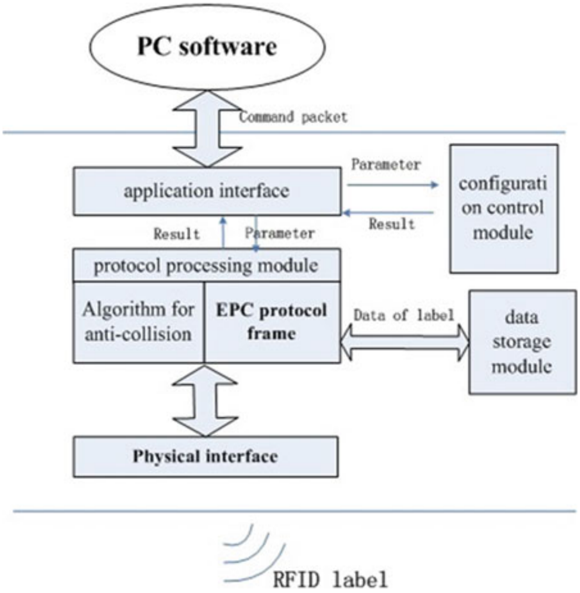
Software Structure

The structure of RFID system software part is shown below. The entire software part is mainly divided into the application interface module, the configuration control module, the protocol processing module and the data storage module (Fig. 3).

Function of each module.

The application interface module: This part provides the connection between protocol layers and the host computer application layer. Operation from PC passed

Fig. 3 Software system



to the protocol layer through the application interface module. After get information from the packet from PC, the protocol processing module completes the corresponding task.

The configuration control module: The entire RFID systems program is burned into the ARM9 processor by CodeWarrior and then the ARM control module will do control and communicate with AS3992. According to the different needs of the user, the reader configured for different environments. Moreover, software design automatic mode for reader. ARM control module get configuration parameters from application interface module. When configuration finish, result will send to PC.

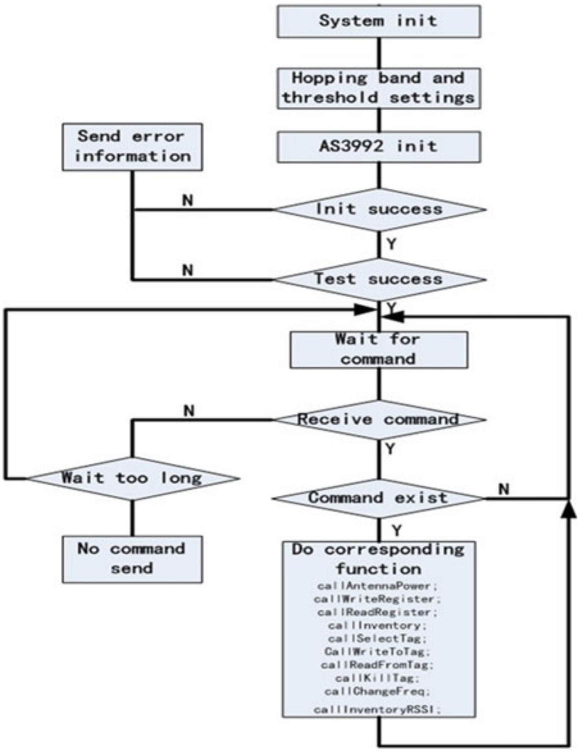
The protocol processing module: This part is the core module of the reader software system and mainly deal with RFID protocol, the related anti-collision algorithm. The reader uses a different protocol processing module can match different RF protocols to operate different protocol version label. This module exchange data with others and the identification work complete in this part.

The data storage module: This module mainly used for the storage of temporary label data. Because the speed of reading label is faster than outputting speed, the reader cannot guarantee outputting data timely when numbers of labels has been read. So system needs to store the data and output at spare time. Storage management obtains each label input and read record. The module cooperates with the corresponding memory interface to form a complete storage.

The RF label identification software implementation process.

RFID software process starts with creating connection between the reader and PC. The reader receives command from PC and completes corresponding work

Fig. 4 Software implementation process



such as read information from label and write information to label. As the core of the reader, ARM coordinates each module and makes the EPC protocol work. The system works as the following Fig. 4.

Test and Conclusion

Author puts 30 RFID labels at positive direction to the antenna and makes the distance between labels and antenna different.

First test is to confirm the limit distance that the reader can indentify whole labels successfully. The conclusion is: when the distance is less than 4.9 m, all of the label can be identified and when the distance is longer than 4.9 m, the identification rate decrease with the increase of the distance.

Second test is to confirm the limit distance that the reader can write information to labels successfully. The conclusion is: when the distance is less than 2.0 m, all of the label can receive the new data from the reader and when the distance is longer than 2.0 m, the information in some labels cannot be changed.

Third test is to test the identification speed. The conclusion is: when the label in the normal distance, the time of recognize all 30 label is less than 1 s.

The result shows that the RFID identification system can basically meet the requirements of the automated identification of the goods in real life.

Prospect

With the in-depth development of the Internet of Things, intelligent storage will no longer be confined to the closed environment of automated identification. In the future, the information of the goods will form a database. The reader can connect to the database and download or update the information. That is more convenient and makes warehousing intelligent.

References

1. Lee J (2007) A UHF mobile RFID reader IC with self-leakage canceller. Radio Freq Integ Cir (RFIC) Symposium. IEEE, 273–276
2. International Standard Organization radio frequency identification for item management-part6: parameters for air interface communications at 860–960 MHz 2004
3. Li M, Qian ZH, Zhang X, Wang Y-J (2011) Slot-predicting based ALOHA algorithm for RFID anti-collision. J Commun 32 (12):43–50
4. International Standard (2005) ISO-IEC_CD 18000-6C. 1
5. Zhang X-P, Zhu Y-l, Luo H (2005) Design of UHF RFID interrogator. Chinese J Elec Dev 28(3):542–545

An Efficient Detecting Mechanism for Cross-Site Script Attacks in the Cloud

Wei Kan, Tsu-Yang Wu, Tao Han, Chun-Wei Lin, Chien-Ming Chen and Jeng-Shyang Pan

Abstract Cloud computing is one of the most prospect technologies due to its flexibility and low-cost usage. Several security issues in the cloud are raised by researchers. Cross-site script (XSS) attack is one of the most threats in the Internet. In the past, there are many literatures for detecting XSS attacks were proposed. Unfortunately, fewer studies focus on the detection of XSS attacks in the cloud. In this paper, we propose a mechanism to detect XSS attacks in cloud environments. The framework is also presented. In particular, our mechanism is not need to modify browsers and applications. We demonstrate our mechanism has higher accuracy rate and lower impact on performance of applications in the experiment. It sufficiently shows our mechanism is suitable for real-time detection in XSS attacks for cloud environments.

Keywords XSS attack · Cloud computing · Detection · Real-time

W. Kan · T.-Y. Wu (✉) · T. Han · C.-W. Lin · C.-M. Chen · J.-S. Pan
Innovative Information Industry Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China
e-mail: wutsuyang@gmail.com

W. Kan
e-mail: hbrkanwei@163.com

T. Han
e-mail: 284763655@qq.com

C.-W. Lin
e-mail: jerrylin@ieee.org

C.-M. Chen
e-mail: chienming.taiwan@gmail.com

J.-S. Pan
e-mail: jengshyangpan@gmail.com

C.-W. Lin · C.-M. Chen · J.-S. Pan
Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen 518055, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_76, © Springer Science+Business Media Dordrecht 2014

Introduction

Nowadays, cloud computing is obviously one of the most prospect technologies due to its flexibility and low-cost usage. Different from the traditional Internet environments, enterprises can exempt from building expensive infrastructure by the cloud techniques. Hence, many IT companies deploy their web applications into the cloud environments. As we all know, the cloud environments are based on the Internet. Hence, the threats in the traditional Internet environments still exist in the cloud environments.

Cross-site script (XSS) attack is one of the most threats in the Internet. It can grab the user's privacy information and leads other attacks such as fishing, SQL injection, and DDoS. This attack is caused by the script language embed into web applications. In general, the web applications are adopted HTML language, script language, hyperlinks, and other languages to provide resources operating and interaction between the client and the server. However, these languages and methods lead web applications are vulnerable to XSS attacks.

Typically, the XSS attack contains the three attacks models: (1) Reflected XSS, (2) Stored XSS, and (3) Dom-based XSS. Up to now, several approaches were proposed to prevent XSS attack such as static analysis [1–3], black-white list [4], taint and flow analysis [5–8], string injection [9–12], machine learning [13–16], rewriting [17, 18]. For the static analysis approach, it addresses XSS attacks by means of static source analyses such as Jovanovic et al. [1] presented a method called alias analysis for PHP language in 2006 and Wassermann et al. [2] proposed the method which combines tainted information flow with string analysis in 2008. For the back-white list issue, it uses back and while list between the client and server to prevent XSS attacks. In 2007, Jim et al. [4] present a method called BEEP. In BEEP, scripts written by developer are stored in whitelist and malicious scripts are stored in blacklist. Before browser executes JavaScript codes, it will check the black and white lists. For the taint and flow analysis approach, it addresses XSS attacks by tracking the sensitive data such as Vogt et al.'s method [5]. For the string injection issue, Shahriar and Zulkernine [11] proposed a method by inserting comment tags containing rules into HTML page to restrict the content in 2011. For the rewriting approach, it addresses XSS attacks by inserting some tokens into the request and the response messages as Cookie and Headers. In 2011, Putthacharoen and Bunyatnoparat [18] put forward a dynamic cookies rewriting method. Recently, Shar and Tan [19] proposed a detecting method which can automatically audit and remove the XSS attacks.

With the fist growing of cloud computing, the research on the protection of application system in cloud platform becomes an important issue. XSS attack is one of the most threats in traditional application system. Hence, the protection of XSS attack should be considered in the cloud. However, the traditional detection methods for XSS attacks are not suitable in the cloud environments because most methods need to modify applications source codes or browsers. Meanwhile, most methods also cannot provide real-time protecting. In this paper, we first define the framework and related notions of XSS attacks in cloud environments. A concrete detecting mechanism is then proposed. Our mechanism can solve the mentioned disadvantages of traditional detecting methods of XSS attacks. In the experiments, our mechanism has a higher precision rate and a low impact on the performance of the application systems in the cloud.

The remainder of this paper is organized as follows. In section Framework, we propose the framework of detecting mechanism. A concrete detecting mechanism is presented in section Concrete Mechanism. In section Experiments and Comparisons, we describe the experiments and the conclusions are given in section Conclusions.

Framework

In this section, we present a sketch framework which is depicted in Fig. 1. Our framework consists of three entities: User, App Server, and XSS Detection Server as follows.

(1) **User**: The user can play the following three roles: app user, app administrator, and platform administrator. We assume the user visits App Server in first time which is not malicious. This assumption is called *F-secure*.

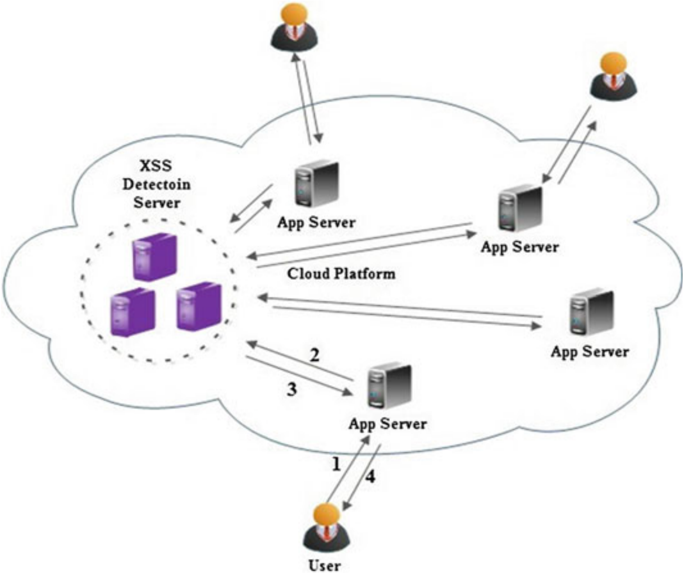


Fig. 1 Sketch framework

(2) **XSS Detection Server** (**XDS**): XDS is responsible to dispose the request messages sent by user and checks the messages whether they are malicious script languages.
(3) **App Server**: In App Server, there is a filter (named XFD) which is responsible to gather the request messages sent by user and then sends the gathered messages to XDS. Meanwhile, we define the following two rules for app in App Server.

Rule 1: *HTML pages should not include comment tags.*
Rule 2: *Script tags should not be dynamically generated in HTML pages.*
The flowchart of our framework contains the following four steps:

Step 1: User sends the request messages to App Server.
Step 2: XFD collects the request messages sent by user and sends them to XDS.
Step 3: After receiving the messages from XFD, XDS verifies them whether they are malicious script languages. Then, XDS outputs the result and sends to XFD. We use the symbol NO_XSS is denoted by the request messages are secure and the symbol HAVE_XSS is denoted by the request messages are malicious script languages.
Step 4: Upon receiving the result form XDS, App Server provides the service for the user if the result is *NO_XSS*. Otherwise, App Server terminates the connection and responds a warning to the user.

Concrete Mechanism

Parsing Phase

Here, we first define a new data structure called *levelgroup* which is used to represent the structure of HTML pages. Then, we propose an algorithm named *FP* algorithm which is used to transform HTML pages into *levelgroup*. By the *F-secure* assumption, we can obtain an initial *levelgroup* in which the elements are not attacked.

Definition 1 Let *levelgroup* be a set which contains $n + 1$ subsets G_i for $i = 0, 1, \ldots, n$, where each G_i is an ordered set. Here, each element of G_i is consists of tag name and attribute.

For convenience to describe *FP* algorithm, we define *Page* by HTML page visited by user and pseudo code of *FP* algorithm is proposed as follows:

Algorithm 1: FP algorithm

```
Input: Page
Output: levelgroup
(1) PStructure ← obtaining the structure of Page;
(2) levelgroup ← defining a map<key, G>, where key denotes the index of G;
(3) For node n in PStructure using depth-first method do
      i ← calculating the corresponding level of n;
      If levelgroup contains key i then
        $G_i$ ← getting value from levelgroup by key i;
        Storing n into $G_i$;
      Else
        $G_i$ ← defining a subset of levelgroup;
        Storing n into $G_i$;
        Putting entity(i, $G_i$) into levelgroup;
(4) Return levelgroup;
```

Identifying Phase

Here, we first define some functions as follows:

Definition 2 (*H function*) *H* function is defined by $H\ (G^{old}, G^{new})$. It responsible to calculate added and deleted elements between the two subsets of *Levelgroup*. Then, it outputs two lists $List_{add}$ and $List_{delete}$. Formally,

$$\text{add } E^{new} \text{ into } List_{add} \text{ and add } E^{old} \text{ into } List_{delete}$$

Definition 3 (*C function*) *C* function is defined as $C\ (G^{old}, G^{new})$. It is responsible to check the name and attributes of elements in the same position of two subsets G^{old} and G^{new}. Then, it outputs the corresponding results *HAVE_XSS*, *NO_XSS*, or *UNCERTAIN*. The formal definition is described below. Note that E_j is a *j*th element in subgroup *G* and is denoted by the number of attributes of element E_j. Formally,

If ∃ outputting *HAVE_XSS*
If outputting *NO_XSS* for $\in G^{new}$ and $\in G^{old}$
Else adding into $List_{UNCERTAIN}$ Outputting *UNCERTAIN*

Definition 4 (*F Function*) *F* function is defined by $F\ (List_{add}, List_{delete}, G^{old}, G^{new})$. It is responsible to verify the elements of G_{new} whether they contain malicious scripts. Then, it outputs the corresponding results *HAVE_XSS*, *NO_XSS*, or *UNCERTAIN*. Formally,

If $\exists\ E \in List_{add}$ contains malicious scripts (by **Rule1** and **Rule2**) outputting
HAVA_XSS

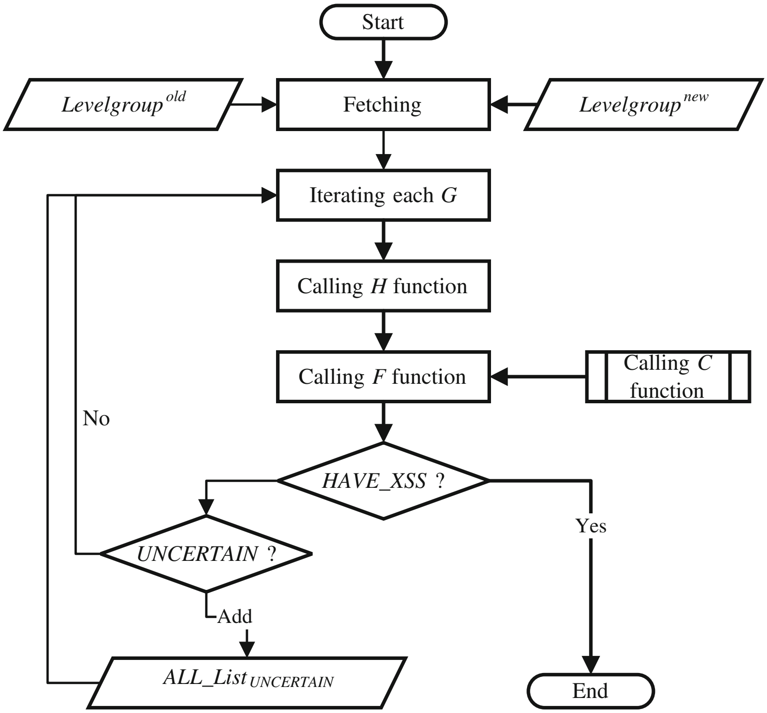


Fig. 2 The flowchart of LC algorithm

Else if $List_{delete} \neq \emptyset$ outputting *NO_XSS*

Else if $List_{add} \neq \emptyset$, adding $List_{add}$ into $List_{UNCERTAIN}$ and outputting *UNCERTAIN*

Else call the *C* function with (G^{old}, G^{new})

Here, we propose an algorithm named *LC* algorithm which is used to identify *Page* whether it suffered from XSS attacks. The pseudo code of *LC* Algorithm and the flowchart in Fig. 2 are presented as follows:

Algorithm 2: LC algorithm

```
Input: Page
Output: HAVE_XSS or NO_XSS
(1)  Calculating $levelgroup^{new}$ of Page;
(2)  Fetching $levelgroup^{old}$ of Page;
(3)  Defining a list, $ALL\_List_{UNCERTAIN}$;
(4)  For each $\in levelgroup^{new}$ and $\in levelgroup^{old}$ do
       Calling H function with (, );
       Calling F function with ($List_{add}$, $List_{delete}$, , );
         If F function returns HAVE_XSS, then breaking ;
       Else if F function return NO_XSS, then continue;
         Else if F function return UNCERTAIN, adding $List_{UNCERTAIN}$ into $ALL\_List_{UNCERTAIN}$, then continue;
(5)  If $ALL\_List_{UNCERTAIN}$ is not empty, calling the CSE algorithm;
      Else return NO_XSS;
```

Verifying Phase

Here, we propose an algorithm named *CSE* algorithm which is used to verify the list *ALL_List*$_{UNCERTAIN}$ by rules.

```
Algorithm 3: CSE algorithm
INPUT: $ALL\_List_{UNCERTAIN}$, application ID
OUTPUT: $HAVE\_XSS$ or $NO\_XSS$
(1)  Fetching the rules by ID;
(2)  Transferring the rules to automata;
(3)  Iterating all the elements in $ALL\_List_{UNCERTAIN}$;
     Running the automata with the element;
     If the result of automata is $HAVE\_XSS$ then breaking and returns $HAVE\_XSS$;
        Else continue;
(4)  Returning $NO\_XSS$;
```

Experiments and Comparisons

The experiment environment is simulated by a standard PC, where the processor Intel(R) Core(TM) i5-2400 3.10 GHz, the RAM is 8 GB, and the operating system is Windows 7. Meanwhile, we use VMware Workstation 7 to customize three virtual PCs to be the XSS Detection Server, one virtual PC to be the App Server. Here, we design an application (app_xss) which suffered from XSS attack.

In Table 1, we demonstrate the accuracy rate of our proposed mechanism with 32,560 malicious URLs [20]. The pre-treatment of URLs are consisted of following steps:

- Decoding and deleting invalid parameters in URLs.
- Changing parameter names and encoding URLs.

The details of experiment are described in the following steps:

1. We honestly visit app_xss in App Server.
2. We fetch disposed URLs to request app_xss in App Server.
3. XDS verifies them and outputs the accuracy rate as shown in Table 1.

In Fig. 3, we demonstrate the accuracy rate of our proposed mechanism with 26 websites such as Sina.com, QQ.com, ifeng.com, etc. It is easy to see that our mechanism provides an insignificant performance overhead of applications. Thus, our mechanism is suitable for real-time detection in XSS attacks.

Table 1 The accuracy rate of our mechanism

Total URLs	Malicious URLs	None malicious URLs	Error URLs
32,560	32,195 (99.18 %)	216 (0.67 %)	49 (0.15 %)

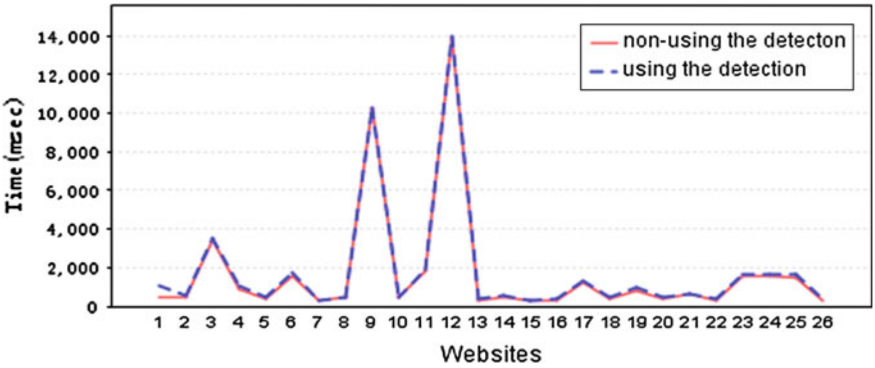


Fig. 3 The efficiency of our mechanism

Table 2 Comparing with other detection methods

Method	BSC	MSC	MB
Static analysis [1–3]	Yes	No	No
Black-white list [4]	Yes	Yes	Yes
Taint and flow analysis [5–8]	Yes	No	No
String injection [9–12]	Yes	Yes	Yes
Machine learning [13–16]	No	No	No
Rewriting [17, 18]	No	No	Yes
Our method	No	No	No

In Table 2, we list the comparison between some of previous XSS detecting methods in terms of whether modifying browsers (MB), browsing source codes (BSC), and modifying source codes (MSC). For instance, Black-white list method needs MB to insert black-white list parsing codes into the JavaScript engine of browsers. Meanwhile, this method needs BSC and MSC to insert black and white list into HITML pages. Obviously, our method is based on the structure of HTML pages such that we need not to do extra operation for the languages. Thus, it sufficient demonstrates the advantage of our mechanism.

Conclusions

In this paper, we have proposed a detecting mechanism of XSS attacks and defined the corresponding framework in the cloud. Our mechanism is not need to modify browsers and applications. The experiment results are demonstrated our method has higher accuracy rate and lower impact on performance of applications.

Acknowledgments The authors thank the referees for their valuable comments and constructive suggestions. This research was partially supported by Shenzhen peacock project, China under contract No. KQC201109020055A and Shenzhen Strategic Emerging Industries Program, China under Grants No. ZDSY20120613125016389 China.

References

1. Jovanovic N, Kruegel K, Kirda E (2006) Precise alias analysis for static detection of web application vulnerabilities. In: 2006 workshop on programming languages and analysis for security. ACM press, New York, pp 27–36
2. Wassermann G, Su z (2008) Static detection of cross-site scripting vulnerabilities. In: 30th international conference on software engineering. IEEE press, New York, pp 171–180
3. Zhang XH, Wang ZJ (2010) A static analysis tool for detecting web application injection vulnerabilities for asp program. In: 2nd international conference on e-business and information system security. IEEE press, New York, pp 1–5
4. Jim T, Swamy N, Hicks M (2007) Defeating script injection attacks with browser-enforced embedded policies. In: 16th international conference on World Wide Web. ACM press, New York, pp 601–610
5. Vogt P, Nentwich F, Jovanovic N, Kirda E, Christopher K, Vigna G (2007) Cross-site scripting prevention with dynamic data tainting and static analysis. In: international symposium on network and distributed system security. IEEE press, New York, pp 201–210
6. Lam MS, Martin M, Whaley J (2008) Securing web applications with static and dynamic information flow tracking. In: 2008 ACM SIGPLAN symposium on evaluation and semantics-based program manipulation. ACM press, New York, pp 3–12
7. Zhang Q, Chen H, Sun J (2010) An execution-flow based method for detecting cross-site scripting attacks. In: 2nd international conference on software engineering and data mining. IEEE press, New York, pp 160–165
8. Guarnieri S, Pistoia M, Tripp O, Dolby J, Teihet S, Berg R (2011) Saving the World Wide Web from vulnerable JavaScript. In: 11th international symposium on software testing and analysis. ACM press, New York, pp 177–187
9. Gundy M, Chen H (2009) Noncespaces: using randomization to enforce information flow tracking and thwart cross-site scripting attacks. In: International symposium on network and distributed system security. IEEE press, New York, pp 123–130
10. Johns M, Engelmann B, Posegga J (2011) S2XS2: a server side approach to automatically detect XSS attacks. In: International conference on computer security applications. IEEE press, New York, pp 335–344
11. Shahriar H, Zulkernine M (2009) Injecting comments to detect JavaScript code injection attacks. In: 35th international conference on computer software and applications. IEEE press, New York, pp 104–109
12. Wurzinger P, Platzer C, Ludl C, Kirda E, Kruegel C (2009) SWAP: mitigating XSS attacks using a reverse proxy. In: 2009 ICSE workshop on software engineering for secure systems. IEEE press, New York, pp 33–39
13. Komiya R, Paik I, Hisada M (2011) Classification of malicious web code by machine learning. In: 3rd international conference on awareness science and technology. IEEE press, New York, pp 406–411
14. Choi J, Kim H, Choi C, Kim Pk (2011) Efficient malicious code detection using n-gram analysis and SVM. In: 14th international conference on network-based information systems. IEEE press, New York, pp 618–621
15. Nunan AE, Souto E, Santos EMD, Feitosa E (2012) Automatic classification of cross-site scripting in web pages using document-based and URL-based features. In: 2012 IEEE symposium on computers and communications. IEEE press, New York, pp 702–707

16. Shar LK, Tan HBK (2012) Mining input sanitization patterns for predicting SQL injection and cross site scripting vulnerabilities. In: 2012 ICSE international conference on software engineering. IEEE press, New York, pp 1293–1296
17. Iha G, Doi H (2009) An implementation of the binding mechanism in the web browser for preventing XSS attacks. In: international conference on availability, reliability and security. IEEE press, New York, pp 996–971
18. Putthacharoen R, Bunyatnoparat P (2011) Protecting cookies from cross site script attacks using dynamic cookies rewriting technique. In: international conference on advanced communication technology. IEEE press, New York, pp 1090–1094
19. Shar LK, Tan HBK (2012) Auditing the XSS defence features implemented in web application programs. IET Software 6(4):377–390
20. XSS Attacks Information. http://www.xssed.com

A Novel Clustering Based Collaborative Filtering Recommendation System Algorithm

Qi Wang, Wei Cao and Yun Liu

Abstract Traditional collaborative filtering algorithms compute the similarity of items or users according to a user-item rating matrix. However, traditional collaborative filtering algorithms face very severe data sparsity, which causes a discount of the performance of recommendation. In this paper, we proposed an improved clustering based collaborative filtering algorithm for dealing with data sparsity. We first clustered the users set into k clusters using K-means algorithm. Then we presented a formula to estimate those absent ratings in the user-item rating matrix and acquired a high density matrix. After that, we use the new rating matrix to calculate the similarity of items and predict the ratings of a target user on items which have not been rated and recommend Top-N items to the target user. We also implemented experiments and demonstrated that our proposed algorithm has better accuracy than traditional collaborative filtering algorithms.

Keywords Collaborative filtering · Recommendation systems · K-means algorithm · Data sparsity

Q. Wang · Y. Liu (✉)
School of Communication and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: liuyun@bjtu.edu.cn

Q. Wang
e-mail: 12125042@bjtu.edu.cn

Q. Wang · Y. Liu
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China

W. Cao
China Information Technology Security Evaluation Center, Beijing, China
e-mail: caow@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_77, © Springer Science+Business Media Dordrecht 2014

Introduction

Recommendation systems can automatically recommend to users what they might be interested in. Usually we divide recommendation system algorithms into content-based algorithms [1, 2] and collaborative filtering algorithms [3–6]. Content-based algorithms recommend to users items which are similar to what users have already bought or rated by analyzing the features of users or items. These algorithms can solve the problem called "cold start" and also won't face the challenge of data sparsity because they don't depend on the rating matrix. But they have a serious drawback that they can't deal with pictures, video, music and other products difficult to be analyzed and extracted features from. On the contrary, collaborative filtering algorithms utilize a user-item rating matrix to calculate the similarity between users or items and then predict those items which have not been rated or bought depending on the ratings of neighbors which have high similarity with the target users. However, the number of items which each user has bought is usually less than 1 % of the total number of items in a site, which causes severe data sparsity and a decrease of the performance.

In this paper, we proposed an improved clustering based collaborative filtering algorithm for dealing with data sparsity. We combined K-means algorithms and a formula dealing with data sparsity of the user-item matrix. After that we implemented some experiments and it was shown that our proposed algorithms have a better performance than the traditional algorithms.

The Proposed Algorithm

Traditional collaborative filtering algorithms [4] create a user-item rating matrix and calculate similarity between users or items, but very often this matrix is sparsely populated leading to poor coverage of the recommendation space and ultimately limiting recommendation effectiveness. Our proposed method combines K-means algorithm [7] and a formula which can assign an estimation rating to an unrated item, so we can get a high-density matrix and resolve the data sparsity.

User Clustering

The first step is user clustering, and clustering is a preliminary step for the subsequent step to gather those similar users. In this paper, we use K-means algorithm to cluster our user set in the user-item rating matrix. Furthermore, we use Euclidean distance to represent the distance between users. Let a centroid is denoted by $U_{mid} = (r_{01}, r_{02}, \ldots, r_{0n})$ and user i can be denoted by $U_i = (r_{i1}, r_{i2}, \ldots, r_{in})$, where r_{mn} states the rating of user m on item n, then the distance between the two users is given by,

Fig. 1 K-means clustering algorithm

```
Input: k original user centroids
for each user vector U
    for the k^th centroid C
        distance[k]=Euclidean(U,C);
    end
    find the shortest distance—distance[i]
    assign user U to cluster i
end
```

$$D = \sqrt{\sum_{j=1}^{n} (r_{ij} - r_{oj})^2} \tag{1}$$

Figure 1 shows the process of the K-means algorithm in our recommendation system and Fig. 2 shows the result of the K-means algorithm.

Constructing the Rating Matrix

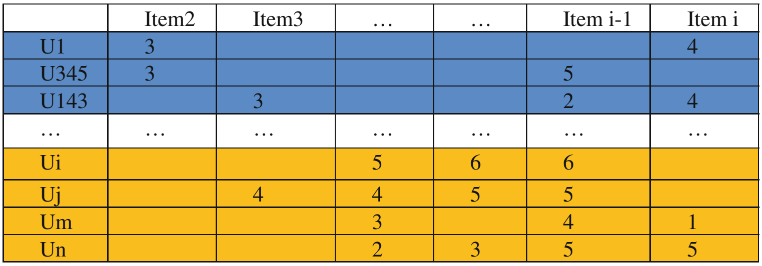
Assuming there are M users and N items in the rating matrix, we define the sparsity level of the matrix as 1 – the number of ratings/M*N. Usually the sparsity level is very high, so in order to resolve this problem, we proposed a formula to calculate an estimation rating for an absent rating. The estimation rating of user *c* on item *s*, $R_{c,s}$ is computed as,

$$R_{c,s} = \bar{R}_c + \frac{1}{|U|} \sum_{\hat{c} \in U} (R_{\hat{c},s} - \bar{R}_{\hat{c}}) \tag{2}$$

where $\bar{R}_c$ is the average rating of the user *c*, *U* is the set of users that belong to the same cluster which is formed through the K-means algorithm and moreover have rated item *s*. Because $R_{c,s}$ is computed by the users in one same cluster and users belonging to one same cluster have higher similarity with each other, the estimation rating is more accurate. In addition, the rating scales of different users are varied, so in order to eliminate this inaccuracy, we let a rating subtract the average rating of a user and utilize the difference to calculate the predicting rating. Apparently, this method can eliminate data sparsity.

Similarity Computation

Similarity computation is the most important step for collaborative filtering algorithms. There are three different ways to compute the similarity between items. They are cosine-based similarity, Person correlation-based similarity and adjusted cosine similarity respectively as shown in Eqs. (3)–(5). We define the set



	Item2	Item3	…	…	Item i-1	Item i
U1	3					4
U345	3				5	
U143		3			2	4
…	…	…	…	…	…	…
Ui			5	6	6	
Uj		4	4	5	5	
Um			3		4	1
Un			2	3	5	5

Fig. 2 The result of clustering

of users that have rated both item i and item j as U, and the average rating of user u is denoted by $\bar{R}_u$. $\bar{R}_i$ and $\bar{R}_j$ denote the average rating of item i and item j respectively.

- Cosine-based Similarity

$$sim\,(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2} \tag{3}$$

- Person Correlation-based Similarity

$$sim\,(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_i)(R_{u,j}-\bar{R}_j)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_i)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_j)^2}} \tag{4}$$

- Adjusted Cosine Similarity

$$sim\,(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_u)(R_{u,j}-\bar{R}_u)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_u)^2}} \tag{5}$$

In the paper [3], Sarwar et al. have demonstrated that adjusted cosine similarity performs best among them in the recommendation system through their experiments. So, in this paper, we will use adjusted cosine similarity as the similarity computation method.

Prediction and Recommendation

After the former similarity computation, we will get a $N * N$ similarity matrix, where N represents the total number of items. The last step is predicting and recommendation. Firstly, we predict the items which have not been bought or rated by the target user, after that we recommend the top-N items to the target user basing the predicting. We use weighted sum to calculate the predicting ratings. Let the target user is u and the unrated item is i. $N = \{$all the items that have high similarity with i and have been rated by user u in the mean time$\}$.The predicting rating of user u on the item i is denoted by the Eq. (6).

$$R'_{u,i} = \frac{\sum_N (sim_{i,N} * R_{u,N})}{\sum_N (|sim_{i,N}|)} \tag{6}$$

Experimental Evaluation

Our data set is from the Movielens which is a web-based research recommender system. The data set includes 100,000 ratings of 943 users on 1,682 items and each user has rated 20 or more movies. The data set is divided into a training set and a test set. The 80 % of the data is used as the training set and the rest 20 % is used as the test set. All our experiments were implemented at Matlab7.0 and run in a PC with Intel Core processor having a speed of 2.2 GHz and 2 GB of RAM.

Metric

We use Mean Absolute Error (MAE) as our evaluation metric, and the MAE is defined as,

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |P_i - R_i| \tag{7}$$

where N states the total number of predicting ratings in the test set, P_i is the ith predicting rating, and R_i is the ith actual rating in the test set.

Experimental Results

First, we implemented an experiment to see the performance of recommendation with the increasing of the density of rating matrix. We let the density of rating

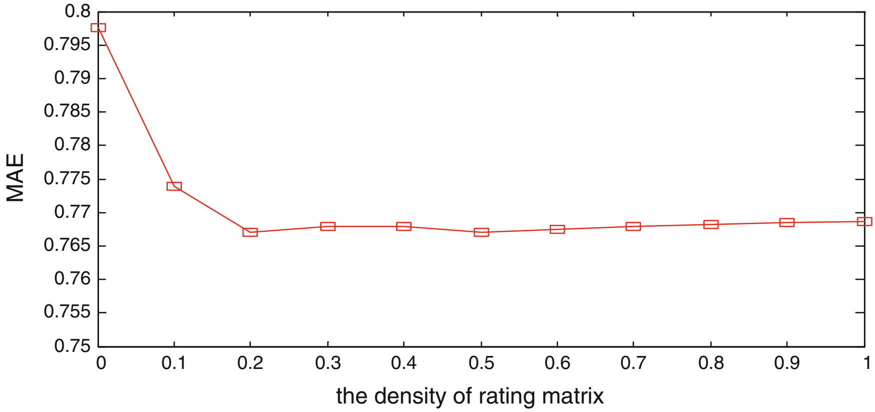


Fig. 3 Performance of the algorithm under different data sparsity

matrix increase at the speed of 10 % and then computed their MAE of predicting ratings. In addition, in this experiment, we set the parameter k = 30 in K-means algorithm and we didn't design an extra experiment to determine the optimal k. The results are shown in Fig. 3. From the Fig. 3 we can observe that with the increasing of the density of rating matrix the MAE of predicting decreases, which means the performance of recommendation getting better. Furthermore, the performance of recommendation improves as we increase the density of rating matrix from 0 to 20 %, after that the curve tends to be flat. With the increasing of the density of rating matrix, the computation also will increase rapidly, so we select the 20 % as the optimal choice of the density. On the other hand, the optimal value of sparsity level is 80 %.

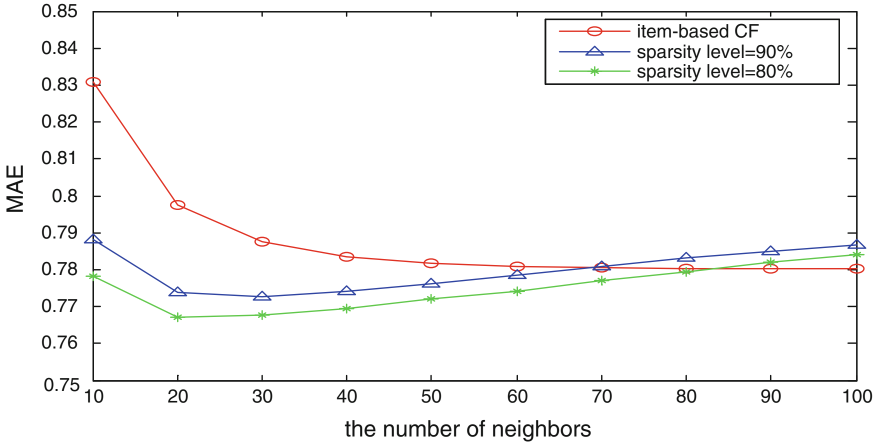


Fig. 4 Performance of the item-based algorithm and our proposed algorithm under the different number of neighbors

Then, we performed an experiment where we varied the number of neighbors and compared the results of the proposed algorithm with the traditional item-based collaborative filtering algorithm. The results are shown in Fig. 4. There are three curves in the figure which represent the normal item-based algorithm, the proposed algorithm where sparsity level equals 90 %, and the proposed algorithm where sparsity level equals 80 %. From the figure, we can observe that our proposed algorithm performs better than the item-based collaborative filtering when the number of neighbors is from 10 to 70, and our proposed algorithm can acquire the minimum MAE when the sparsity level = 80 % and the neighbors = 20.

Conclusion

In this paper, we proposed a new CF recommendation algorithm which introduces the K-means algorithm and estimation ratings to resolve data sparsity. By employing the experiments we observe the performance of our proposed algorithm is much better than the traditional CF algorithm and we also find when de sparsity level = 80 % and the number of neighbors = 20 the MAE of predicting ratings is minimum and the performance is the best. On the other hand, our proposed algorithm resolves data sparsity and acquires a more accurate recommendation. In addition, we can implement more experiments to determine the optimal number k of clusters in the K-means algorithm.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Balabnovic M, Shoham Y (1997) Fab: content-based, collaborative recommendation. Comm ACM 40(3):66–72
2. Mooney RJ, Bennett PN, Roy L (1998) Book recommending using text categorization with extracted information. In: Procedings of recommender systems papers from 1998 workshop, technical report WS-98-08
3. Linden G, Smith B, York J (2003) Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet Comput 7(1):76–80
4. Sarwar B, Karypis G, Konstan J et al (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of 10th international WWW conference, Hong Kong, pp 1–5
5. Deshpande M, Karypis G (2004) Item-based top-N recommendation algorithms. ACM Trans Inf Syst 22(1):143–177

6. Koren Y (2010) Factor in the neighbors: scalable and accurate collaborative filtering. ACM Trans Knowl Disc Data 4:1–24
7. Kim KJ, Ahn HC (2008) A recommender system using GA K-means clustering in an online shopping market. Expert Syst Appl 34:1200–1209

A Secure and Flexible Data Aggregation Framework for Smart Grid

Lun-Pin Yuan, Bing-Zhe He, Chang-Shiun Liu and Hung-Min Sun

Abstract Smart grids are electrical grids that take advantage of information and communication technologies to achieve energy-efficiency, automation and reliability. Smart grids include renewable energy, electrical vehicles, phasor measurement unit (PMU) and advanced metering infrastructure system (AMI) etc. The system's availability can be achieved via data aggregation technique by reducing the overhead of networks. However, since smart grids have become more popular in recent years, many researches have been done on the security issue of the smart grid such as confidentiality, integrity and availability. For these security issues, many researchers adopt secure data aggregation algorithms to protect the data transmission and to reduce the overhead of networks. In this paper, we propose a secure data aggregation framework, which provides multi-level security for different kinds of applications.

Keywords Smart grid · Data aggregation

L.-P. Yuan · B.-Z. He · C.-S. Liu · H.-M. Sun (✉)
Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan,
Republic of China
e-mail: hmsun@cs.nthu.edu.tw

L.-P. Yuan
e-mail: lunpin@is.cs.nthu.edu.tw

B.-Z. He
e-mail: ckshejrho@is.cs.nthu.edu.tw

C.-S. Liu
e-mail: monkey10020@is.cs.nthu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_78, © Springer Science+Business Media Dordrecht 2014

Introduction

Smart grid is the next generation of electricity gird which can help us to monitor and to manage energy usage so that we can accomplish energy conservation and carbon reduction. In a smart grid, meters will report the usage of reading periodically to the energy producer via wireless or power line communication (PLC). In addition, smart grids are required to be self-healing and to protect customers' privacy. Nowadays, there are many research issues [3, 8, 9] need to be discussed in smart grids such as security and network issues.

The network overhead in the smart grid is a serious problem, which needs to be addressed carefully. The metering messages are able to waste the bandwidth of the network due to the increasing requests of sending similar messages. To resolve above problem, many researchers apply secure data aggregation to reduce the overhead in smart grids. The data aggregation algorithms can be divided into end-to-end based and hop-by-hop based algorithms according to whether the aggregator can obtain the messages or not. In 2010, Li et al. [4] proposed a secure data aggregation scheme which is based on the homomorphic encryption. Later, many researchers proposed the end-to-end based data aggregation schemes for smart grids. For instance, in 2011, the Elster group [1] provides a proposal for privacy enhancing technology implementation on smart grids. Later, Lu et al. [7] proposed efficient and privacy-preserving aggregation (EPPA) scheme for smart grid communication. Besides the above works, Kamto et al. [2] proposed an aggregation protocol for advanced metering infrastructure (AMI) system, which support both end-to-end model and hop-by-hop model. Furthermore, Li et al. [11] studied the signature issues on smart grids in 2012.

Although homomorphic cryptosystems can protect customers' privacy, generating metering packages has expensive electricity costs which need to be taken into account. Sometimes, the real-time property is more important than the security property. As an illustration, a phasor measurement unit (PMU) submits the data to the server, and the server should be able to analyze the data in a very short period of time such as 30 ms [5]. In this particular case, a lightweight aggregation approach needs to be designed to satisfy real-time requirement. In this paper, we propose a secure and flexible data aggregation framework, which can satisfy different requirements in smart grids such as security property and real-time property. The proposed framework consists of end-to-end aggregation and hop-by-hop aggregation. Compared to hop-by-hop based approach, the end-to-end based approach can achieve more security since the aggregator cannot learn the plaintext in the aggregation phase. On the other hand, the hop-by-hop based approach can do the aggregation faster than the end-to-end based approach because of using symmetric encryption.

The rest of the paper is organized as follows: In the section Framework, we describe our framework in detail. Then, we analyze the proposed framework in section Analysis. Finally, we summarize our results in section Conclusion.

Framework

In this section, the workflow of our framework is described. The proposed framework consists of four parts: secure end-to-end data aggregation, signature aggregation, secure hop-by-hop data aggregation, and MAC aggregation. Figure 1 shows the architecture of the proposed framework.

The secure end-to-end data aggregation permits the nodes to gather information (i.e. power consumption) from other nodes without disclosing their privacy. Additionally, we propose a signature aggregation approach which is compatible with the secure end-to-end data aggregation. Therefore, the overheads can be minimized while delivering security services. Besides, since all real-time information (i.e. PMU information) should be aggregated by a faster approach, we propose a secure hop-by-hop data aggregation and a MAC aggregation approach.

Secure End-to-End Data Aggregation

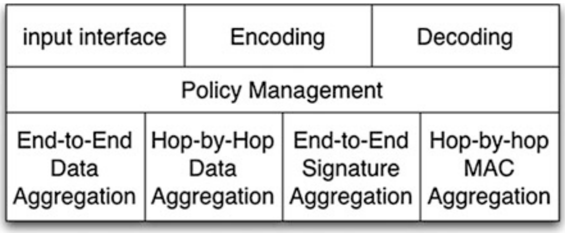
In this approach, one node (either a meter or a sensor) generates a data to be sent to a server. Data from nodes have to be aggregated in order to minimize the overheads. The data are aggregated into one result and sent in cipher to an aggregator using homomorphic encryption with server's public key. Once an aggregator receives the data, the aggregator can apply operations (i.e. summation or multiplication) without knowing server's private key. After calculation, the aggregator sends the aggregated ciphertext to another aggregator, until reaching a control server.

For instance, we can aggregate power consumption information by summing them up together. Therefore, we may use EC-Elgamal encryption [10], which is an additive homomorphic encryption, so that the secure data aggregation can be done without decryption.

EC-Elgamal encryption mainly consists of the following four parts: KeyGen, Encrypt, Aggregate, and Decrypt.

- *KeyGen*(τ): Given a security parameter τ, outputs an elliptic curve $E(F_q)$, where F_q is a finite field, and q is a large prime which is related to τ. Points on the

Fig. 1 The architecture of the proposed framework


input interface
Encoding
Decoding
Policy Management
End-to-End Data Aggregation
Hop-by-Hop Data Aggregation
End-to-End Signature Aggregation
Hop-by-hop MAC Aggregation

elliptic curve $E(F_q)$ form an additive cyclic group G_1, and $ord(G_1) = n$. Choose a generator $G \in G_1, n * G = \infty$. Randomly choose a server's private key $Priv = x$, where $x \in Z_q^*$. Calculate the server's public key $Pub = Y = x * G$. Finally publish (E, G, q, n) as system parameters.

- $Encrypt(M, Pub)$: Given a message M to be encrypted, where $M \in [1, \ldots, q-1]$, and given the server's public key Pub and system parameters (E, G, q, n). One can calculate the ciphertext C by applying the following steps:

1. Choose a random number $k \in Z_n^*$.
2. Ciphertext $C = (R, S) = (k * G, M * G + k * Pub)$.

- $Aggregate(C_1, C_2)$: Given two (or more) ciphertexts C_1 and C_2, the output of aggregating ciphertexst can be calculated by applying the following step:

$$\begin{aligned} C' &= C_1 + C_2 = (R_1 + R_2, S_1 + S_2) \\ &= ((k_1 + k_2) * G, (M_1 + M_2) * G + (k_1 + k_2) * Pub) \end{aligned}$$

- $Decrypt(C, Priv)$: Given a ciphertext C and the server's private key Priv, one can calculate the aggregated plaintext as the equation, $M * G = -Priv * R + S$. The consistency is shown in the following equation:

$$\begin{aligned} M * G = -Priv * R + S &= -x * (\sum_i k_i * G) + (\sum_i M_i * G + \sum_i k_i * Pub) \\ &= -x * (\sum_i k_i * G) + (\sum_i M_i * G + \sum_i k_i (x * G)) \\ &= \sum_i M_i * G = M * G \end{aligned}$$

Signature Aggregation

In this subsection, we show an example of signature aggregation, which is a simple short signature based on bilinear pairings. These kinds of signatures are shorter than traditional ones. Furthermore, signatures can be aggregated and verified in batch processing.

In this approach, every node must sign the message with its private key. Once an aggregator receives messages, the aggregator can either verify them one by one, or verify them in once by applying batch verification process.

In addition, in some particular cases, a node is directly connected to another node instead of an aggregator. And the fact that these nodes have less computational power than an aggregator, all in all affects the execution time. For instance, a group of meters are connected to an aggregator through another meter. In these situations, it is impractical to perform secure data aggregation over a meter. However, it can perform signature aggregation to reduce overheads.

In this example, by using EC-Elgamal encryption, we use short signature, which is based on bilinear pairings, so that the number of system parameters on curve can be reduced.

The simple short signature is generally consists of five parts: KeyGen, Sign, Verify, Aggregation, and Batch–Varify.

- *KeyGen*(τ): This part is quite similar to the *KeyGen* part described in the previous subsection. Therefore, if a proper τ is given, the number of system parameters on curve may not be increased. In Addition, choose a secure hash function $H : \{0,1\}^{*} \rightarrow h,\ h \in G$. Each user U_i, should choose a random number $x_i \in Z_q^*$ as his private key $Priv_i$, and calculate $Pub_i = x_i * G$ as his public key.
- *Sign*$(M, Priv_i)$: Given M and $Priv_i$, where M is a message to be signed and $Priv_i$ is the private key of user U_i.. The result of signing a signature σ can be calculated by the following steps:

1. Calculate the message digest $h_i = H(M)$.
2. Signature $\sigma = Priv_i * h_i$.

- *Verify*(M, σ, Pub_i): Given M, σ, and Pub_i, where M is the message, σ is the signature to be verified, and Pub_i is the public key of user U_i. Determine the validity of the signature by the following steps.

1. Calculate the message digest $h_i = H(M)$.
2. Determine whether $e(\sigma, G) = e(h, Pub_i)$ is equal or not. If the equation holds, then the message M is acceptable. The consistency of a valid signature is shown as follows:

$$e(\sigma, G) = e(Priv_i * h_i, G) = e(h_i, Priv_i * G) = e(h_i, Pub_i)$$

- *Aggregate*(σ_1, σ_2): Given two (or more) signatures σ_1 and σ_2, the output of aggregating signatures can be calculated as follows:

$$\sigma' = \sigma_1 + \sigma_2 = (Priv_1 * h_1) + (Priv_2 * h_2)$$

- *Batch–Verify*$(\sigma, \{M_i\}, \{Pub_i\})$: Given σ, $\{M_i\}$, and $\{Pub_i\}$, where σ is an aggregated signature, $\{M_i\}$ is a list of messages, and $\{Pub_i\}$ is a list of public keys of users. Determine the validity of the signatures as follows:

1. Calculate every message digests $h_i = H(M)$.
2. Determine whether $e(\sigma, G) = \prod_i e(h_i, Pub_i)$ is true or not. If it is true, then accept the messages. The consistency of valid signatures is shown as follows:

$$
\begin{aligned}
e(\sigma, G) &= e\left(\sum\nolimits_i (Priv_i * h_i), G\right) = \prod\nolimits_i e(Priv_i * h_i, G) \\
&= \prod\nolimits_i e(h_i, Priv_i * G) = \prod\nolimits_i e(h_i, Pub_i)
\end{aligned}
$$

Secure Hop-by-Hop Data Aggregation

The previous approach cannot satisfy real-time communication. Some metering messages (i.e. PMU information) should be transmitted in real time (e.g., less than 30 ms). Security and performance are in a trade-off. Therefore, we propose a secure hop-by-hop data aggregation, which is faster than the secure end-to-end data aggregation but less secure, and MAC aggregation, which is also faster than the signature aggregation approach.

In this approach, a node sends the data to a server. The data from nodes have to be aggregated in order to minimize the overheads. The node encrypts the data by symmetric encryption, and then sends the ciphertext to an aggregator. Once an aggregator receives packets, the aggregator first decrypts the packets, and then the aggregator can apply operations (i.e. calculate sum, average, deviation). After encrypting the results, the aggregator then sends the ciphertext to a server through other aggregators.

We use Advanced Encryption Standard (AES) [6] in this particular example. Let $sk_{A_1 n_1}$ denote the session key between an aggregator A_1 and a node n_1.

- n_1 encrypts a message M_1 using AES: $C_1 = E_{sk_{A_1 n_1}}(M_1)$.
- Once A_1 receives i packets C_i, A_1 decrypts ciphertext C_i to its corresponding message: $M_i = D_{sk_{A_1 n_1}}(C_i)$.
- Afterwards, A_1 can use the plaintexts obtained from the previous step to apply operations, such as calculating sum, average, and deviation. Then A_1 stores the result into M_A.
- A_1 encrypts the result message M_A using AES: $C_A = E_{sk_{A_2 A_1}}(M_A)$, where $sk_{A_2 A_1}$ is a session key between A_1 and another aggregator A_2.
- A_2 can repeat step 2 to step 4 until reaching a control server.

MAC Aggregation

In this subsection, we propose an aggregated message authentication code (MAC) method for data integrity. As mentioned before, every node must sign his messages. However, using homomorphic signature is not practical when dealing with transmitting data real-timely. Therefore, we propose a MAC aggregation approach.

In this approach, every node must send their message along with a MAC. The MAC function takes a shared secret key and an arbitrary-length message as inputs. Applying the same session key that is used in secure hop-by-hop data aggregation is also possible.

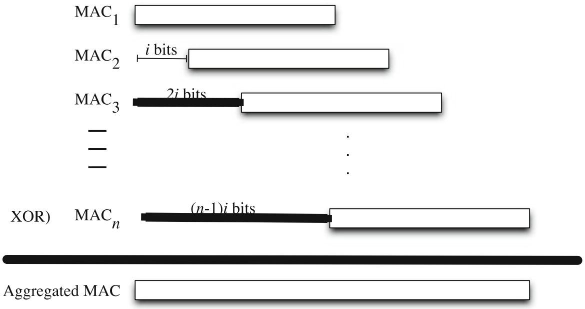


Fig. 2 An example of an aggregated MAC

Once an aggregator receives packets, the aggregator can either verify them one by one, or verify them in once by aggregating MACs.

In some particular cases, a node is directly connected to another node instead of an aggregator. For instance, a group of sensors are connected to an aggregator through another sensor. In these situations, it is impractical to perform secure data aggregation over a sensor. However, it can perform MAC aggregation to reduce overheads. An example of aggregated MAC is shown in Fig. 2.

- A node n_1 calculates MAC_1 by performing MAC algorithm with M and $sk_{A_1 n_1}$.
- An aggregator A_1 can aggregate MAC_1 and MAC_2 by performing the following steps. Basically, we are aware of the order of MACs by applying bitwise shifts as shown in Fig. 1. As a result, in every step of verification process, we know which MAC_i fails. Therefore, we can reduce the time to be spent for re-verification. The aggregation method is described briefly in the following steps:

1. Shift MAC_2 by i bits.
2. Performing exclusive-or (XOR) on MAC_1 and shifted MAC_2.

- A_1 can verify MACs by performing the same MAC algorithm with same inputs to check whether the outputs are equal or not.

Analysis

In this section, we compare our framework with related work in order to show the strength of our scheme.

Examine the Proposed Framework

As mentioned in the previous parts, our framework consists of two main parts: end-to-end based algorithm and hop-by-hop based algorithm. The end-to-end model consists of data aggregation and signature aggregation whereas the hop-by-hop model includes data aggregation and message authentication code (MAC) aggregation instead.

The hop-by-hop model is more efficient but less secure than the end-to-end model, while the hop-by-hop model uses symmetric encryption, and as a result, encryption/decryption phase will be executed faster. The fact that the aggregator can access the users' data, thus, the aggregator must be fully trusted.

On the other hand, the end-to-end model is based on public key cryptosystems where the aggregator can aggregate data without knowing the plaintext even under the honest- but-curious model. Therefore, the end-to-end based can provide more security for users' privacy than the hop-by-hop model. The end-to-end based approach has more computational costs because of inheriting properties from public key cryptosystems.

In order to satisfy data integrity and non-repudiation properties, our framework offered signature aggregation for end-to-end model and message authentication code aggregation for hop-by-hop model.

Compare Our Framework with Previous Works

In [2], Kamto et al. established a protocol for advanced metering infrastructure system (AMI). This protocol assumes that the relay gateways are fully trusted. The relay gateway can decrypt and collect the data from several meters into a single data to be sent to the next substation. But their protocol is insecure under the honest-but-curious model because the relay gateways are able to decrypt the data.

Li et al. [4] designed a secure data aggregation protocol using Paillier cryptosystem to encrypt data. In this protocol, each meter will send the data in ciphertext to the parent node. By using homomorphic encryption, the parent node can aggregate all the ciphertexts from other nodes into one ciphertext to be sent to the collector. This protocol assumes that all smart meters follow the honest-but-curious model. But this protocol is not suitable in some real-time required environment because the end-to-end model requires more computational cost.

Lu et al. (2012) [4] provided the efficient and privacy-preserving aggregation (EPPA) scheme which is constructed under the end-to-end model for smart grid communications. This scheme addressed a multidimensional data aggregation approach based on the homomorphic Paillier cryptosystem. Comparing with the traditional methods, EPPA can reduce computational costs and improve the throughput of communication. Moreover, EPPA satisfies the real-time high-frequency data collection requirements in smart grid communications. Above all, the EPPA scheme can recover the aggregated data into individual data.

Elster group proposed the privacy enhancing technologies (PETs) [1] for the smart grid. These technologies employ the homomorphic cryptographic technique for aggregating data from several nodes. Because these technologies are for the end-to-end model, PETs are not compatible with some real-time systems.

Table 1 Comparison between our framework and previous works

	[2]	[4]	[7]	[1]	Our scheme
End-to-end data aggregation	√	√	√	√	√
End-to-end signature aggregation			√		√
Hop-by-hop data aggregation	√				√
Hop-by-hop MAC aggregation					√
Honest-but-curious model		√	√	√	√
Real-time	Δ		Δ		√

Δ is partially offers the benefit

In this paper, we present a framework that gives a variety of configurations according to different requirements. We can apply the end-to-end model when we need more security, since it is more secure than the hop-by-hop model. On the other hand, for some real-time system, such as PMU system, where the data are delivered in less than 30 ms, we can choose the hop-by-hop model.

We summarize the comparisons of our scheme based on the result of Table 1. As shown in Table 1, all of them can achieve the end-to-end data aggregation. The honest-but-curious model is achieved by all except [2]. However, [4] and Elster [1] cannot provide data integrity and non-repudiation services while our framework can use signature aggregation to deliver these services. Although EPPA [7] scheme can work with many options that listed in the table, it is not appropriate for the PMU system where real-time data transmission is required. By adopting the hop-by-hop model, our framework ensures real-time transmission. Compared with the previous studies, our framework is more flexible in most situations than the others.

Conclusion

In this paper, we propose a secure and flexible data aggregation framework for smart grids. Our framework consists of end-to-end based and hop-by-hop based aggregation approaches. The end-to-end based approach is more secure but less efficient since the aggregator cannot learn any plaintext from all collected data. For this reason, we also design the hop-by-hop based approach which is suitable for real-time requirements. However, the hop-by-hop based approach is less secure than the end-to-end based approach. In addition, in order to resolve data integrity and non-repudiation, we also construct an end-to-end based signature algorithm and a hop-by-hop based MAC aggregation. Consequently, our framework is more flexible in most situations, compared to the previous works.

Acknowledgments The authors would like to thank anonymous reviewers for their valuable comments and suggestions that certainly led to improvements of this paper. This research was partially supported by the National Science Council of the Republic of China under Contract NSC-101-2221-E-007-026-MY3 and NSC-100-2628-E-007-018-MY3. The corresponding author is Professor Hung-Min Sun.

References

1. Elsters proposal for privacy enhancing technology implementation. http://www.elster.com/en/privacy-enhancing-technologies-for-the-smart-grid
2. Kamto J, Qian L, Fuller J, Attia J, Qian Y (2012) Key distribution and management for power aggregation and accountability in advance metering infrastructure. In: Smart grid communications (SmartGridComm), 2012 IEEE international conference
3. Khurana H, Hadley M, Lu N, Frincke D (2010) Smart-grid security issues. Secur Priv IEEE 8(1):81–85
4. Li F, Luo B, Liu P (2010) Secure information aggregation for smart grids using homomorphic encryption. In: Smart grid communications (SmartGridComm), 2010 first IEEE international conference, pp 327–332

11. Li F, Luo B (2012) Preserving data integrity for smart grid data aggregation. In: Smart grid communications (SmartGridComm), IEEE third international conference, pp 366–371

5. Lin H, Deng Y, Shukla S, Throp J, Mili L (2012) Cyber security impacts on all-pmu state estimator: a case study on co-simulation platform GECO. In: Smart grid communications (SmartGridComm), 2012 IEEE international conference. pp 587–592
6. Deamen J, Rijmen V (1998) AES proposal: Rijndael. In: First advanced encryption standard (AES) conference
7. Lu R, Liang X, Li X, Lin X, Shen X (2012) EPPA: an efficient and privacy-preserving aggregation scheme for secure smart grid communications. Parallel Distrib Syst IEEE Trans 23(9):1621–1631
8. Metke A, Ekl R (2010) Security technology for smart grid networks. Smart Grid IEEE Trans 1(1):99–107
9. Parikh P, Kanabar M, Sidhu T (2010) Opportunities and challenges of wireless communication technologies for smart grid applications. In: Power and energy society general meeting, 2010 IEEE. pp 1–7
10. Rabah K (2005) Elliptic curve elgamal encryption and signature schemes. Inf Technol J 4(3):299–306

Data Integrity Checking for iSCSI with Dm-verity

Rui Zhou, Zhu Ai, Jun Hu, Qun Liu, Qingguo Zhou, Xuan Wang, Hai Jiang and Kuan-Ching Li

Abstract With the ever increasing popularity of web service and e-commerce, there is a high demand on data storage. Because of the development of Internet infrastructure and the low cost of deployment, implementing storage over IP has become a trend. For the utilization of network storage, one important issue is the way to achieve data integrity. Usually, the application of Internet Small Computer System Interface (iSCSI), which is a kind of network storage technology, is to store some read-only or important data remotely. I/O requests, which may cause data loss and data error, are frequent in a traditional distributed network storage system like

R. Zhou · Z. Ai · J. Hu · Q. Liu · Q. Zhou (✉)
School of Information Science and Engineering, Lanzhou University Lanzhou, Lanzhou, People's Republic of China
e-mail: zhouqg@lzu.edu.cn

R. Zhou
e-mail: zr@lzu.edu.cn

Z. Ai
e-mail: aiz11@lzu.edu.cn

J. Hu
e-mail: huj11@lzu.edu.cn

Q. Liu
e-mail: liuq11@lzu.edu.cn

X. Wang
School of Science, Lanzhou University of Technology, Lanzhou, People's Republic of China
e-mail: wangxuan2010@lut.cn

H. Jiang
Dept. of Computer Science, Arkansas State University, Arkansas, USA
e-mail: hjiang@astate.edu

K.-C. Li
Dept. of Computer Science and Information Engineering (CSIE), Providence University, Taichung, Taiwan
e-mail: kuancli@pu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_79, © Springer Science+Business Media Dordrecht 2014

iSCSI. In this paper, the data integrity of iSCSI is analyzed and Dm-verity mechanism is utilized to provide read-only transparent integrity checking for iSCSI, which could avert data loss and data error, increasing overall system reliability.

Keywords Dm-verity · iSCSI · Data integrity

Introduction

Dm-verity [1] stands for *device-mapper verity*, which aims to provide read-only transparent integrity checking of block devices. It was originally developed by Google Chromium OS team and introduced later in Linux kernel 3.4.0. Dm-verity is based on device-mapper [2] used to verify the integrity of the root filesystem on boot and supported in many applications, such as LVM, RAID, and multi-path. The core component of dm-verity mechanism is a cryptographic hash tree, in which the leaf nodes of the tree store data blocks, so the hash nodes of the intermediary nodes are calculated based on all of its child nodes and hash function. When some data blocks are accessed, related hash nodes will be verified. In case one of hash nodes fails in the verification, the access will be denied. Therefore, dm-verity could ensure the integrity of data blocks. iSCSI, is an IP based network storage standard for linking data storage facilities. Also like IP-SAN (Storage Area Networks), iSCSI is a standard to encapsulate one SCSI command into a TCP/IP (Ethernet) packet, while users can access the storage system with commodity IP devices only. There are some security issues in iSCSI, since it is a type of network storage and data are transferred by network. In this paper, dm-verity is applied to iSCSI, arming it with an additional layer of security. iSCSI provides two disks, one for data disk of dm-verity, and another one for hash disk of dm-verity. Once the performance is estimated, the research shows that it pays to cost verification overhead to improve the security of iSCSI.

The paper is organized as follows. Section Related Work introduces related work and analyzes the past work of iSCSI security. Section Design and Implementation describes the ways to deploy dm-verity into iSCSI, while section Performance Evaluation evaluates the reading performance of iSCSI with dm-verity embedded, and finally some conclusions are discussed in section Conclusions.

Related Work

There exist limitations in security in present iSCSI, based on the way it works. Mainly they can be categorized as [3]:

1. Active attack (modifying/deleting data, inserting illusive data),
2. Passive attack (listening to Internet lines, data analysis),

A lot of research on storage security has been done to deal with attacks from the above two categories, such as identification and protection of TCP/IP package. As identification, connection authentication is the iSCSI way to determine trustworthiness via CHAP, SRP, SPKM-1 or SPKM-2. This method only protects the identification of target and initiator. On the other hand, in data encryption schemes, two alternatives are considered, IPSecurity Protocol (IPSec) and Secure Socket Layer (SSL). iSCSI uses Cyclic Redundancy Check (CRC) digest to solve the security problems when the 32-bit check word algorithm can ensure end to end checking. iSCSI can have digests for iSCSI headers and data. Header digest can ensure correct operation and data placement, whereas data digest can ensure that data is unmodified throughout network path. Moreover, the performance of iSCSI with CRC has been evaluated in [4].

In particular, some previous works have put forward schemes to ensure the security of network storage. A framework of storage security was proposed in [5]. Analyzing the framework shows that encrypt-on-disk systems not only are more secure but also provide better performance than encrypt-on-wire systems. Other frameworks about iSCSI security have also been brought. For example, a security encryption storage system named ANGLE, which contains two major parts—the Key Management System (KMS) and the Encryption Engine (E-Engine) in [6]. Related transfer protocols have also been target of researches, as analysis of the iSCSI protocol presented in [7], enabling accessing SCSI I/O devices over an IP network.

A number of recent storage security works have been focusing on protecting communication between servers and clients in a distrustful networked world. In particular, the focus is on data integrity: preventing unauthorized modification of data or commands, replaying of requests and modification of requests in transit. Some of these systems further address the issue of privacy, or confidentiality of data transfer: preventing the leaking of data in transit by snooping on the network. Those typical solutions could solve the security problems of identification, provide data integrity to transferred data and evaluate the security of iSCSI, but there are no good measures to make sure whether the important and read-only data of iSCSI application is unmodified or not when being visited. Therefore, it is introduced in this paper a new mechanism, called dm-verity, to protect iSCSI data from being polluted, and mainly fighting against active attacks.

Design and Implementation

Dm-verity, aimed to data integrity protection, has been applied to Chromebooks to protect OS information from being modified. Based on this mechanism, we consider applying it to iSCSI for data integrity.

The iSCSI environment configured with dm-verity is depicted in Fig. 1. It is build up with two targets (target1 and target2) and one initiator. The terminals are connected by a switch and both of the targets will map a disk into the initiator, so

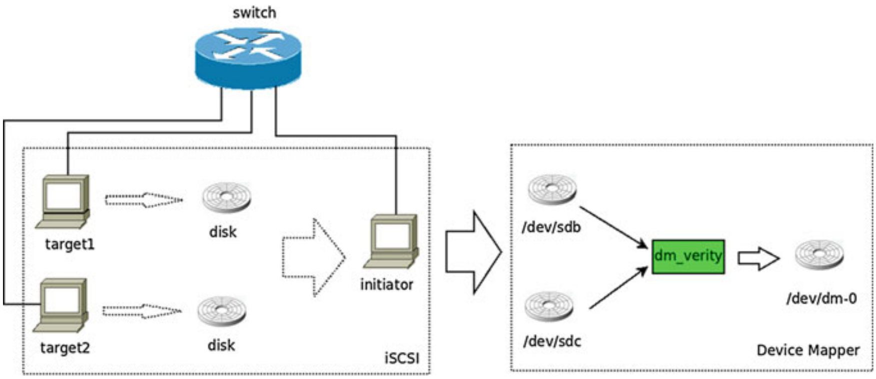


Fig. 1 Experimental system setup

the initiator would have two disks, */dev/sdb* and */dev/sdc*. Then dm-verity mechanism is utilized to configure the integrity checking environment with the disks in initiator.

The configuration information of three computers is as follows:

target1: Ubuntu OS 10.04, Linux kernel 3.0.0, CPU: AMD, 64-bit, 2.2 GHz
target2: Ubuntu OS 10.04, Linux kernel 2.6.35, CPU: AMD, 64-bit, 2.2 GHz
initiator: Ubuntu OS 11.10, Linux kernel 3.4.0, CPU: AMD, 64-bit, 2.2 GHz

iSCSI Configuration

iSCSI technology is very mature and the process of configuration is easy and straightforward. In the experiments, both of the two targets provide two disks for the initiator.

Dm-verity Over iSCSI Configuration [8]

There are four steps to configure dm-verity over iSCSI:

Step 1. Loading related module and preparing for the disks

We insert the *dm-verity.ko* module into kernel. It will be available once the kernel is re-compiled with dm-verity enabled. Also a data disk and a hash disk should be specified. We regard */dev/sdb1* (from target1) as the data disk and */dev/sdc1* (from target2) as the hash disk.

Step 2. Preparing hash device (formatting)

All of the following steps are taken in the initiator. We format the hash device by *veritysetup* command, and then get the root hash value, salt value and other information. After that a hash tree will be generated. The root hash value is the root node of the hash tree.

Step 3. Activating verity data device

Once data device is activated by *veritysetup*, the mapping device will be created. Then there could be a mapped device named */dev/dm-0* from data disk. The most difference between */dev/dm-0* and normal disk is that, the sector capacity of */dev/dm-0* is 4,096 bytes, which is the same as a data block of most system, but the normal size is 512 bytes, and *dm-0* is read-only.

Step 4. Verification

We verify all data block by checking the integrity of block devices through related commands. This process will take several minutes until all nodes of the hash tree are checked. Integrity checking for any specific block does not happen until that block is read, and other data blocks will not be verified. If some blocks to be read are polluted, it will fail to read. Once some blocks that are to be read are correct, while other blocks are broken, the verification will succeed.

Performance Evaluation

In order to evaluate reading performance (writing performance is not available because the disk is read-only) of iSCSI with dm-verity, we take two experiments. First, we evaluate iSCSI without dm-verity. In this experiment, we just set up the iSCSI environment with two targets and one initiator. The formatted and partitioned */dev/sdb* storage device is the mapper device from one target. Finally, we test the reading capability of */dev/sdb1* by *hdparm* tool. Second, we evaluate the reading performance of iSCSI with dm-verity. After configuring the iSCSI with dm-verity mechanism, we would get a virtual disk named */dev/dm-0*, and test the reading capability of */dev/dm-0* in the same method.

We carry out three groups of tests, for each of which the capacity of data disk and hash disk is 2, 4 and 8 GB. The testing results are shown as Fig. 2, in which horizontal axis stands for the times of testing, while vertical axis stands for the throughput of reading. The average throughput without verification are 11.0788, 9.3048 and 10.6691 MB/s in Fig. 2a, b and c respectively. Meanwhile, the average throughputs with verification are 11.042, 9.155, and 10.3053 MB/s, so the average performance overhead are 0.33, 1.61 and 3.41 % respectively based on the three-condition set, which implies a very small performance loss with application of dm-verity.

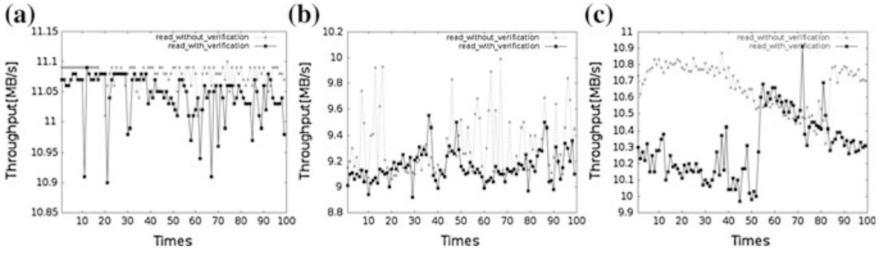
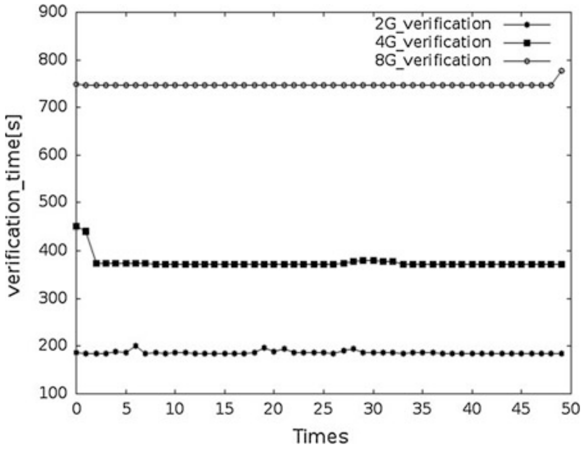


Fig. 2 Reading performance (the capacity of each disk is 2G in (**a**), 4G in (**b**) and 8G in (**c**))

Fig. 3 Verification time



The overhead results from the process of verification costing extra time. Considering the improvement of security, this overhead is acceptable.

Additionally, the bigger the capacity of the disk is, the more time of verification will cost and the larger their margin of read throughput will be. In order to evaluate the relation between verification time and capacity of disk, we carry out another three group experiments—the capacity of data disks also are 2, 4 and 8 GB. The process costs some time, since it must verify from every leaf nodes to root node to assure each hash node of the tree to be verified. This verification could test whether all the data are correct, and testing results are depicted in Fig. 3.

In Fig. 3, horizontal axis stands for the times of verification, and vertical axis stands for verification time. The verification time depends on the size of the data, hash algorithm and network speed [9], and the results are almost stable. In tests performed, the hash algorithm is SHA256. The verification time of 2 GB disk is about 180 s, and the verification speed is about 11.37 MB/s. The verification time of 4 GB disk is about 370 s, and the verification speed is about 11.03 MB/s. The verification time of 8 GB disk is about 750 s, and the verification speed is about 10.9 MB/s. We can see the verification speed is similar to reading throughput.

Conclusions

This paper analyzes the data integrity of iSCSI and utilizes dm-verity to ensure the data integrity of iSCSI. It is an efficient way to protect important data from being polluted. Experiments have shown that application of dm-verity mechanism to iSCSI is feasible with tiny and ignorable performance loss. As some significant data are visited, it can protect the data from being modified and ensuring that callers obtain correct data. Dm-verity for iSCSI not only improves data integrity of iSCSI, but also eases the data management of iSCSI.

Acknowledgments We would like to thank the anonymous reviewers for their constructive comments. This work is supported in part by National Natural Science Foundation of China under Grant No. 60973137, Program for New Century Excellent Talents in University under Grant No. NCET-12-0250, Gansu Sci.&Tech. Program under Grant No. 1104GKCA049, 1204GKCA061 and 1212RJYA003, The Fundamental Research Funds for the Central Universities under Grant No. lzujbky-2012-44, lzujbky-2013-k05, lzujbky-2013-43 and lzujbky-2013-44, Google Research Awards, Google Faculty Award, China and National Science Council (NSC), Taiwan, under grant NSC101-2221-E-126-002.

References

1. Corbet J (2011) dm-verity overview: http://lwn.net/Articles/459420/ LWN: 9/19
2. Device-mapper Resource Page: http://sourcewre.org/dm/
3. Kaufman C, Perlman R, Speciner M (2002) Network security: private communication in a public world, 2nd edn. Prentice Hall, New Jersey
4. Wendt J, Thaler P, Satran J, Shimony I, Makhervaks V (2010) iSCSI-R data integrity: http://www.research.ibm.com/haifa/satran/ips/iscsi-r-data-integrity-v1d.pdf
5. Riedel E, Kallahalla M, Swaminathan R (2002) A framework for evaluating storage system security. In: The 1st USENIX Conference on File and Storage Technologies, pp 15–30, USENIX Association, Berkeley, CA
6. Di CY, Li KC, Hung JC, Yu Q, Zhou R, Hung CH, Zhou QG (2013) A case of security encryption storage system based on SAN environments: intelligent technologies and engineering systems lecture notes in electrical engineering, vol 234. Springer, Heidelberg, pp 27–32
7. Meth KZ, Satran J (2003) Design of the iSCSI protocol. In: The 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSS'03), pp 116–122, IEEE Computer Society, Washington, DC
8. Linux Unified Key Setup (2012) http://code.google.com/p/cryptsetup/wiki/DMVerity 6/12
9. Wu GF, Wu HH, Chang L (2011) On performance optimization of iSCSI SAN. In: 3rd International Conference On Information Technology and Computer Science (ITCS 2011), pp 430–433, ASME, New York

Opportunistic Admission and Scheduling of Remote Processes in Large Scale Distributed Systems

Susmit Bagchi

Abstract The large scale loosely-coupled distributed systems such as, grid and cloud computing systems employ opportunistic execution mechanism of remote processes in order to utilize computing resources of idle nodes. The opportunistic admission and scheduling of remote processes at a node need to balance the enhanced resource utilization and the performance of local processes at the node. This paper proposes the design and implementation of a novel Admission Control and Scheduling (ACS) algorithm for opportunistic execution of remote processes in a distributed system based on online estimation method. The experimental results illustrate that the algorithm can schedule the CPU-bound and IO-bound remote processes without degrading overall performance of a node. The CPU-utilization and memory-utilization of a node are enhanced by 26.65 and 24.5 % respectively on the average without degrading the performance of local processes executing at the node.

Keywords Distributed systems · Opportunistic scheduling · Online estimation

Introduction

The large scale loosely-coupled distributed computing systems such as, grid computing and cloud computing systems are gaining attention as next generation computing platforms due to the availability of computing machines as well as Internet at lower cost [1, 2]. The large scale distributed computing systems can be viewed as the virtual supercomputer infrastructure offering location transparent execution and faster response time [1, 3–6]. However, the statistical observations

S. Bagchi (✉)
Department of Informatics, Gyeongsang National University, Jinju, South Korea
e-mail: susmitbagchi@yahoo.co.uk

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_80, © Springer Science+Business Media Dordrecht 2014

about distributed resource utilizations have revealed that 50 $\sim$ 70 % of computing resources remain idle or unutilized on the average in the large scale grid and cluster computing systems [7–10]. In the large scale distributed computing systems, the scheduling of the remote processes remains an important research issue [2, 3, 11–13]. The global scheduling algorithms are often employed in grid computing to enhance throughput, to scale up the global resource utilization and to manage idle nodes [7, 8]. However, the global scheduling of processes does not consider the dynamics of available resources at individual nodes. Due to this reason, the opportunistic scheduling of processes tends to degrade the performance of existing local processes at a node. The local scheduling of remote processes in large scale distributed computing systems has received less attention [7, 9]. It has been proposed that scheduling of processes should be done at two levels such as, at local cluster level and at grid level [3, 14]. Thus, admission control of remote processes and local scheduling of the remote processes at a node are the two important research areas with the aim to maximize the resource utilization of a node while maintaining the throughput of the local processes. This paper proposes the novel Admission Control and Scheduling (ACS) algorithm based on online estimation of available resources and the process-load in a system. The proposed architecture utilizes the kernel address-space virtual devices and user-space application to create a remote process execution framework in a node in the distributed systems. The goals of the proposed design framework are, (1) *enhancing the utilization of idle resources in a node* and, (2) *maintaining load-balance between local as well as remote processes and system resources of a node preventing the system from thrashing or resource-starvation*. Rest of the paper is organized as follows. Section Related Work describes related work. Section Resource Estimation and Algorithm Design depicts the design model and description of the algorithm. Section Experimental Evaluations presents implementation and experimental evaluation. Section Conclusion concludes the paper.

Related Work

The wide array of computing applications employ large scale loosely-coupled distributed computing systems. The examples of such systems are grid, cluster and cloud computing systems [1, 2, 8, 15]. In the grid computing systems, the global scheduling of jobs in the batches is one of the main research challenges. The scheduling of jobs in such large scale distributed systems can be broadly classified into three classes namely, periodic resource allocation based scheduling, opportunistic scheduling and, reliability-based scheduling. Furthermore, the reliability-based scheduling algorithms can be classified into two broad groups such as, hierarchical scheduling and, genetic algorithm based scheduling. In the Periodic Resource Reallocation (PRR) based global scheduling technique, the resource broker maps the jobs to the available computing nodes in multiple stages [1]. The periodic reinforcing and reclamation algorithm schedules the jobs at the nodes to minimize the execution time of the jobs at each stage [1]. However, the algorithm is prone to frequent process migration through checkpointing, which is an expensive operation. Moreover, the migration of processes between heterogeneous systems would degrade the system performance. On the other hand, the opportunistic scheduling model aims to utilize the resources of idle nodes in a distributed system [7, 8, 16]. The dilation of response time of short jobs is reduced by opportunistic scheduling (OPS) algorithm proposed in [8]. The OPS algorithm is an augmented form of Up-Down algorithm [7]. The algorithm reduces the mean slowdown of processes in a system by rotating the jobs in a batch within a queue [8]. However, the algorithm is based on the assumptions that, local processes at a node have exponential inter-arrival time and, the distribution of local process-load in a system is hyper-exponential in nature [8]. These are fairly rigid assumptions in case of real-life execution environments. If the arrival rate of processes varies significantly then, the performance of the opportunistic algorithm degrades quickly with increasing remote process-load. The Condor distributed system employs Up-Down algorithm to implement fair access to remote resources for executing processes [7]. The performance of the system is further improved by introducing Multilevel Queue Opportunistic (MOF) policy [16]. However, the policy is dependent on workload model, where the processing load estimation assumes hyper-exponential distribution.

The hierarchical reliability-driven (HRD) job scheduling algorithms have gained research attention in recent time. The hierarchical reliability-driven scheduling algorithm employs two stages to schedule the jobs in a grid based on the distributed network reliability model [2, 5]. The first stage scheduling decision is based on local schedulers controlling a subset of nodes and the second stage of scheduling is done by global scheduler located at a central server [2]. The hierarchical scheme aims to enhance system reliability through distributed decision making in a computing system. The main two disadvantages of the hierarchical scheme are employment of non-preemptive scheduling model by the local scheduler and, possibility of single-point of failure due to centralized global scheduler system. In another approach, the reliability-aware genetic algorithm is proposed, where the transmission time and waiting time in resource queue are taken into consideration while making scheduling decisions [15]. The sampling model employed in the genetic algorithm is stochastic in nature in order to speed up the convergence. The stochastic universal sampling enhances the performance of the algorithm over the other genetic algorithms [15]. However, the algorithm is computationally expensive. The master–slave architecture is often employed for the realization of grid scheduling of jobs in batches. A multilevel hybrid scheduling algorithm is proposed based on master–slave model [3]. The algorithm uses two different schemes such as, single queue model and dual queue model of task scheduling. Although the algorithm enhances performances however, one of the major disadvantages of the algorithm is the centralized master–slave architectural model, which is prone to the single point of failure.

Resource Estimation and Algorithm Design

The main two computing resources required by any process are CPU and main memory, where the availability of both can vary randomly in time. The online estimation of the computing resources in a system is carried out based on the time-windows having fixed quanta $\varepsilon > 0$ and having total memory M, a constant. Let, P(t) is the total number of processes existing in a node in the distributed computing system at time t, whereas n(t) and p(t) represent the remote process load (RPL) and local process load (LPL) in the node at time t, respectively. Let, the instantaneous memory-load of the node is denoted by function m(t) and the CPU-load is denoted by function c(t), where both are dimensionless. Hence, the instantaneous ratio of number of remote processes and local processes including the administrator process (i.e. the process which admits the remote processes in a node) can be given by,

$$y(t) = [n(t) + 1]/[p(t) + 1] \quad (1)$$

The ratio of number of local processes including the administrator process and total number of processes at time t is given by,

$$z(t) = [p(t) + 1]/P(t) \quad (2)$$

Accordingly, the memory-load at time t is given by, m(t) = u(t)/M, where u(t) is the amount of used memory at time t. Thus, the process-load at a node is computed as,

$$c(t) = [n(t) + 1]/p(t) \quad (3)$$

From Eqs. (1) and (3),

$$c(t) = [y(t)(p(t) + 1)]/[P(t) - y(t)(p(t) + 1)] \quad (4)$$

From Eqs. (2) and (4),

$$c(t) = [y(t)(p(t) + 1)]/[P(t)z(t) - 1] \quad (5)$$

Now, from Eqs. (4) and (5), it can be further derived as,

$$P(t) = [(y(t)(p(t) + 1)) - 1]/[1 - z(t)] \quad (6)$$

It is evident from Eq. (6) that, it is a balancing equation representing the dynamics of process-load on a CPU of a system. Thus, dynamics of the co-variation of the instantaneous memory-load and process-load in a system at time t is computed by,

$$h(t) = m(t) - c(t) \quad (7)$$

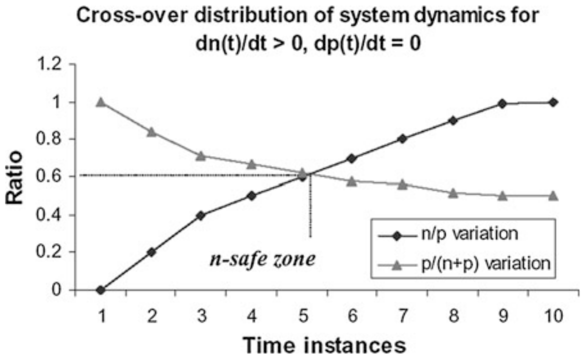


Fig. 1 Dynamics of cross-over distribution for monotonically increasing RPL

The difference equation for any time-window $\varepsilon = t_2 - t_1$ can be computed from equation (7) as,

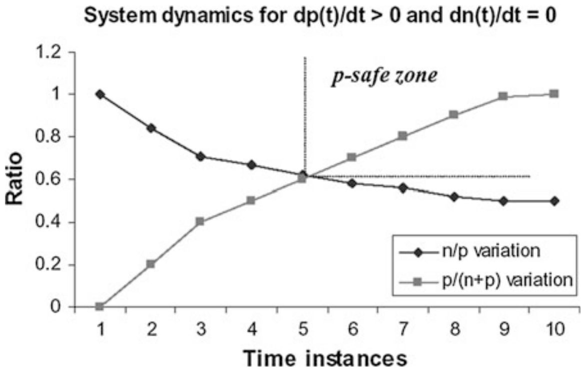
$$\Delta h(\varepsilon) = h(t_2) - h(t_1) = [m(t_2) - m(t_1)] - [c(t_2) - c(t_1)] \qquad (8)$$

Hence, for the sufficiently small ε, $lim_{\varepsilon \to 0}(\Delta h(\varepsilon)/\varepsilon) = lim_{\Delta t \to 0}(\Delta h(\Delta t)/\Delta t) = d/dt(h(t))$, where $\Delta t = t_2 - t_1$. This indicates that Eq. (8) computes the dynamics of resource-load in a system as the differential on continuous time plane for sufficiently small quanta. However, for larger quanta, Eq. (8) transforms into difference equation computing the dynamics of resource-load variations in the discrete time plane. In the real-life systems, the LPL and RPL can vary randomly in time. The cross-over distribution of monotonically increasing RPL with respect to time t (dn(t)/dt > 0) along with static LPL (dp(t)/dt = 0) is illustrated in Fig. 1. It is evident that, on the left side of cross-over point (n-safe zone), the resource utilization by remote processes is ensured without creating the possibility of thrashing. On the other hand, if LPL increases monotonically in time along with static RPL, then mainly the local processes utilize the system resources on the right side of the cross-over point (p-safe zone) as illustrated in Fig. 2. Thus, a computing node requires the balance of the admixture of load by keeping local processes on upper right side of the cross-over point and remote processes on the lower left side of the cross-over point on their respective curves.

Designing the Algorithm

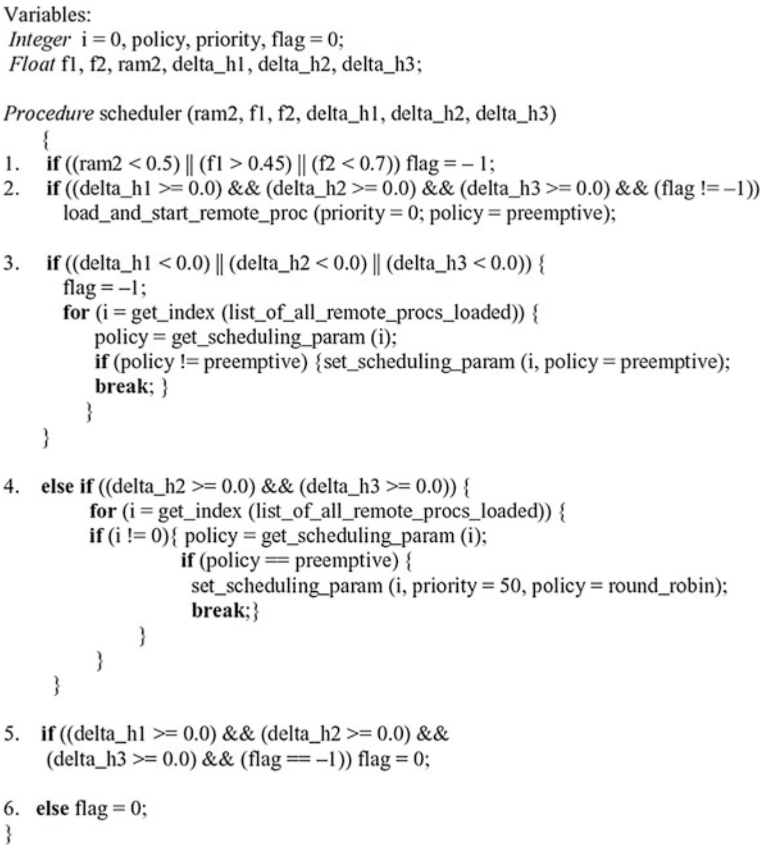
The ACS algorithm is comprised of two logical components namely, the online periodic estimation and the scheduler. The estimator periodically makes online estimation of system resources available to the local as well as remote processes. On the other hand, the scheduling algorithm takes the estimated parameters as input and controls the admission as well as scheduling of the remote processes in a node. The pseudo-code representation of the ACS algorithm is illustrated in Fig. 3,

Fig. 2 Dynamics of cross-over distribution for monotonically increasing LPL



where delta_h1, delta_h2 and delta_h3 represent estimated values of Eq. (8) in three time-windows.

The algorithm does not assume any fixed number of local processes in a node and, it controls the admission of remote processes based on the instantaneous number of local processes in a node at any point of time. One of the aims of the algorithm is to restrict the possibility of creation of monopoly of remote processes in a system leading to starvation of the existing local processes. On the other hand, the algorithm aims to enhance the overall resource utilization in a system by admitting the remote processes following the controlled proliferation technique. According to the resource estimation model, 50 % of instantaneously available main memory (i.e. in n-safe zone below the cross-over point) is kept free as upper bound (ram2 in Fig. 3) in order to avoid memory-overload and thrashing due to sharp rise of local processes in a short duration of time. However, the ratio of number of remote processes and local processes (f1 in Fig. 3) is kept near to cross-over point at 0.45 at any point of time in order to control the load-balancing in a node. On the contrary, if the ratio of number of local process and total number of processes in a system (f2 in Fig. 3) falls below 70 % (i.e. in p-safe zone above the cross-over point), the algorithm stops admission of remote processes into the system to achieve load-balancing. If any of the aforesaid three conditions are satisfied, then the algorithm temporarily stops admitting remote processes in a node by changing the flag (line 1). If the estimated values of co-variance of resource-load in a system are non-negative and system is lightly loaded, then remote processes are loaded one by one from the queue of processes (line 2). Initially, all the newly loaded remote processes are allocated zero priority and the remote processes run following the preemptive scheduling policy. However, if the system is on the process of getting lightly-loaded due to completion of some of the local or remote processes, then the blocking of the remote-process admission into the node is cleared by setting the flag (line 3). The set_scheduling_param procedure in Fig. 3 changes the scheduling parameters of the corresponding process as mentioned in the argument. In order to allow greater time-slicing between local and remote processes in a node, the remote processes running in non-



```
Variables:
 Integer  i = 0, policy, priority, flag = 0;
 Float f1, f2, ram2, delta_h1, delta_h2, delta_h3;

Procedure scheduler (ram2, f1, f2, delta_h1, delta_h2, delta_h3)
      {
1.    if ((ram2 < 0.5) || (f1 > 0.45) || (f2 < 0.7)) flag = – 1;
2.    if ((delta_h1 >= 0.0) && (delta_h2 >= 0.0) && (delta_h3 >= 0.0) && (flag != –1))
        load_and_start_remote_proc (priority = 0; policy = preemptive);

3.    if ((delta_h1 < 0.0) || (delta_h2 < 0.0) || (delta_h3 < 0.0)) {
        flag = –1;
        for (i = get_index (list_of_all_remote_procs_loaded)) {
            policy = get_scheduling_param (i);
            if (policy != preemptive) {set_scheduling_param (i, policy = preemptive);
            break; }
           }
      }

4.   else if ((delta_h2 >= 0.0) && (delta_h3 >= 0.0)) {
          for (i = get_index (list_of_all_remote_procs_loaded)) {
          if (i != 0){ policy = get_scheduling_param (i);
                      if (policy == preemptive) {
                        set_scheduling_param (i, priority = 50, policy = round_robin);
                        break;}
                  }
             }
        }

5.   if ((delta_h1 >= 0.0) && (delta_h2 >= 0.0) &&
      (delta_h3 >= 0.0) && (flag == –1)) flag = 0;

6.  else flag = 0;
}
```

Fig. 3 Pseudo-code of the admission control and scheduling algorithm

preemptive mode are changed to preemptive mode of execution. On the other hand, if the system continues to be in lightly loaded state in three consecutive time-window frames, then the priorities of remote processes are increased to 50 one by one and the scheduling policy is changed to round-robin in order to speed up the processing (line 4). Lastly, if the system had entered into highly-loaded stage previously and currently is in lightly-loaded stage in three consecutive time-window frames, then the blocking of the admission of remote process into a node is cleared by resetting the flag (line 5).

Experimental Evaluations

The implementation of the algorithm and other software components are made in Linux 2.6 kernel using C language. The implementation is comprised of one user-space daemon and one kernel-space module having secured communication channel between them. Experiments are carried out on the platforms having dual core Pentium processor equipped with 4 GB RAM and 250 GB disk drive connected to 100 Mbps wired network. The list of concurrent local process load (LPL) and, multiple instances of user applications are illustrated in Table 1, where the processes are characterized as IO-bound and CPU-bound depending on their nature. The experimental observations are made at a node in four phases such as, Phase I: RPL < 50, Phase II: $50 <$ RPL < 100, Phase III: $100 <$ RPL < 150 and, Phase IV: Stable system. A system is called stable when ACS algorithm temporarily stops admission of remote processes depending upon the dynamics of resource-load in the system and waits for some of the executing processes to complete execution before admitting any remote processes further in the system.

Table 1 The distribution of LPL and RPL in a system

Process load	Admission	Type	Locality
• System processes • Adobe photoshop Processes with multiple instances: • Multimedia editor • Spawning application • Media streaming	Pre-existing for concurrent execution	• CPU and IO bound • CPU-bound and memory-loading • CPU-bound and memory-loading • CPU-bound • IO-bound	LPL
Type I (multiple instances): • Video encoder and streaming • DSP simulator • Big-data encryption algorithm • Continuous-fork application Type II (multiple instances): • Repeated file IO application • Streaming server application • Memory-grabber application	From queue for concurrent execution	Type I: CPU-bound and memory-loading Type II: IO-bound and memory-loading	RPL

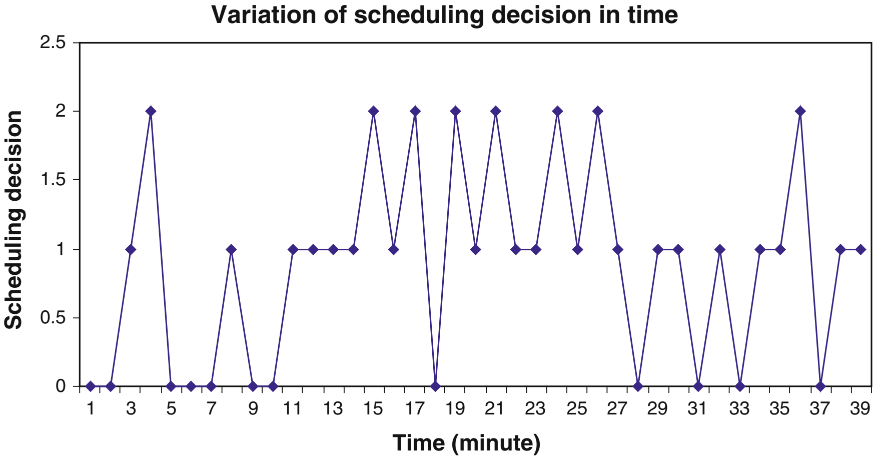


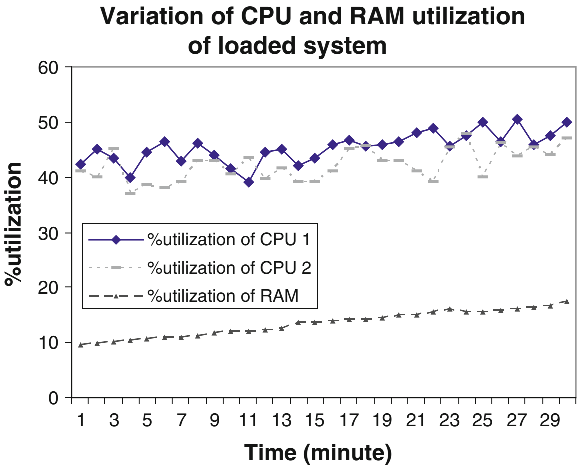
Fig. 4 Snapshot of variations of scheduling decisions for RPL <50

Phase I

The snapshot of the variations of admission and scheduling decisions taken by ACS algorithm is illustrated in Fig. 4, where on y-axis numerical value 0 denotes the reduction of priority and changing the scheduling policy of remote processes, 1 denotes the admission of remote processes and, 2 denotes the increment of priority and changing the scheduling policy of remote processes. The variation of average % utilization of the system resources by LPL and RPL is illustrated in Fig. 5.

It is evident from Fig. 4 that, in Phase I, initially a node behaves as highly unstable system due to repeated admission of remote processes in the system,

Fig. 5 Snapshot of variation of %utilization of CPUs and RAM for RPL <50



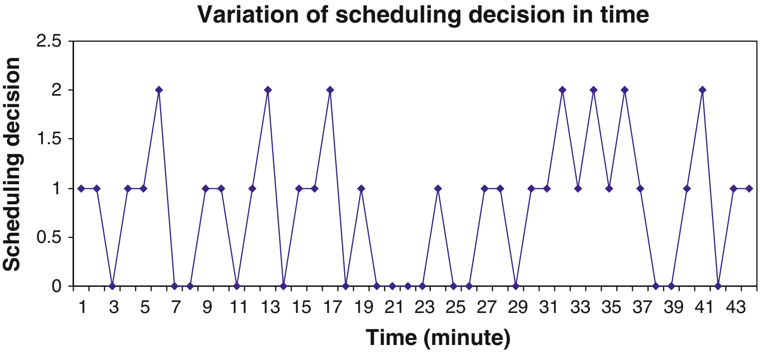


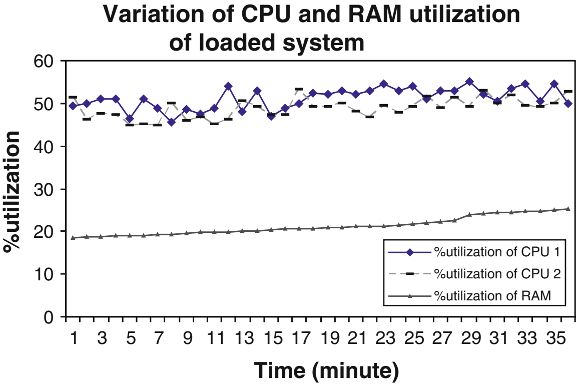
Fig. 6 Snapshot of time variation of scheduling decision for 50 < RPL < 100

whereas the resulted utilization of main memory of the system increases monotonically. However, occasionally, the ACS algorithm reduces the priority of remote processes in order to maintain the resource-demand and performance of the local processes. In this phase, on the average CPU 1 and CPU 2 are utilized up to 50 and 46.7 %, respectively. However, the main memory utilization is 21.4 % on the average.

Phase II

The snapshot of the variations of admission and scheduling decisions taken by ACS algorithm in this phase is illustrated in Fig. 6, where on y-axis numerical value 0 denotes the decrement of priority and changing scheduling policy of remote processes, 1 denotes the admission of remote processes and, 2 denotes the

Fig. 7 Snapshot of variation of %utilization of CPUs and RAM for 50 < RPL < 100



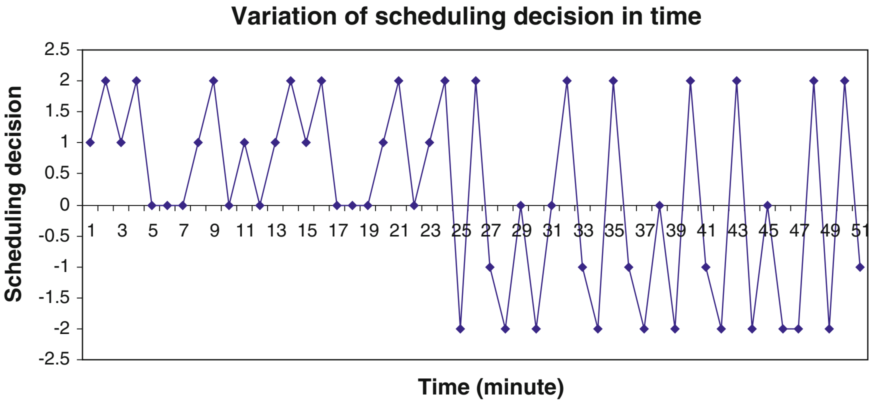


Fig. 8 Snapshot of time variation of scheduling decision for 100 < RPL < 150

increment of priority and changing scheduling policy of remote processes. The average % utilization of the system resources by LPL and RPL in Phase II is illustrated in Fig. 7. In this phase, Fig. 6 illustrates that, due to enhanced processing-load and resource-demand on the system, the kernel is reclaiming the idle resources (such as, fragmented memory pages) and making them available to LPL as well as RPL executing in the system. This phenomenon induces the frequent transitions between two extreme scheduling decisions as depicted in Fig. 6.

However, according to Fig. 7, the overall utilization of resources is monotonically increased. In this phase, on the average CPU 1 and CPU 2 are utilized up to 51.5 and 45.3 %, respectively. However, the main memory utilization is increased to 22.7 % on the average.

Phase III

The snapshot of the variations of admission and scheduling decisions taken by ACS algorithm in this phase is illustrated in Fig. 8, where on y-axis numerical value 0 denotes the reduction of priority and changing scheduling policy of remote processes, 1 denotes the admission of remote processes, 2 denotes the increment of priority and changing scheduling policy of remote processes, −1 denotes the readmission of remote processes and, −2 denotes stopping admission of remote processes. The average % utilization of the system resources by LPL and RPL in Phase III is illustrated in Fig. 9. According to Fig. 8, through the kernel activation the resource-availability is increased and the ACS algorithm rapidly admits remote processes into the system while increasing scheduling priorities for faster execution of remote processes. However, due to variations of LPL, the ACS algorithm later stops admitting any RPL further when the scheduling decision makes frequent transitions between two extremes in the highly-loaded system in Phase III.

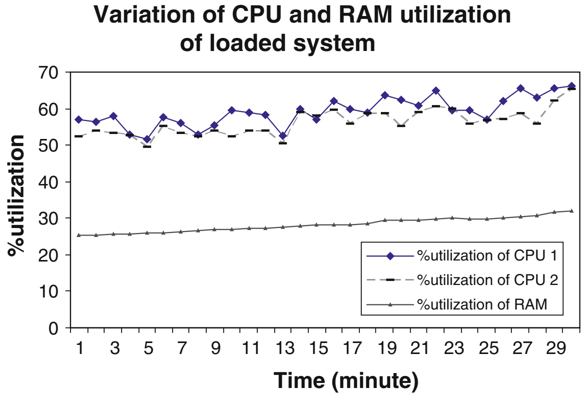
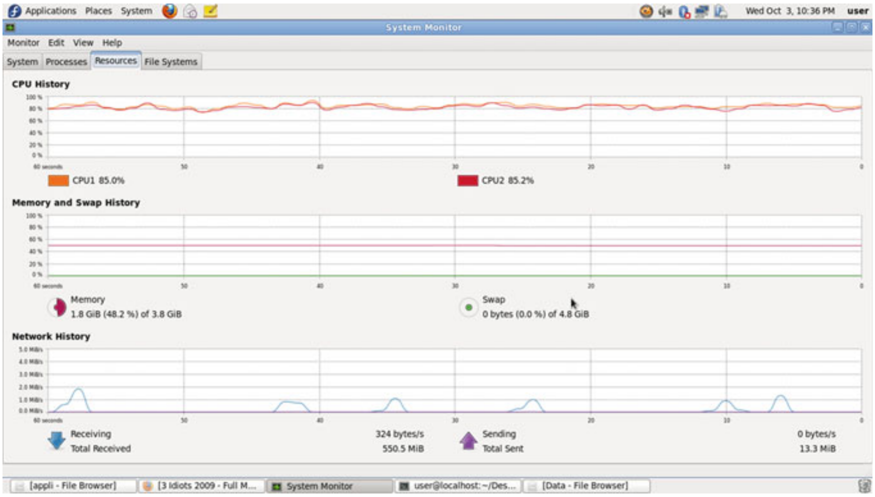


Fig. 9 Snapshot of variation of %utilization of CPUs and RAM for 100 < RPL < 150

In this phase, utilization of CPUs has increased to 78.6 and 73.4 % on the average, whereas the main memory utilization is increased to 40.5 % on the average scale.

Phase IV

This is the phase of temporary stability, where the ACS algorithm temporarily stops any admission of remote processes into the system allowing the LPL and existing RPL to execute enhancing resource utilization without causing memory-overloading and thrashing. The screenshot of resource monitor of the system in



🟧 CPU1 85.0%   🟥 CPU2 85.2%

Memory
1.8 GiB (48.2 %) of 3.8 GiB

Swap
0 bytes (0.0 %) of 4.8 GiB

Receiving 324 bytes/s
Total Received 550.5 MiB

Sending 0 bytes/s
Total Sent 13.3 MiB

Fig. 10 Screenshot of resource monitor for stable system with RPL equal to 150

maximally-loaded stable phase is illustrated in Fig. 10, where RPL is equal to 150 after 3 h of execution.

The CPU utilization is enhanced to 85 and 85.2 % at maximum for CPU 1 and CPU 2, respectively. The main memory utilization is increased up to 48.2 %. The performances of local processes remain unaffected.

Conclusion

The global scheduling algorithms are employed to determine the location of execution of remote processes in large scale distributed computing systems. However, the admission control and scheduling of remote processes locally at the nodes are required to maintain the load-balance between local processes and remote processes, while enhancing resource utilization of a node. The proposed lazy invocation based ACS algorithm implements opportunistic admission and scheduling of remote processes in a node. The algorithm enhances the over all CPU and memory utilizations up to 85.2 and 48.2 % during the phase of temporary stability. The average value of resource utilization is increased by 26.65 and 24.5 % for CPU and memory, respectively, without affecting the performance of the local processes. The algorithm is easy to realize and computationally inexpensive.

References

1. Lin CC, Shih CW (2008) An efficient scheduling algorithm for grid computing with periodical resource reallocation. In: 8th IEEE International Conference on Computer and Information Technology, IEEE
2. Tang X, Li K, Qiu M, Sha EHM (2011) A hierarchical reliability-driven scheduling algorithm in grid systems. J Parallel Distribut Comput 72(4):525–535
3. Shah SNM, Mahmood AKB, Oxley A (2010) Analysis and evaluation of grid scheduling algorithms using real workload traces. In: ACM MEDES'10, ACM
4. Wieczoreka M et al (2009) Towards a general model of the multi-criteria workflow scheduling on the grid. Future Generation Computer Systems Journal, Vol. 25, No. 3, Elsevier (2009)
5. Fan H, Sun X (2010) A multi-state reliability evaluation model for P2P networks. Reliab Eng Syst Saf J 95(4): 402–411
6. Fernandez-Baca D (1989) Allocating modules to processors in a distributed system. IEEE Transact Softw Eng 15(11): 1427–1436
7. Litzkow MJ, Livny M, Mutka MW (1988) Condor: a hunter of idle workstations. In: 8th IEEE International Conference on Distributed Computing Systems, IEEE
8. Ghare GD, Leutenegger ST (2000) Improving small job response time for opportunistic scheduling. In: 8th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, IEEE
9. Mutka M, Livny M (1991) The available capacity of a privately owned workstation environment. Perform Eval J 12(4), 269–284

10. Tchernykh A et al (2009) Idle regulation in non-clairvoyant scheduling of parallel jobs. Discrete Appl Math J 157(2): 364–376
11. Livny M, Basney J, Raman R, Tannenbaum T (1997) Mechanisms for high throughput computing. SPEEDUP J 11(1)
12. Zhihong X, Xiangdan H, Jizhou S (2003) Ant algorithm-based task scheduling in grid computing. In: CCECE'03, IEEE
13. Kiran N, Maheswaran V, Shyam M, Narayanasamy P (2007) A novel task replica based resource scheduling algorithm in grid computing. In: The 14th HiPC Conference
14. Li H (2009) Workload dynamics on clusters and grids. J Supercomputing 47(1): 1–20
15. Abdulal W, Ramachandran S (2011) Reliability-aware genetic scheduling algorithm in grid environment. In: International Conference on Communication Systems and Network Technologies, IEEE
16. Abawajy JH (2002) Job scheduling policy for high throughput computing environments. In: 9th International Conference on Parallel and Distributed Systems, IEEE

Feasible Life Extension Concept for Aged Tactical Simulation System by HLA Architecture Design

Lin Hui and Kuei Min Wang

Abstract The military owned simulation systems are facing legacy situation causing sky-high maintenance cost and unstable condition. It would be costly and risky for retiring them without considering life extension alternative. The objective of this research is to provide a feasible concept for migrating the structure of aged Tactical Training Simulation (TTS) system to High Level Architecture (HLA), which not only can retain the original feature but also offer an extra function of interoperability. This research offers a solution for military legacy simulation systems and paves the way for creating the synthetic battlefield environment that military can benefit its training, planning, acquisition and resource allocation.

Keywords HLA · Interoperability · PDU · RTI · Simulation · TTS

Introduction

The original Tactical Training Simulation (TTS) system has been built for over decades being not only uneasy to maintain but also very hard to have the merit of interoperability and reusability for this synthetic environment. The causes of that are the out of date of aging hardware and software system for being uneasy maintained which is even costly, imbedded models in the system with its limitation in function expansion which includes new forces add in and exercises size

L. Hui
Department of Innovative Information and Technology, Tamkang University, Tamsui, Taiwan, Republic of China
e-mail: amar0627@gmail.com

K. M. Wang (✉)
Department of Information Management, Shih Chien University, Taipei, Taiwan, Republic of China
e-mail: willymarkov@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_81, © Springer Science+Business Media Dordrecht 2014

expanding, as well as unable to interact with other simulation systems which is one of the most important abilities in the joint exercise simulation environment.

This paper proposes a way to migrate aging simulation system to the interoperability-based High Level Architecture (HLA). In addition to the hardware that still in use in the system, the function of interoperability for meeting the reality of live (real people with real equipment in the field), virtual (real people with simulated equipment) and constructive (simulated people using simulated equipment) in simulation world is also a major concern.

The latest trend with high fidelity synthetic environment, based on HLA support trainings and exercises with virtual forces, has replaced the most of live forces in the field that not only can cut down the military training cost hugely but also reduce the probable casualties in the unexpected incidents.

Based upon the migration strategy of legacy system [1], study of TTS has been carried out in sequence as follows: trace TTS framework for understanding its original design concept and function of modules; the second is to check the way of message interchanged in the system; the third is to analyze HLA framework for the purpose of transforming it into the TTS. The purpose of the first two sequences is to restore TTS original technical and functional features for upgrades or migration concerns. The third is mainly for interoperability concern dressing TTS with interaction capability.

The upgraded TTS will have a new system procedure according to the six services of HLA Run Time Infrastructure (RTI). In the process, the primitive TTS military functions would be all retained, other than that the more flexible and convenient feature will be presented with the function of add-in, interoperability and reusability.

Literature Review

The aging Tactical Training Simulation system has fatal defects from the perspective from this decade, which is very hard to keep it operational when any system-malfunctions, such as expanded size of exercises and the new military assets that designed and setup in the system, occurred. The U.S.A. Army's Enhanced REmoted Target System (ERETS) has the operational problems caused by over the system life cycle that system components had no longer been available or repaired [2]. Most of the Tactical Training Simulation systems are unique and independent from each other due to they came from different vendors under without standards set by government authorities that make these systems become lower performing value, decomposability, obsolescence and deterioration, after years in service and it is called the legacy system [3].

U.S. Department of Defense (DoD) regards the migration of legacy systems as critical issues for there are so many systems have been in use by military for so many years. The legacy systems produced by various companies with different software that most of the systems are required to be upgraded or migrated to new hardware and software platforms. Therefore, "DoD Legacy System Migration Guidelines" is issued for being as an aid in the migration. There are 10 guidelines in total for risk-reduction while in the process of the migration of legacy systems. Guideline #6 indicates: Make software architecture a primary reengineering consideration, and define that a methodical evaluation of the software architectures of the legacy and target systems should be a driving factor in the development of the reengineering technical approach [1, 4, 5].

Well-known protocols for this paradigm include Distributed Interactive Simulation (DIS) [6], Aggregate Level Simulation Protocol (ALSP) [7], and the High Level Architecture (HLA) [8]. The HLA is an ameliorative technique that combines the operational concepts of DIS and ALSP, as well as other distributed simulation issues, into a standardized infrastructure. The technical innovation of HLA allows a radical upward scaling of the number of distributed simulation nodes to participate in a single scenario.

The HLA is a project initiated by the U.S. DoD to support interoperability among geographically distributed simulators. It later became an international standard, IEEE 1516, in 2001 [8]. The HLA defines an infrastructure to support the reusability and interoperability among heterogeneous simulations. To achieve this goal, the HLA defines a time management service for a simulation system to coordinate its execution pace with others. The time management service defines the synchronization mechanism to ensure causal order consistency of the event sequence among distributed simulation nodes. In HLA terminology, each distributed node of a simulation is called a federate. Furthermore, a federation is a simulation environment that consists of a set of federates. Depending upon the simulating model, a federate can use the time-stepped or event driven time advancing technique [9, 10].

Concept of TTS Operation

After the current TTS operation being examined, the documented concept of fundamental operation can be classified into five steps that they are Pre-setup before system initialization, initialization, communication link establishment, message transmission and termination.

The exercise is initiated after the system completion of its initialization. There are a series of weapon deployments, detection message transmitting etc. are followed for being set up. When an exercise is about to finish, the 'end' button can be chosen on the exercise screen of each computer. Then the system transmits finish message to server for the logout in the client end computer. The training simulated exercise flow chart is as Fig. 1.

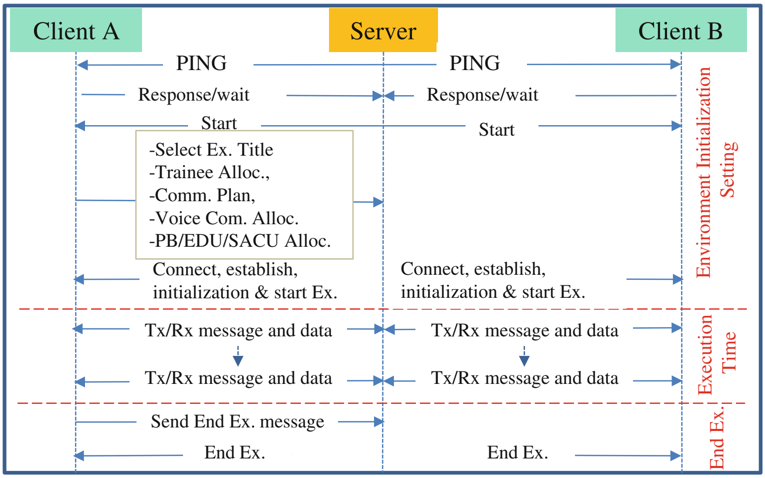


Fig. 1 TTS system flow chart

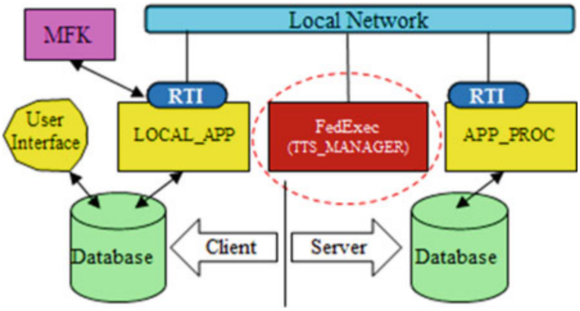
Upgrades Analysis

TTS and HLA architectures are analyzed for developing the concept of TTS migration that will follow the upgrades on the communication and application (APP).

HLA Architecture

There are advantages of HLA framework: first, it provides an interface for all computers through network that increases the function of interaction among users and data requesters (including system designer and analysts). Second, the modeling and simulation environment is under the open system framework allowing adopting reusable components to construct simulation that will maximize the utility rate and increase the flexibility. Third, HLA framework can integrate all simulators and simulation systems, which can be homogeneous or heterogeneous type that makes the expansion of the simulation range possible. Forth, the RTI mechanism reduces the quantity of data transmission in network. The data transmission for original TTS was in Protocol Data Unit (PDU) format, but PDU format is complicated and most of PDUs were defined for military purpose that would be a bottleneck for TTS with a very limited simulation performance. Also in the procedure of data transmission, due to PDU format, lots of unchanged data are still transmitted in packet by broadcasting that will waste bandwidth hugely. Fifth, the new system will be modulated for making it more convenient to have the following functions such as add in, modify and delete that can increase system flexibility. Sixth, DIS has time

Fig. 2 HLA architecture



managing problem that is prone to make time disorder in data receiving but with complete time managing functional services, HLA can illuminate the time disordering problem.

The HLA-based architecture is as Fig. 2. The comparison between TTS_MANAGER and FedExec is as Table 1.

Upgrade Communication and APP Module

Two major issues are essential with top priority to concern in the process of migration toward HLA architecture. The first concern is the modification of the communication that can be done by using RTI to take charge of all interchanged message. The advantage is to allow tactical module concentrating on its tactics design and mission planning without the distraction of taking care of data transfer event. The second concern would be the modification of APP on both of local computer and main computer for their Local_APP and APP_PROC, respectively. For current APP modules, they will be improved by having additional functions on addition, modification and deletion under HLA architecture that allows the system to edit all of the weapon systems, equipment and platforms base on the users' need.

Fig. 3 depicts the contrast of APP module in TTS and HLA. In TSS APP module, the send-out from client A to client B must by server, therefore some APPs would appear at both of main computer and local computer, such as Dynamics, MSG_OUT, EX_MNG and Detection, etc. In fact, the function of APPs are not the same at both main and local computer, e.g. at main computer, the tactical APP is simply responsible for receiving and send-out the message to the clients but it's not the case at local computer that would make mistakes that is ambiguous.

Table 1 TTS_MANAGER versus FedExec

TTS_Manager in TTS	FedExec in HLA
• Checking the status of all joined computers • Handling additional joining computer • Message transmission • Exercise recovery	• Managing the joining and leaving federate in federation • Helps Message transmission among federates in their federation • Restored function in federation management • RTI's time management

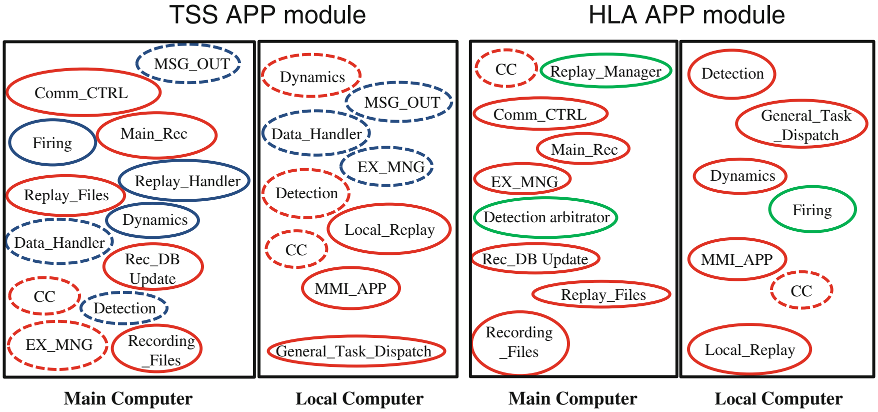


Fig. 3 APP in TTS and HLA (*dash line* indicates the same APP at both computer)

Conclusion

Most of military simulation systems which ranged from tactical simulations to technical simulators have been facing legacy problem and lack of interoperable capability. To bridge the current legacy TTS to HLA based concept under the considerations of life extension with interoperability, cost-effectiveness and risk, analyzing TTS by tracing codes thoroughly for restoring its original feature is substantial, from where the way of migration become clearer and the logic is developed with less difficult.

The serious issues encountered in migrating TTS to be under HLA architecture are repetitive sending message over the network and the ambiguous function emerged with APP module. This paper presents the solution with using RTI to handle all the in and out message instead of server, and simplify APP module with clear function in order to avoid the ambiguity.

Contribution of this paper is to provide a way solving not only the issue of life extension with the required more powerful additional functions but also the problem of making a legacy simulation system interoperable. With the migrated system being part of the military's synthetic environment, it become possible for having live, virtual and constructive (LVC) exercise tightly integrated that multiplier on training which would beyond uses' expectation. Moreover, from the Ministry of National Defense perspective, the LVC environment can also benefit the issues of military planning, acquisition and doctrine development.

References

1. Samuel ARB (2007) Legacy system: migration strategy. Technowave, Inc., pp 4–5
2. Smith J, Todd J, Kahl R (2009) Training range modernization: new technology on old infrastructure. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). Orlando, Floria: National Training and Simulation Association, pp 1–13
3. Fronckowiak D (2008) Extending the life of legacy software systems: a decision process model. In: Proceedings of Student-Faculty Research Day. New York City: Pace University, D1.2–D1.6
4. Bass L, Clements P, Kazman R (2003) Software architecture in practice, 2nd edn. Addison-Wesley Professional, MA
5. Bergey J, Smith D, Weiderman N (1999) DoD Legacy System Migration Guidelines. Carnegie Mellon Software Engineering Institute, Pittsburgh, pp 2–12
6. IEEE, Std-1278.1 (1995) IEEE standard for distributed interactive. Institute of Electrical and Electronics Engineers, Inc., New York
7. Wilson AL, Weatherly RM (1994) The aggregate level simulation protocol: an evolving system. In: Proceedings of the Winter Simulation Conference, IEEE. pp 781–787
8. IEEE, Std. -1516.1 (2000) IEEE standard for modeling and simulation (M&S), high level architecture (HLA)—federate interface specification. New York: IEEE, p.467
9. Huang J (2007) Design of an online decision-making support system for joint training simulation. JDMS, pp 43–53
10. Löfstrand B (2006) HLA Standard and Certification. Pitch Technologies, Rome

Mobile Agents for CPS in Intelligent Transportation Systems

Yingying Wang, Hehua Yan, Jiafu Wan and Keliang Zhou

Abstract Recently, cyber-physical systems (CPS) have emerged as a promising direction to enrich the interactions between physical and virtual worlds. Because of the large-scale features of CPS, mobile agents (MA) technology can promote the performance of CPS. In this article, we first introduce the concept and characteristics of CPS, MA, and intelligent transportation system (ITS). Then, we propose the structure of intelligent transportation CPS (ITCPS). On this basis, giving the case of mobile agents for ITCPS, we exploit a mobile agent by three levels (node level, task level, and combined task level) to reduce the information redundancy and communication overhead. Finally, we in brief outline the technical challenges for ITCPS.

Keywords Adaptive mobile agents · Cyber-physical systems · Intelligent transportation system

Introduction

In recent years, along with the rapid development and matures of Machine-to-Machine (M2M) technology, wireless sensor technology, cloud computing technology, and distributed real-time control technology, Cyber-Physical Systems (CPS) integrating computing, communications, and control in one fusion system has become one of the forefront interdisciplinary areas. CPS is closely related to

Y. Wang · H. Yan (✉) · J. Wan
School of Information Engineering, Guangdong Jidian Polytechinc, Guangzhou, China
e-mail: hehua_yan@126.com

K. Zhou
College of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou, China
e-mail: nyzkl@sina.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_82, © Springer Science+Business Media Dordrecht 2014

the development of human life and society, ranging from the small nanoscale biological robots to global energy coordination and management systems and involving human infrastructure construction. Currently, CPS applications include many domains, such as intelligent transportation and security, advanced automotive systems, industrial process control, distributed robotics, environmental control, and avionics, forming the basis to build smart city and smart planet of the future [1, 2].

CPS has emerged as a promising direction to enrich human-to-human, human-to-object, and object-to-object interactions in the physical world as well as in the virtual world, and typically involves multiple dimensions of sensing data, crosses multiple sensor networks and the Internet, and aims at constructing intelligence across these domains. So, dynamic participation and departure of a sensor network is possible. CPS imposes the same requirements for a WSN, but different levels of connectivity and coverage for different WSNs [3]. Furthermore, cross-domain communications may happen quite frequently in CPS applications [4]. In CPS applications, sensing data may be collected from static and mobile sensor nodes (such as those in vehicles, smart phones, etc.,) with both controllable and uncontrollable mobility. Apparently, we need to address the basic problems about the large data traffic, heterogeneous network communication, and real-time multi-task in CPS applications.

A mobile agent (MA) is a composition of computer software and data which is able to migrate from one machine to another autonomously and continue its execution on the destination machine, with the feature of autonomy, social ability, learning, and most importantly, mobility [5–8]. The feasibility of using mobile agents in CPS is reflected in the following aspects:

- CPS exists the great data traffic, while MA can move to the machine to calculate, reducing the network load.
- CPS is composed by a plurality of heterogeneous network, and MA is conducive to parallel processing to be executed on multiple heterogeneous network machines.
- CPS autonomously adapts to changes in the physical environment, and MA can react according to the state of the machine environment. For example, by the load condition of the current machine, MA decides whether to move to another machine.
- CPS meets the reliability requirements, and MA is tolerant to network faults and able to operate without an active connection between client and server.
- CPS includes a variety of applications, and the use of MA may reduce the deployment of applications and enhance portability, consequently maintenance is more flexible.

Intelligent transportation system (ITS) is the effective integration of advanced information technology, communications technology, sensor technology, control technology and computer technology, over a large distance area and in all directions. ITS is a real-time, accurate, efficient and integrated transport and management system [9–11]. The ITS research is in full swing in the United States, Europe, Japan, with amazing scale of development and alarming rate. Studies have shown that the use of CPS technology is feasible to achieve intelligent transportation system.

Intelligent Transportation CPS

Intelligent transportation CPS (ITCPS) needs to coordinate the different particle size of manned and unmanned equipment. To build reasonable CPS traffic architecture is the basic work of the future transportation system. The ITCPS is a typical CPS system, and has the typical characteristics of CPS, such as high flexibility, security, stability, reliability, efficiency, and seamless coupling. As Fig. 1 shows, the architecture of ITCPS can be divided into the following three layers [12–15]:

- Physical layer: In ITCPS, car, road equipment is no longer just a simple mechanical device, and will be embedded in a large number of sensors, calculators, and controllers. There are a lot of smart traffic devices distributed in the environment, such as smart cars, smart traffic lights, smart roads, smart bridge and so on, directly interacting with the physical environment. These transport equipment having sensing, computing and control functions constitute the physical layer of the ITCPS.
- Network layer: Network layer including satellite communications, base station communications, and other means of communications. All devices of the physical layer share the communication information through the network layer.

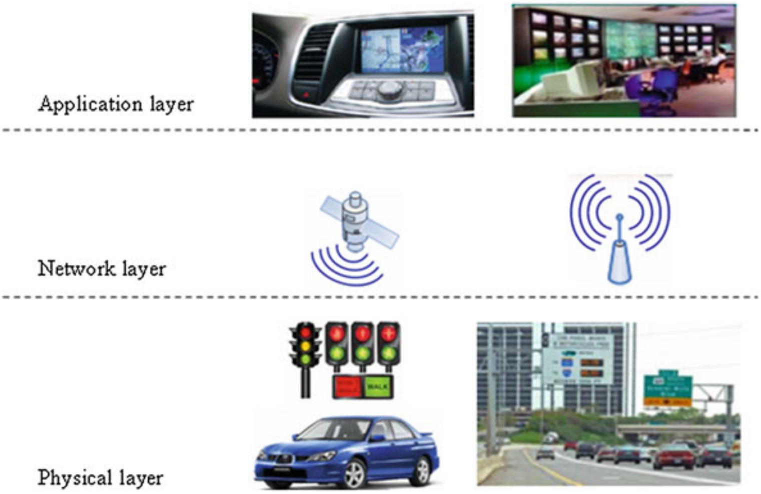


Fig. 1 A simple architecture of ITCPS

A smart car not only obtains surrounding environmental information through its own sensors, but also gets geographical environment information sensed by other vehicle sensors through the network layer, in order to facilitate the path forecast planning. Shielding the heterogeneity of the physical layer, the network layer seamlessly connects to share resource network-based for the application layer, and provide plug-and-play service for the user with the fully transparent way.

- Application layer: The application layer is the application of the user-oriented services, such as Smart car software, centralized management platform for traffic management department. Smart does not mean to exclude people, but in order to better serving people. The application layer is the part interacting with people directly; it needs to consider the people-oriented issues of design, implementation and confirmation, such as comfort, availability and correctness.

Case Study: Mobile Agents for ITCPS

In ITCPS, vehicles are furnished with on-board equipment (OBE) which includes computing, positioning, sensing, and wireless communication devises. As such, these vehicles are able to communicate with other equipped vehicles as well as with similarly instrumented roadside equipment (RSE). Thus, these components form vehicular ad-hoc networks (VANET), within which connected vehicles will be able to “talk” to each other. Belonging to the wireless sensor network, the link bandwidth of VANET is usually much lower than the wired network. However, ITCPS has frequent tasks, and the great data traffic is likely to exceed the capacity of wireless networks.

A mobile agent is a process that can transport its state from one environment to another, with its data intact, and be capable of performing appropriately in the new environment. MAs execute parallel processing on multiple heterogeneous network machines. MA’s actions are dependent on the state of the machine environment. Thus, MA may reduce the deployment of applications and enhance portability, and maintenance is more flexible. In conclusion, MA has the mobility, autonomy, collaboration and intelligence. Therefore, the use of the Mobile Agent technology in VANET can effectively reduce redundant information and communication overhead, thereby reducing network load.

As Fig. 2 shows, the RSE produces MA to perform specific tasks as a server. The MA should adjust its own behaviors depending on quality of service needs (e.g., data delivery latency) and the network characteristics to increase network lifetime while still meeting those quality of service needs. The information contained in a MA packet is shown in Table 1. A MA is dedicated to reduce the information redundancy and communication overhead in three levels so as to prolong VANET lifetime [16–23].

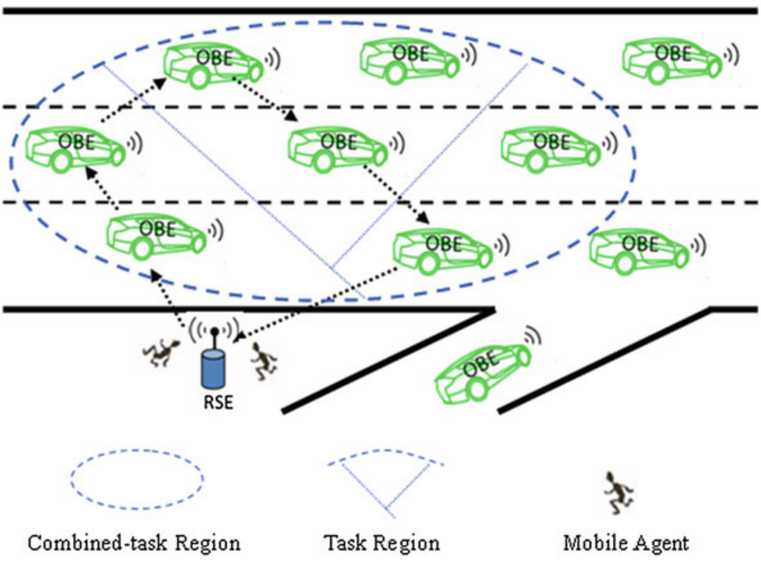


Fig. 2 A simple architecture of MA based VANET

Table 1 MA structure

Fixed attributes	Variable attributes	Payload
RSE_ID	NextOBESrc	Processing code
MA_SeqNum	NextHop	Accumulated data and result
FirstOBESrc	ToRSEFlag	
LastOBESrc	OBESrcList	
RoundIdx		
LastRoundFlag		

- Node Level: Application Redundancy Eliminating by local Processing.

In Fig. 2, the RSE can assign the processing code (behavior) of MA to the target OBE, based on the requirement of a specific application (e.g., measuring road surface conditions). The MA in OBE processes the local raw data as requested by the application, only to transmit, extract and send relevant information. So the capability enables a reduction in the amount of data transmission.

- Task Level: Spatial Redundancy Elimination by Data Aggregation.

The MA aggregates individual sensed data when it visits each target OBE in the task region. Such as navigation, the RSE assigns a MA one-by-one to access the target OBE and aggregate individual location data to achieve precise positioning.

- Combined Task Level: Communication Overhead Saved by Multiple Tasks' Data Concatenation.

The packet unification technique that unifies the several short data packets to one longer packet can reduce the communication overhead in combined-task level. Due to data concatenation, the duty cycle and communication overhead of intermediate OBE can be reduced so as to prolong VANET lifetime. In order to reduce the delay, multiple tasks should be executed simultaneously by replicated MAs in the combined-task region to decrease the execution time of all tasks. Since copying MAs will bring additional overhead, the MA number of the combined-task region should be carefully designed depending on application's requirements (e.g., intersection traffic safety, dangerous road traffic safety).

The MA performing tasks algorithm is as follows:

```
    If RSE has new tasks then
       RSE creates MA;
       Set RSE_ID, MA_SeqNum, FirstOBESrc, LastOBESrc, LastRoundFlag,
         NextOBESrc, NextHop, ToRSEFlag, OBESrcList;
  RSE dispatch MA to FirstOBESrc;
  FirstOBESrc sets RoundIdx = 1;
  MA processes Data and gets accumulated data result;
  Do While ToRSEFlag is false
    FirstOBESrc sets LastRoundFlag;
    MA sets NextOBESrc;
    If NextOBESrc = RSE then
       MA sets ToRSEFlag true;
    elseif NextOBESrc = LastOBESrc
         MA migrates toward LastOBESrc;
         MA processes Data and gets accumulated data result;
         MA sets ToRSEFlag true;
    elseif NextOBESrc = FirstOBESrc
         MA migrates toward FirstOBESrc;
         RoundIdx = RoundIdx + 1;
         MA processes Data and gets accumulated data result;
    else
         MA migrates toward NextOBESrc;
         MA processes Data and gets accumulated data result;
    endif
  End While
  MA migrates back to RSE;
endif
```

Discussion and Outlook

CPS research is still at a preliminary stage, so adaptive mobile agents for CPS also face many key technical issues, such as optimal allocation of resources, tasks real-time assurance, security control, standards development [21, 24–26]. Below, we in brief explain the technical challenges:

- Optimal allocation of resources: Multiple tasks are executed simultaneously by replicated MAs in the combined-task region to decree the execution time of all tasks. But copying MAs will bring additional overhead. Hence we should carefully design the MA number of the combined-task region and the RSE number according to application's requirements in order to avoid waste of resources.
- Real-time capabilities: We must ensure that the real-time performance must meet the specific application requirements. However, MAs execute the operation associated with migration, communication, control and safe, along with the additional running load. Therefore need to find a solution, such as better migration technology, better communication mechanisms, and multi-agency coordination mechanism.
- Security and privacy challenges: Since sensing data is no longer owned by local devices, security and privacy issues become more critical. MAs and machines (e.g., OBE, RSE) are vulnerable to a number of threats, including MA to machine threats, MA to MA threats and machine to MA threats. Security issue has been a hot and difficult research.
- Standards development. ITCPS applications depend on many technologies across multiple industries. Consequently, the required scope of standardization is significantly greater than that of any traditional standards development.

Conclusions

In recent years, CPS has become one of the forefront interdisciplinary areas. In this article, we first analyze the concept and characteristics of CPS, MA and ITS. Next, we present the structure of ITCPS. Then, through the case of mobile agents for ITCPS, we illustrate how a mobile agent is exploited in three levels (node level, task level, and combined task level) so as to reduce the information redundancy and communication overhead. Finally, we outlined the relevant research challenges. We hope to inspire more technological development and progress for CPS applications.

Acknowledgments The authors would like to thank the National Natural Science Foundation of China (No. 61262013), the Natural Science Foundation of Guangdong Province, China (No. S2011010001155), and the High-level Talent Project for Universities, Guangdong Province, China (No. 431, YueCai Jiao 2011) for their support in this research.

References

1. Wen JR, Wu MQ, Su JF (2012) Cyber–physical system. Acta Automatica Sinica 38(4):507–515

2. Wang ZJ, Xie LL (2011) Cyber–physical systems: a survey. Acta Automatica Sinica 37(10):1157–1166
3. Wang X, Xing G, Zhang Y, Lu C, Pless R, Gill C (2003) Integrated coverage and connectivity configuration in wireless sensor networks. In: Proceedings of the 1st international conference on Embedded networked sensor systems, Los Angeles, California, USA, pp 28–39
4. Han L, Potter S, Beckett G, Pringle G, Welch S, Koo SH, Tate A (2010) FireGrid: an e-infrastructure for next-generation emergency response support. J Parallel Distrib Comput 70(11):1128–1141
5. Ranjan S, Gupta A, Basu A, Meka A, Chaturvedi A (2000) Adaptive mobile agents: modeling and a case study. In: Proceedings of 2nd workshop on distributed computing IEEE Ind CFP, WDC
6. Zhao J, Liu B (2010) An overview of mobile agent (Mogent). Micro Process 31(1):1–5
7. Chen M, Gonzalez S, Leung V (2007) Applications and design issues for mobile agents in wireless sensor networks. IEEE Wirel Commun 14(6):20–26
8. Wang R, Zhou C (2001) The study of mogent (mobile agent): an overview. Appl Res Comput 18(6):9–11
9. Beresford AR, Bacon J (2006) Intelligent transportation systems. IEEE Pervasive Comput 5(4):63–67
10. Dimitrakopoulos G, Demestichas P (2010) Intelligent transportation systems. IEEE Veh Technol Mag 5(1):77–84
11. Weiland RJ, Purser LB (2000) Intelligent transportation systems, transportation in the new millennium
12. Miller J (2008) Vehicle–to–vehicle–to–infrastructure (V2V2I) intelligent transportation system architecture. IEEE intelligent vehicles Symposium, pp. 715–720
13. Wang F (2010) Parallel control and management for intelligent transportation systems: concepts, architectures, and applications. IEEE Intell Transp Syst 11(3):630–638
14. Chen M, Wan J, Li F (2012) Machine-to-machine communications: architectures, standards, and applications. KSII Trans Internet Inf Syst 6(2):480–497
15. Li LI, Liu YA, Tang BH (2007) SNMS: an intelligent transportation system network architecture based on WSN and P2P network. J China Univ Posts Telecommun 14(1):65–70
16. Suo H, Wan J, Huang L, Zou C (2012) Issues and challenges of wireless sensor networks localization in emerging applications. In: Proceedings of 2012 international conference on computer science and electronic engineering, Hangzhou, China, pp 447–451
17. Wan J, Li D (2010) Fuzzy feedback scheduling algorithm based on output jitter in resource–constrained embedded systems. In Proceedings of international conference on challenges in environmental science and computer engineering, Wuhan, China, pp 457–460
18. Chen M, Gonzalez S, Zhang Q, Leung VC (2010) Code-centric RFID system based on software agent intelligence. IEEE Intell Syst 25(2):12–19
19. Tseng YC, Kuo SP, Lee HW, Huang CF (2004) Location tracking in a wireless sensor network by mobile agents and its data fusion strategies. Comput J 47(4):448–460
20. Zou C, Wan J, Chen M, Li D (2012) Simulation modeling of cyber-physical systems exemplified by unmanned vehicles with WSNs navigation. In Proceedings of the 7th international conference on embedded and multimedia computing technology and service, Gwangju, Korea, pp 269–275
21. Chen M, Leung V, Mao S, Kwon T (2009) Receiver-oriented load-balancing and reliable routing in wireless sensor networks. Wirel Commun Mob Comput 9(3):405–416
22. Wan J, Yan H, Suo H, Li F (2011) Advances in cyber-physical systems research. KSII Trans Internet Inf Syst 5(11):1891–1908
23. Chen M, Gonzalez S, Zhang Y, Leung V (2009) Multi-agent itinerary planning for wireless sensor networks. Qual Serv Heterogen Netw 22:584–597
24. Wu FJ, Kao YF, Tseng YC (2011) From wireless sensor networks towards cyber physical systems. Pervasive Mob Comput 7(4):397–413

25. Liu J, Wang Q, Wan J, Xiong J (2012) Towards real-time indoor localization in wireless sensor networks. In: Proceedings of 12th IEEE international conference on computer and information technology, Chengdu, China, pp 877–884
26. Yan H, Wan J, Suo H (2011) Adaptive resource management for cyber–physical systems. In: Proceedings of international conference on mechatronics and applied mechanics, HongKong, pp 747–751

Improving Spectator Sports Safety by Cyber-Physical Systems: Challenges and Solutions

Hehua Yan, Zhuohua Liu, Jiafu Wan and Keliang Zhou

Abstract With the support of wireless networking, cloud computing, and advanced control technology, cyber-physical systems (CPS) applications are gradually becoming a reality. However, different applications (e.g., spectator sports safety) are confronted with diverse challenges and issues. In this article, we innovate an autonomous dynamic spatial panoramic video surveillance system that is a typical CPS to provide effective technologies and systems to protect spectators in safety and security. We first introduce the system model and components, and then propose the methodologies for quality of service (QoS) improvement from the following aspects: distributed real-time panoramic video stitching, autonomous unmanned aerial vehicles, autonomous unmanned ground vehicles, and wireless networking. The proposed methodologies can help the system design and improve the QoS.

Keywords Cyber-physical systems · Wireless networking · Unmanned aerial vehicles · Unmanned ground vehicles · Challenges

H. Yan · Z. Liu (✉) · J. Wan
School of Electrical Engineering, Guangdong Jidian Polytechinc, Guangzhou, China
e-mail: cnhope@tom.com

H. Yan
e-mail: hehua_Yan@126.com

J. Wan
e-mail: jiafu_wan@ieee.org

K. Zhou
College of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou, China
e-mail: nyzkl@sina.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_83, © Springer Science+Business Media Dordrecht 2014

Introduction

The importance of spectator sports safety and security is often overlooked in research and education. The large sports events may face various threats. For example, recall the bombing threat in the 2004 Summer Olympic Games [1] and the Munich massacre during the 1972 Olympics [2]. Other threats can come from accidents, such as fire and gas explosions. Confronting these challenges, sports spectators are hard to be secured with the capabilities of traditional sport security management mainly relying on human labor. In our view, an interdisciplinary effort is required. This paper takes the multidisciplinary research initiative for a cyber-physical system (CPS) of dynamic spatial video surveillance to effectively protect spectators.

In recent years, the research on CPS has made much progress in prototype testbed, resource allocation, control methods, system model, etc., [3–11]. Depending on the previous research results, we innovate an autonomous dynamic spatial panoramic video surveillance CPS to provide effective technologies and systems to protect spectators in safety and security. This system consists of two primary instruments: (1) an autonomous spatial video surveillance platform consisting of heterogeneous unmanned physical objects (unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs)) networked by wireless communication, and (2) a distributed multi-tier hierarchical real-time panoramic video stitching solution overlaying across the spatial surveillance platform. As shown in Fig. 1, the autonomous platform is driven by surveillance missions and collaborates with the distributed video stitching to accomplish the surveillance missions. The system communication and the video are secured with confidentiality and integrity.

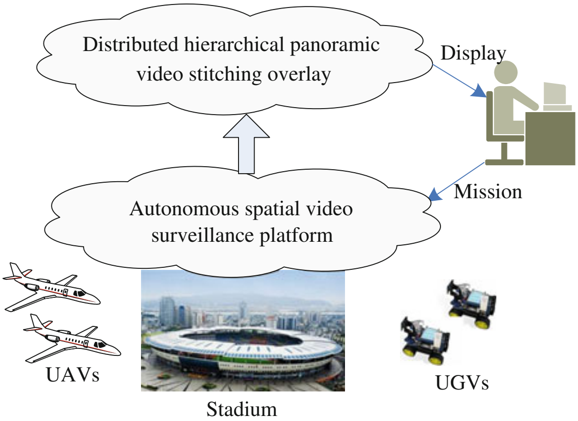

Distributed hierarchical panoramic video stitching overlay
Display
Mission
Autonomous spatial video surveillance platform
UAVs
Stadium
UGVs

Fig. 1 Conceptual system

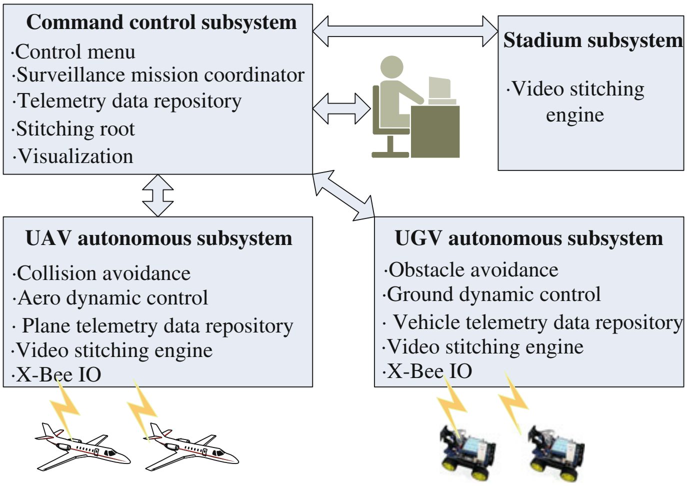


Fig. 2 System model

System Model and Components

A loosely conceptualized system model of the proposed CPS is shown in Fig. 2. The model consists of three major subsystems: command control subsystem, UAV autonomous subsystem and UGV autonomous subsystem, and one minor subsystem: stadium subsystem. Actually, the whole system is an autonomous one with three major subsystems being semi-autonomous in that:

- The command control subsystem has the top-level mission-based intelligence, but not to every detail, to coordinate the UAVs and UGVs.
- The UAV and UGV subsystems have their own low-level intelligence on autonomous coordination to accomplish their assigned missions, but they have to follow the control and missions from the command control subsystem.
- They all can revise their control during a mission based on the closed-loop feedback of physical surroundings.
- Command Control Subsystem: This subsystem provides the interface to the commander or system administrator and is the "brain" of the entire system in coordinating UAVs and UGVs subsystems to fulfill a surveillance mission. The commander initiates a surveillance mission through the *control menu*, for example, the concerned surveillance area/object is specified. The mission can be interpreted into a set of parameters such as the target Geo location, concerned objects, expected video resolution, etc. Then, the mission is delivered to the *surveillance mission coordinator* that computes the preliminary paths for the UAVs and UGVs to move based on their current physical information such as battery (power) and current Geo location from the *telemetry data repository*. The computed path data are delivered via wireless links to the UAV and UGV

autonomous subsystems for further process and control. The *telemetry data repository* polls the current physical system information of UAVs and UGVs from their subsystems. Thus, it also supports the visualization of the UAVs and UGVs on the map with Google Maps KML [12] to the commander. The *stitching root* stitches the panoramic videos from the UAV and UTV subsystems into the final panoramic real-time video and visualizes it to the commander for surveillance.

- UAV Autonomous Subsystem: This is a distributed system working across all UAVs. When the UAV subsystem receives UAV control messages with path data from the *command control subsystem*, the *collision avoidance* inspects any potential collision and makes necessary updates of the data before passing them to the *aero dynamic control* module. It keeps running with the continual updates of the latest UAV physical system information in flights. The *aero dynamic control* computes the aero dynamic parameters for UAVs and then notify all UAVs through the *X-Bee IO* that is a tri-band wireless interface on 2.4/5/60 GHz. Meanwhile, all UAVs continually update their flight physical information such as location, speed and battery to the *plane telemetry data repository* in the subsystem via the *X-Bee IO* so that the system is up-to-date of their status. In addition, the surveillance video captured by the UAVs will be hierarchically stitched by the *video stitching engine* in a distributed mode. The eventual stitched panoramic video of the UAV subsystem is pushed to the *command control subsystem* for a final stitching.
- UGV Autonomous Subsystem: This subsystem works similarly to the UAV subsystem except that the UGVs have to avoid obstacles in their paths and have different control and mechanical systems for their motions on the ground. *X-Bug IO* is also a tri-band wireless interface that exchanges the messages containing UGV physical system information or surveillance video.
- Stadium Subsystem (Immobile Camera): This subsystem refers to the traditional video surveillance system deployed in the venue or its surroundings. This subsystem does not require any mechanical control as the UAV and UGV systems do. It accepts inputs such as video resolutions from the *command control subsystem* and runs the distributed *video stitching engine* to generate the panoramic real-time video of its coverage to command control office. It might also provide wireless communication replay to the UAVs and/or UGVs when possible.

Distributed Real-Time Panoramic Video Stitching

The stitching solution currently only runs in standalone mode on a single computer that is short of stitching videos from a large number of surveillance cameras in events like sporting games. To extend the standalone stitching to a distributed real-time video stitching solution, we address a couple of challenges.

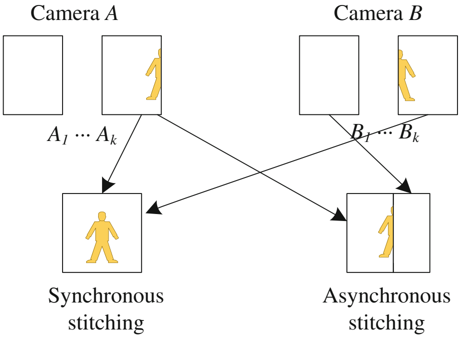


Fig. 3 Synchronization on stitching

Synchronization: Synchronization between video frames to be stitched is critical to the stitching result. Stitching asynchronous video frames may lead to the displacement of objects, which is illustrated in Fig. 3 where two video streams come from neighboring surveillance cameras A and B. Suppose that an object is moving from A to B at the time of frame k. If both camera videos arrive synchronously at the stitching point, frame A_k is stitched with frame B_k and the object will appear naturally in the stitched frame. Otherwise, in an asynchronous case where frame A_k is stitched with frame B_1 that arrives late, partial of the moving object disappears in the stitched result as in the right bottom on the figure. In the distributed video stitching, the synchronization of video frames from different cameras is significantly challenged even if the cameras are tightly synchronized because of

- The different delays that video streams may undergo when transmitted over the network.
- The different processing time at stitching nodes at different tiers.

We consider a delay-tolerant stitch-or-store solution to address the asynchronous issue. At stitching nodes, the video frames are verified of the synchronization in a two-step procedure before stitched. The timestamps of the frames are first inspected. When the system time of the cameras is tightly synchronized, identical timestamps hint synchronization. Otherwise, mobile objects in the overlapping coverage among frames are examined. Synchronous frames should have the same mobile object at the same place in their overlapped coverages at the same time. If the frames are synchronous, then they are stitched. If only some frames are synchronous, these frames are first stitched and stored with other asynchronous frames locally for a short period to wait for other frames synchronous to this stitching set.

Scalability: Another challenge is the scalability of the stitching due to the large number of cameras in the spatial surveillance platform. We could run the standalone video stitching solution on a central powerful server such as a high performance cluster to achieve large scale stitching. But, we have to transmit all videos across the network to the central computing server. As a result, the huge amount of data will definitely overwhelm the network and the resulted traffic congestion will lead to severely asynchronous videos. Meanwhile, the overlapping information in video frames wastes the network bandwidth. Instead of transmitting all videos to a central server for processing, we will take a strategy similar to clouding computing that pushes the computation to the fronts. This strategy leads to a distributed multi-tier hierarchical stitching architecture. The camera nodes form into clusters and each cluster has an elected cluster lead that stitches the local panoramic video from the cameras in the same cluster. Then, the cluster leads at the same tier form into higher-tier clusters and do the election and stitching again. This process repeats till reaching the stitching root, where the final panoramic video is generated. The size of a cluster depends on the computation resources of the embedded devices available at the distributed stitching nodes.

Autonomous UAVs

At this time, we already know how to fly UAVs through a set of predetermined or uploaded waypoints. The proposed CPS requires UAVs fly autonomously and collaboratively to accomplish tasks in taking and stitching videos. The flight paths are often determined by the surveillance tasks. In this paper, we may design task-driven algorithms that allow genuine autonomous flight of a large number of cooperative UAVs.

One challenge is that the *Multiplex Easystar* UAV is not appropriate for surveillance purpose due to its high speed. For surveillance, we often need to browse slowly or even stop to focus on suspicious events. For such constraints, a multicopter (quad) is more appropriate. The DIY Drone community developed an open source auto-pilot for a quadcopter: the *ArduCopter*. The autopilot uses the *Arduino* microcontroller. Given Auburn familiarity with this platform and the versatility of the *ATTRACT-ROS* software architecture, it is easy to adapt the collision avoidance algorithms designed initially for the *Multiplex Easystar*. The maneuverability of the *ArduCopter* greatly simplifies the problem of collision avoidance in that: (1) the speed can vary, (2) it can stop, and (3) it can take sharp turns. With this architecture, *ArduCopters* can be assigned surveillance missions with physical flexibility for collisions avoidance. Another challenge is in the UAV capability. The current *ArduCopter* can fly only about 10 min. We need to increase the flight time and the payload capacity to carry a camera.

Autonomous UGVs

The UGVs in the proposed system face various challenges in the populated venues. The largest challenge comes from the mobility of human beings in sports games. A UGV has to detect and avoid any obstacles in its path. To this day, it is not a problem anymore for a robot to avoid static obstacles in indoor or outdoor environments with many-year research. However, the situation becomes extremely complicated at a sport stadium with a big crowd: people often walk around. It is challenging for a UGV to act in such an environment where obstacles are mobile. An appropriate moving path of a UGV must be dynamically computed based on the mobility of obstacles, the stadium environment and the sensibility of the UGV. New algorithms are required to provide high accuracy, effectiveness and efficiency in UGVs' autonomous driving. We may address this problem by exploiting the vector field based (VFB) guiding model to UGVs with the global knowledge from the UAVs in the sky in the spatial platform. VFB extends the potential based guiding (PBG) model that is widely adopted in obstacle avoidances for robots. In PBG, a robot divides scanned area into *Z* scanned lines that point to different directions. The main idea of the PBG is to repulse (or attract) the robot from or to the obstacles. However, this model relies only on the sensing information obtained by robot itself, which easily leads to a mistaken decision due to the lack of global knowledge. VFB extends the PBG model in such a way that a direction can be defined at every point of the potential surface in PBG model. Therefore, the global information discovered by the UAV system in our spatial platform can be modeled into the UGV's decision-making process. The globe information such as human density, traffic pattern, and moving speed of a swarm can be useful in computing the potential best path data. The data will then be fed into the VFB model so that UGVs can make a motion decision based on both the local and global information in their movement.

Another challenge is in the UGV mechanical capability. In the application of spectator sports safety and security, UGVs confront the complexity of terrestrial surroundings of a venue that is normally in an urban area where there are various physical obstacles such as curbs and steps. In addition, the battery of the Lunabot does not last long. In the surveillance, this is a potential issue.

Wireless Networking

Networking protocols are necessary to drive the wireless communication. We expect to use the IEEE 802.11 as the MAC/PHY protocols. However, special routing protocols are required to: (1) forward the coordination messages of the spatial platform, especially between the command control and the UAV and UGV systems and (2) deliver the surveillance videos for stitching or visualization. The routing faces the challenges of (1) tri-band wireless communication, (2) directional antennas in 60 GHz for power gain, and (3) special hierarchical topologies in distributed video stitching. Now, we have extensive expertise on multi-hop wireless network routing and design and develop a geographical routing protocol to support the communication of the coordination across the system, since the system components have GPS equipped [13, 14]. In addition, a tree-like routing protocol should be developed to support the distributed video stitching and transmission over the hierarchical multi-tier topology. Since the wireless networking is based on tri-band wireless communication, the routing protocols should consider the impact of multiple bands.

Conclusions

Recently, CPS has emerged as a promising direction to enrich the interactions between physical and virtual worlds. This paper integrates the traditional disciplines into a complex system with the clear goal of autonomous dynamic spatial panoramic video surveillance to address the challenges. In order to protect spectators in safety and security, we introduce the system model and components of CPS, and propose the methodologies for QoS improvement from the following aspects: distributed real-time panoramic video stitching, autonomous unmanned aerial vehicles, autonomous unmanned ground vehicles, and wireless networking. We have proposed several research challenges and solutions to encourage more insight into this meaningful application.

Acknowledgments The authors would like to thank the National Natural Science Foundation of China (No. 61262013), the Natural Science Foundation of Guangdong Province, China (No. S2011010001155), and the High-level Talent Project for Universities, Guangdong Province, China (No. 431, YueCaiJiao 2011) for their support in this research.

References

1. Taylor T, Toohey K (2006) Perceptions of terrorism threats at the 2004 Olympic Games: implications for sport events. J Sport Tourism 12(2):99–114
2. Wikipedia http://en.wikipedia.org/wiki/munichmassacre
3. Wan J, Chen M, Xia F, Li D, Zhou K (2013) From machine-to-machine communications towards cyber-physical systems. Comp Sci Inf Syst. doi:10.2298/CSIS120326018W
4. Chen M, Wan J, Li F (2012) Machine-to-machine communications: architectures, standards, and applications. KSII Trans Internet Inf Syst 6(2):480–497
5. Wan J, Yan H, Li D, Zhou K, Zeng L (2013) Cyber-physical systems for optimal energy management scheme of autonomous electric vehicle. Comput J. doi:10.1093/comjnl/bxt043
6. Wan J, Yan H, Suo H, Li F (2011) Advances in cyber-physical systems research. KSII Trans Internet Inf Syst 5(11):1891–1908
7. Suo H, Wan J, Huang L, Zou C (2012) Issues and challenges of wireless sensor networks localization in emerging applications. In: Proceedings of 2012 international conference on computer science and electronic engineering, Hangzhou, China, pp 447–451, March 2012
8. Zou C, Wan J, Chen M, Li D (2012) Simulation modeling of cyber-physical systems exemplified by unmanned vehicles with WSNs navigation. In: Proceedings of the 7th international conference on embedded and multimedia computing technology and service, Gwangju, Korea, pp 269–275, Sept 2012
9. Chen M, Gonzalez S, Zhang Q, Leung V (2010) Code-centric RFID system based on software agent intelligence. IEEE Intell Syst 25(2):12–19
10. Suo H, Wan J, Li D, Zou C (2012) Energy management framework designed for autonomous electric vehicle with sensor networks navigation. In: Proceedings of the 12th IEEE international conference on computer and information technology, Chengdu, China, pp 914–920, Oct 2012
11. Shi J, Wan J, Yan H, Suo H (2011) A survey of cyber-physical systems. In: Proceedings of the international conference on wireless communications and signal processing, Nanjing, China, pp 1–6, Nov 2011
12. Google Inc. http://code.google.com/apis/kml/documentation/whatiskml.html

13. Wang X, Vasilakos A, Chen M, Liu Y (2012) A survey of green mobile networks: opportunities and challenges. Mob Netw Appl 17(1):4–20
14. Chen M, Leung V, Mao S, Li M (2008) Cross-layer and path priority scheduling based real-time video communications over wireless sensor networks. In: Proceedings of IEEE vehicular technology conference, pp 2873–2877

MapReduce Application Profiler

Tzu-Chi Huang, Kuo-Chih Chu, Chui-Ming Chiu and Ce-Kuen Shieh

Abstract MapReduce is a programming model popularized by Google to process large data in clusters and has become a key technology on cloud computing nowadays. Due to the feature of simplicity, MapReduce attracts many application developers to develop related applications. However, MapReduce currently has few solutions to help application developers with the tasks of profiling their applications. In this paper, MapReduce Application Profiler (MRAP) is proposed to facilitate profiling MapReduce applications in clusters.

Keywords MapReduce · Cloud computing · MRAP · Profiler

Introduction

MapReduce [1] is a programming model popularized by Google to process large data in clusters. MapReduce has become a key technology on cloud computing to make an application easily utilize resources in computers. Today, MapReduce has been widely used by many public services such as Google Search Engine, Yahoo

T.-C. Huang (✉) · K.-C. Chu
Department of Electronic Engineering, Lunghwa University of Science and Technology,
Guihan, Taiwan, Republic of China
e-mail: tzuchi@mail.lhu.edu.tw

K.-C. Chu
e-mail: kcchu@mail.lhu.edu.tw

C.-M. Chiu · C.-K. Shieh
Institute of Computer and Communication Engineering, Department of Electrical
Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China
e-mail: cmchiu@ee.ncku.edu.tw

C.-K. Shieh
e-mail: shieh@ee.ncku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_84, © Springer Science+Business Media Dordrecht 2014

Search Engine, Amazon EC2, and Facebook to deal with a huge amount of data. In the future, MapReduce definitely will have more and more related applications.

Due to the feature of simplicity, MapReduce allows application developers to easily develop applications in clusters even though they do not have many experiences on developing distributed and parallelized applications. MapReduce relies on the runtime system to automatically distribute an application over computers in clusters, provide the application with input data at runtime, and collect outputs of the application from computers in clusters. All MapReduce needs is a Map function and a Reduce function developed by application developers to implement the functionality of their applications.

Since simplicity is the main contribution of MapReduce to the development of applications on cloud computing, MapReduce should offer application developers a way to easily profile their applications as well. MapReduce should care about that application developers may have very few experiences on developing distributed and paralleled applications in clusters. Furthermore, MapReduce should care about that application developers may neither know how the runtime system serves their applications nor understand how to improve their applications in order to get a better performance. However, MapReduce currently has few solutions specifically designed to help application developers with the tasks of profiling their applications.

In this paper, MapReduce Application Profiler (MRAP) is proposed to facilitate profiling MapReduce applications in clusters. MRAP can collect runtime information about the application execution progress at runtime and save runtime information in files compatible to Microsoft Excel. Accordingly, MRAP allows application developers to quickly profile their applications, e.g., doing statistics or drawing charts about the application execution progress. MRAP provides application developers with runtime information about the application execution progress, so they can easily improve their applications to get a better performance. In this paper, MRAP demonstrates the performance metrics of Word Count in a generic runtime system as the example.

This paper is organized as follows. Section Background reviews MapReduce as the background. Section MapReduce Application Profiler (MRAP) addresses MRAP. Section Implementation introduces the MRAP implementation. Section Demonstration has a demonstration of MRAP with Word Count. Section Conclusions concludes this paper.

Background

MapReduce is a programming model relying on its runtime system to distribute an application over computers in clusters, provide the application with input data at runtime, and collect outputs of the application from the computers in clusters. For serving an application, MapReduce needs application developers to develop a Map function (a.k.a. Mapper) and a Reduce function (a.k.a. Reducer) to implement the functionality of their applications. Accordingly, MapReduce divides the application execution progress roughly into the Map phase and the Reduce phase.

At the Map phase, the runtime system reads input files from local disks of a master node, i.e., a computer responsible for controlling the entire application execution progress, or certain distributed file systems such as the Google File System (GFS) [2] or the Hadoop Distributed File System (HDFS) [3] specifically designed for cloud computing. The runtime system distributes Mappers over computers in clusters according a certain node selection algorithm, e.g., finding the idle ones to run the Mappers, and provides the Mappers with contents of input files. After Mappers process contents of input files and generate intermediate data in a format of key and value pairs, the runtime system enters the Reduce phase. At the Reduce phase, the runtime system chooses suitable computers to run Reducers and forwards the intermediate data from Mappers to Reducers. Finally, the runtime system collects outputs from Reducers as the result of the application. Because the runtime system manages Mappers and Reducers of applications at runtime on behalf of application developers, application developers can easily develop applications on cloud computing by focusing on the development of Mappers and Reducers.

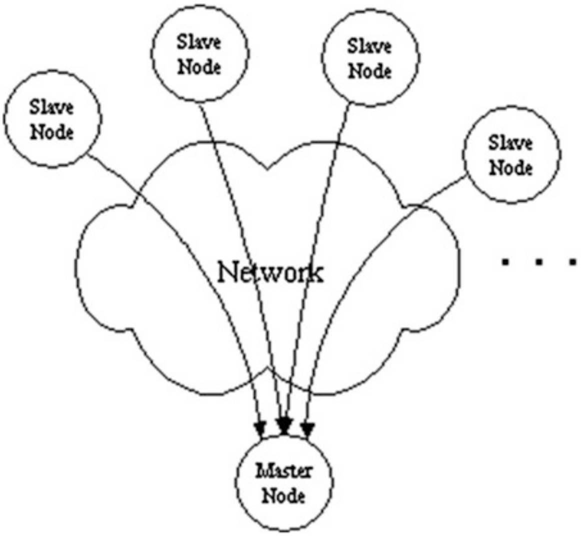
Word Count [1] is a canonical application often used to explain MapReduce on cloud computing. Word Count uses a Mapper to process input data offered by the runtime system and generate a key and value pair as intermediate data for each word found in input data, e.g., “word 1”. When processing intermediate data, Word Count uses a Reducer to merge the key and value pairs generated by a Mapper. Word Count may count words by summing up their values in a Reducer and outputs partial results to the runtime system. Finally, Word Count waits the runtime system to collect all partial results from Reducers to get the final result of the application.

MapReduce Application Profiler (MRAP)

Overview

MRAP has an overview in Fig. 1 where a cluster is composed of a master node and several slave nodes. MRAP uses a master node to not only do the works as the master node in other MapReduce systems but also accept runtime information reported by slave nodes. MRAP uses slave nodes to not only execute tasks dispatched by the master node, e.g., running a Mapper or Reducer, but also collect runtime information of the runtime system and the Operating System (OS) at local nodes. In this paper, MRAP refers to information about the application execution progress in a slave node as runtime information, e.g., CPU utilization, number of Mapper instance, number of Reducer instance, intermediate data size, and quantity

Fig. 1 MRAP overview



of network bandwidth consumption. At the end of the application, MRAP allows application developers to use Microsoft Excel to examine runtime information in files saved by the master node.

System Components

MRAP has components in Fig. 2. MRAP has the Runtime System to MRAP (RS-MRAP) interface between the runtime system and itself, and the MRAP to Operating System (MRAP-OS) interface between the Operating System (OS) and itself. MRAP exports the RS-MRAP interface to the runtime system and communicates with the runtime system in order to get the MapReduce-related information in the runtime system, e.g., number of Mapper instance, number of Reducer instance, and intermediate data size, because the runtime system is responsible for managing the execution of Mappers and Reducers and intermediate data at runtime. MRAP utilizes the MRAP-OS interface (usually embodied with OS calls) to communicate with the OS in order to get the OS-related information in the OS, e.g., CPU utilization and quantity of network bandwidth consumption. Through the MRAP-OS interface, MRAP also can report runtime information to the master node via the network service, collect runtime information from slave nodes via the network service, and save runtime information to files via the file system.

MRAP uses Information Reporter to periodically report runtime information of the local node to the master node via the network service in the OS. MRAP uses Information Recorder to collect runtime information from salve nodes via the

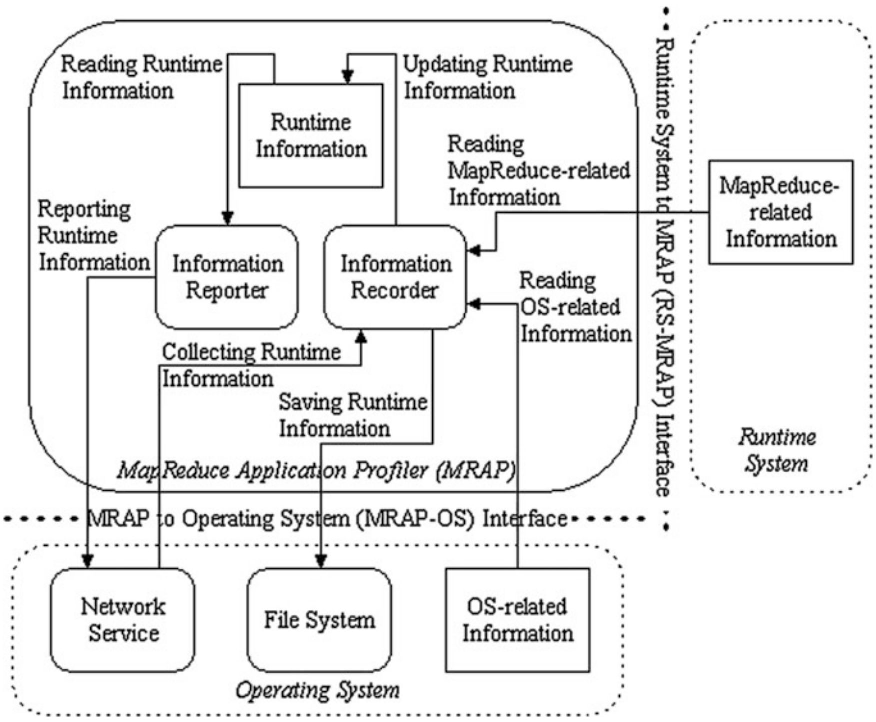

Reading Runtime Information
Runtime Information
Updating Runtime Information
Reading MapReduce-related Information
MapReduce-related Information
Reporting Runtime Information
Information Reporter
Information Recorder
Reading OS-related Information
Collecting Runtime Information
Saving Runtime Information
MapReduce Application Profiler (MRAP)
Runtime System to MRAP (RS-MRAP) Interface
Runtime System
MRAP to Operating System (MRAP-OS) Interface
Network Service
File System
OS-related Information
Operating System

Fig. 2 MRAP components

network service in the OS, save runtime information to files via the file system, read OS-related information in the OS, and read MapReduce-related information in the runtime system. Besides, MRAP uses Information Recorder to update runtime information that is the collection of the OS-related information and the MapReduce-related information. While MRAP usually caches runtime information in the memory in slave nodes to facilitate reporting runtime information to the master node periodically, it saves runtime information to files in the master node directly once receiving runtime information reported by slave nodes from networks.

Implementation

We use the C language to implement MRAP as a library that can be linked to the runtime system. At the initial time of MRAP, we create a thread to periodically query not only the runtime system about number of Mapper instance, number of Reducer instance, and intermediate data size, but also the OS about CPU utilization and quantity of network bandwidth consumption. Next, we arrange the

MapReduce-related information and the OS-related information in a Type-Length-Value (TLV) format [4] and use the UDP socket library to send them to a configurable IP address as the master node. Currently, we set the interval to 1 s for MRAP in a slave node to periodically report runtime information to MRAR in the master node.

For a proof of concept, we implement a runtime system capable of reading input files from disks in the master node, finding the idle nodes to run Mappers and Reducers of an application, automatically feeding Mappers with input files downloaded from the master node, forwarding intermediate data files from Mappers to Reducers on demand, and collecting outputs produced by Reducers from all computers in clusters as the result of the application. Next, we link MRAP to the runtime system and make it periodically query the runtime system about the MapReduce-related information through subroutine calls. Besides, we make MRAP query the OS about the OS-related information through the Windows Management Instrumentation (WMI) APIs [5], because Windows Server 2003 is our current platform.

Demonstration

We construct a cluster with 9 computers connected to each other via Gigabit Ethernet. We setup an AMD Athlon II X4 620 CPU (2.6 GHz) and 4 GB DDR2 RAM in each of the computers. We install a runtime system and MRAP on Windows Server 2003 at each computer. We select one of the 9 computers as the master node while selecting the other computers as slave nodes. We prepare 210 input files having 64 MB in the master node in order to test Word Count, the canonical application widely used to observe MapReduce runtime systems. Accordingly, we give Word Count a total workload of 13,440 MB (i.e., 210 × 64 MB). We use MRAP to profile Word Count and Microsoft Excel to draw charts about various metrics in runtime information as the demonstration in this paper.

We show CPU utilization of all slave nodes in Fig. 3. We observe that Word Count roughly can achieve 60 % CPU utilization in all slave nodes in the runtime

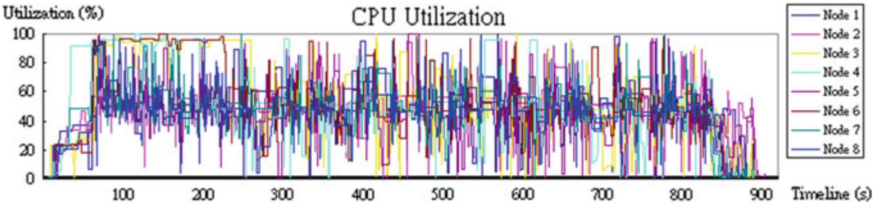


Fig. 3 CPU utilization

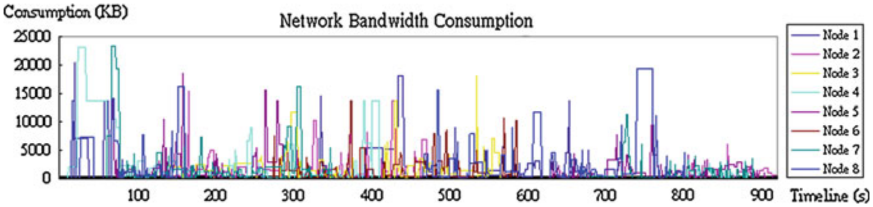


Fig. 4 Network bandwidth consumption

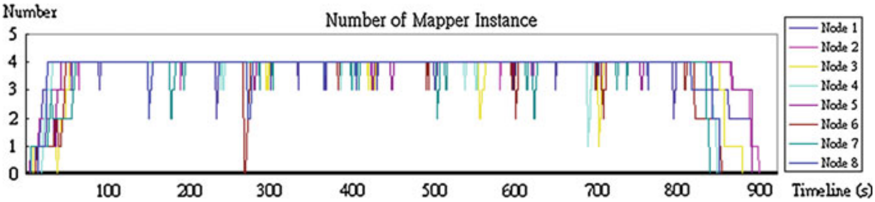


Fig. 5 Number of mapper instance

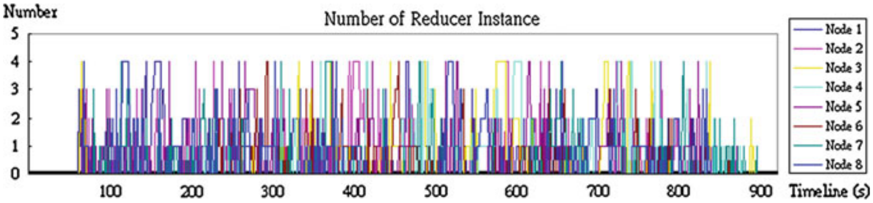


Fig. 6 Number of reducer instance

system. In other words, we use CPU utilization in runtime information offered by MRAP to know that Word Count is not an application costing much CPU time.

In Fig. 4, we observe network bandwidth consumed by each slave in the cluster. We note that network bandwidth consumption in Fig. 4 corresponds to CPU utilization in Fig. 3 because Mappers have to receive input files downloaded by the runtime system from the master node and because Reducers have to receive intermediate data forwarded by the runtime system from Mappers to Reducers before any computation can be made.

With the help of MRAP, we can clearly observe the number of Mapper instances in the runtime system. According to Fig. 5, we observe that Word Count almost has 4 Mapper instances activated in each slave node during the entire application execution progress, which implies that the runtime system can keep the application busy at processing input data with Mappers.

In Fig. 6, we observe the number of Reducer instances in the runtime system. We note that the average number of Reducer instances in each slave node is 2,

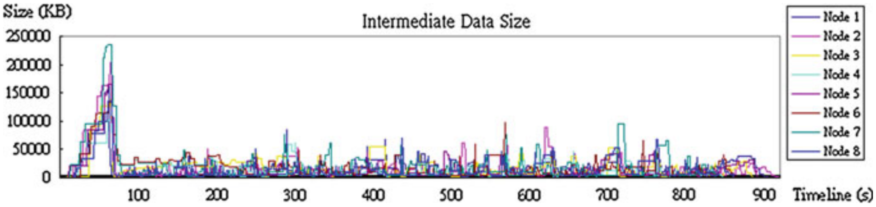


Fig. 7 Intermediate data size

which implies that the runtime system does not use a Reducer to process much intermediate data. Accordingly, we can deduce that Word Count does not generate much intermediate data in Mappers.

Finally, we observe intermediate date size of Word Count in Fig. 7. We confirm that Word Count does not generate much intermediate data, which corresponds to the observation in Fig. 6. Moreover, we note that Word Count generate much intermediate data before 60 s because the runtime system is busy at uploading input files to slave nodes and running Mappers to process input data without invoking Reducers to process intermediate data. We confirm the deduction because Fig. 6 shows that there is no Reducer instance before 60 s in the runtime system.

Conclusions

In this paper, we propose MapReduce Application Profiler (MRAP) to facilitate the profile of an application on cloud computing. We design MRAP to periodically query the runtime system and the Operating System (OS) about runtime information, the collection of the MapReduce-related information and the OS-related information. We design runtime information in MRAP to include information about CPU utilization, network bandwidth consumption, number of Mapper instance, number of Reducer instance, and intermediate data size in each slave node at runtime. We make MRAP collect and output runtime information to files compatible to Microsoft Excel, so application developers can easily profile their applications by observing the application execution progress with statistics or charts. We believe that application developers with the help of MRAP can easily improve their applications to get a better performance.

Acknowledgments We gratefully acknowledge the National Science Council of Taiwan for their support of this project under grant numbers NSC 100-2628-E-262-001-MY2 and 101-2219-E-006-005. We further offer our special thanks to the reviewers for their valuable comments and suggestions, which materially improved the quality of this paper.

References

1. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. In: Communications of the ACM, vol. 51. ACM, New York, USA
2. Ghemawat S, Gobioff H, Leung S-T (2003) The Google file system. In: ACM SIGOPS operating systems review, vol. 37. ACM, New York, USA, pp 29–43
3. Shvachko K, Kuang H, Radia S, Chansler R (2010) The Hadoop distributed file system. In: 2010 IEEE 26th Symposium on mass storage systems and technologies (MSST), IEEE Press, New York, USA, pp 1–10
4. Liu H, Zhang D (2012) A TLV-structured data naming scheme for content-oriented networking. In: 2012 IEEE international conference on communications (ICC), IEEE Press, New York, USA, pp 5822–5827
5. Jayaputera J, Poernomo I, Schmidt H (2004) Runtime verification of timing and probabilistic properties using WMI and NET. In: 30th Euro micro conference, IEEE Press, New York, USA, pp 100–106

Part IX
Cross Strait Conference on Information Science and Technology

Research on Micro-Blog Information Perception and Mining Platform

Xing Wang, Fei Xiong and Yun Liu

Abstract To predict the tendency of Micro-blog information dissemination, provide the early warning of the Internet emergencies, and contribute to the content security of micro-blog, the paper offers a platform for Micro-blog information perceiving and mining. This platform is an integration of Micro-blog data collection and processing module, topic detection and tracking module, user behavior analysis module, trend prediction module, etc. It could access and analyze micro-blog information automatically, leading a positive significance to grasp the emergencies on micro-blog. This paper puts forward methods based on the Latent Dirichlet Allocation (LDA) document clustering and hot topics prediction, which could analysis and predict the micro-blog data effectively, avoiding the problems in the traditional algorithm. Also, these methods have a higher accuracy for clustering and prediction.

Keywords Micro-blog · Data mining · Text clustering · Hot topic detection

X. Wang
China Information Technology Security Evaluation Center, Beijing, China
e-mail: wangx@itsec.gov.cn

F. Xiong (✉) · Y. Liu
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: 08111029@bjtu.edu.cn

Y. Liu
e-mail: liuyun@bjtu.edu.cn

F. Xiong · Y. Liu
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_85, © Springer Science+Business Media Dordrecht 2014

Introduction

Micro-blog, is an integrated, open Internet social networking services in the Web2.0 era. It is an system of immediate message releasing, similar to the blog. Users can use the following ways to announce its messages, such as mobile phones, instant messaging software, and external application programming interface,. The general message is less than 200 words (in practice, generally 140 words). The micro-blog makes more and more people participate in the Internet interactions. Many scholars at home and abroad, have carried out in-depth research on micro-blog, which includes the characteristics, propagation mechanism, development trends, and the social influences.

In recent years, scholars have conducted a lot of research on Internet information processing and mining. Han Ruixia introduced the characteristics and basic concepts of micro-blogging platform [1]. Kang Shulong studied the user network structure and degree distribution characteristics of Sina micro-blog [2]. Wang Rui explained empirically the relationship between the number of user's friends and the popularity level of the user [3]. In the area of text clustering and topic detection, Bruno et al. proposed a logarithmic maximum-likelihood criterion to calculate the word weights, instead of the traditional algorithm [4]. In the field of text vector extraction, Yang et al. improved the similarity formula, and presented a decay function of similarity calculation with time [5]. A lot of research methods for hot topic mining originate from Topic Detection and Tracking (TDT). TDT usually handles the document stream which dynamically changes with the news, rather than a set of texts on Internet [6]. Gabriel Piu et al. proposed a hot topic identification scheme based on free parameters, which determines the distribution of hot topics within a certain period of time by the characteristics of text [7]. In the field of user behavior and trend forecasting, Jiang Zhu extracted the features related to user's retweeting behavior, including content influence, network influence, and time influence [8]. Fei Xiong et al. analyzed the factors associated with the post propagation in online forums and combined these factors to predict the potential hot topics [9].

The recent research mainly aims at information on the Web pages, such as news and online forums. However, information on micro-blog and other social networks is more latent, and user behavior and social relationships are often heterogeneous and variable. Moreover, the text information of a message is very small. Therefore, traditional network monitoring systems based on text analysis cannot work well on micro-blog. We need to develop a content-aware and user behavior mining platform for micro-blog to provide public opinion analysis and early warning. This paper presents a micro-blog information perception and mining platform, which integrates the function of micro-blog user behavior mining, the social relationship mining, the topic detection, and so on. This work can help grasp the online emergencies, guide public opinion towards the right direction, and build a green and healthy online environment.

The remainder of the paper is organized as follows. In section Functional overview of the platform, we give the overview of the platform, and describe its functional modules. In section Research on key technology we introduce the key methods of our platform. We put forward a clustering algorithm for micro-blog short texts, and present a new method to detect hot topics in terms of ensemble learning. We conclude the paper in section Conclusions.

Functional Overview of the Platform

Our system includes the module of micro-blog data acquisition and processing, topic detection and tracking, user relationship analysis, user behavior mining, and propagation trend prediction. It can collect micro-blog information, provide information monitoring and early warning, discover potential hot topics, and detect user abnormal behavior. The functional modules of our platform are shown as Fig. 1.

1. Micro-blog data acquisition and processing

In consideration of the problems in micro-blog data acquisition and processing, this module puts forward a method for micro-blog data collection based on the

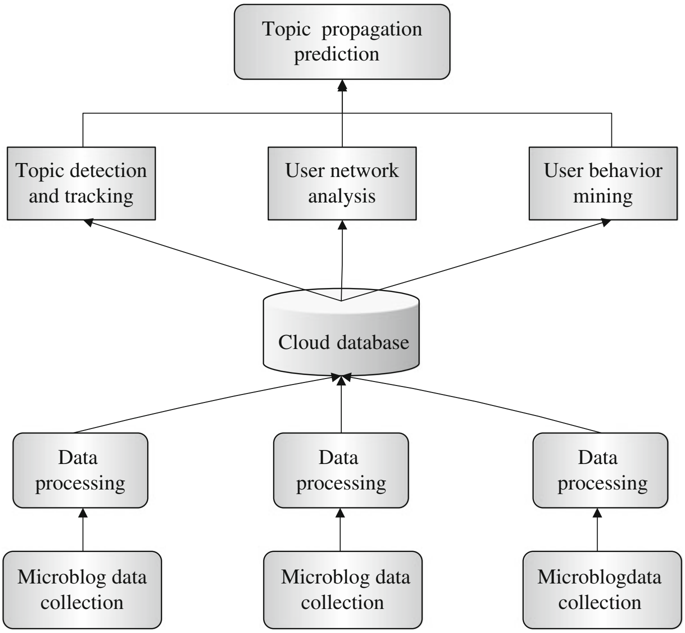


Fig. 1 Functional modules of the platform

combination of open API and page crawling and parsing technique. In the data collection module based on open API, the system obtains the data results that the API requests by parsing JSON object, and controls the query frequency of API by thread scheme. For the data outside the available scope of API, we use the web crawler to collect relevant page files by simulating the user's login process on micro-blog. By establishing different web page templates, the system can parse different categories of web pages downloaded by crawler. Therefore, our platform realizes the comprehensive and efficient data acquisition.

2. Micro-blog topic detection and tracking

This module can identify topics in the information flow of micro-blog media automatically by combining methods of information extraction and content clustering. Considering the feature of small text for micro-blog posts, this module builds a model of the topic expression and detection. The module achieves the mining of meaningful strings and the extraction of network characteristics, and can also identify different sentimental values of posts belonging to the same topic. Moreover, it can detect new and unknown topics including the event occurring first and related events, from the massive information on Internet.

3. Micro-blog user relationship analysis

The user relationship network on micro-blog is different from the traditional social network or interpersonal network. According to complex network theory, this module computes the macroscopic characteristics of network structure, and compares these characteristics in different network communities. This module can also find the heterogeneity of different virtual communities, and analyze user behavior and preferences.

4. Micro-blog user behavior mining

The module processes and analyzes user data, and finds out the potential rules of user behavior. According to the queuing theory, it researches on the interaction between users and micro-blog communities, and simulates user's behavior in the network, and reflects the characteristic distribution of user participation.

5. Micro-blog topic trend prediction

This module uses mathematics, information science, and other interdisciplinary approaches to model and forecast the diffusion process of information and evolution of public opinion. Meanwhile, in order to find general control strategy, this module can simulate the effect of different guide techniques on the opinion formation.

Research on Key Technology

Micro-Blog Short Text Clustering Based on LDA Algorithm

A topic consists of a series of related posts, containing the description of cause of the incident, the evolution, results and impact of the event. In our micro-blog perception and mining platform, micro-blog posts should be clustered into topics, and these posts about the same event are grouped in the same cluster. Only in this way, the impact of the social event on Internet can be measured, and the clustering also provides a basis for hot topic detection and evolutionary trend prediction. Topic clustering is implemented according to the similarity between the contents of posts.

However, all the micro-blog posts are short texts, which are no more than 140 words. After converting these posts to term vectors, any two posts rarely have overlap words. Therefore, the accuracy of the similarity between these posts is not high even if synonym substitution is used in the calculation. Brief contents make the cosine similarity between the posts belonging to the same topic very low. Here, we use the LDA algorithm to carry out document clustering. LDA is a generative model which treats a post as the mixing of multiple topics [10]. Each post has a different function of topic distribution, and therefore a post may belong to more than one topic. LDA does not need to compare the similarity between two posts, greatly improving the accuracy of clustering.

LDA model assumes that each word in posts is generated by topics, each topic corresponds to a word distribution function, and each post also corresponds to a topic function. The topic distribution of post i is defined as θ_i, and the initial distribution of θ_i is Dirichlet distribution for parameter α. The word distribution of topic k is φ_k, and the initial distribution of φ_k is also Dirichlet distribution for parameter β [11, 12]. With these two distribution functions, the post i can be constructed from these topics. For each word in the post, we choose a topic z_{ij} according to multinomial distribution of θ_i, and then in the light of the topic, choose a word w_{ij} following the multinomial distribution of $\varphi_{z_{ij}}$. We repeat this process until we reproduce the entire post. Therefore, if word t is assigned to topic k for n_k^t times, and then the posterior probability of words is

$$P(w|z,\varphi) = \prod_{k=1}^{K}\prod_{t=1}^{V}\left(\varphi_k^t\right)^{n_k^t} \tag{1}$$

K is the total number of topics, V is the total number of words in the lexicon. Similarly, if the topic k appears n_i^k times in the document i, then the posterior probability of topics is

$$P(z|\theta) = \prod_{i=1}^{M}\prod_{k=1}^{K}\left(\theta_i^k\right)^{n_i^k} \tag{2}$$

M is the total number of posts.

For each micro-blog post, we get its topic distribution θ. If the largest component of θ for a post is greater than 0.5, then the post is assigned to the topic corresponding to that component. If the largest component of θ is lower than 0.5, then the post is assigned to the first three topics ranking by probability of topic distribution. Before LDA clustering, we need to determine the total number of topics K based on validity index. Here, we calculate the number of topics by text databases [13]. Figure 2 illustrates the process of short text clustering for micro-blog posts.

Hot Topic Prediction Based on Ensemble Learning

The propagation extent of a micro-blog topic is closely related to the attraction of the topic itself, current active users' behavior, and the competition between the topics. During the process of micro-blog topic dissemination, there are new posts published about this topic continuously, and the most powerful mechanism to diffuse information is post retweeting based on user relationship. Thus, the unique way of micro-blog information retweeting, makes the topic propagation show different characteristics from traditional networks. The diffusion process is more

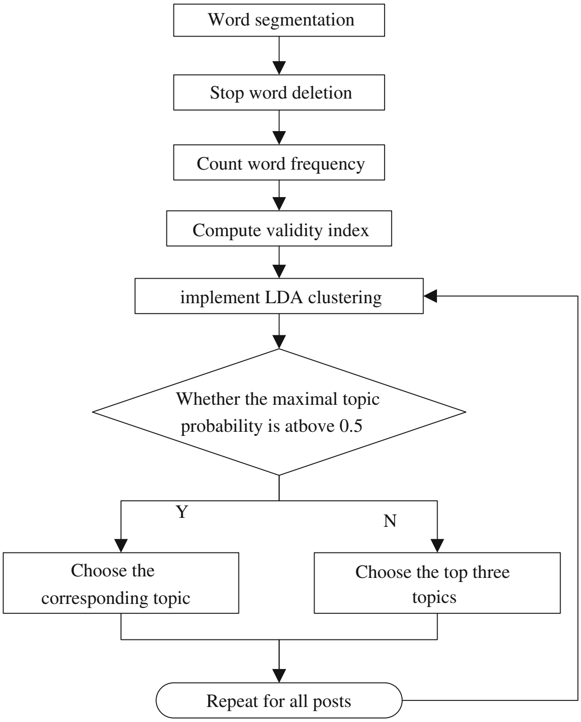


Fig. 2 Flow chart of LDA clustering for short texts

explosive and more capricious. The features of initial topic dissemination reflect the overall trend to a certain extent, and can be used as an indicator for judging whether a topic is popular. However, some topics absorb only a few users at the beginning, but after some influential users retweet it, it becomes a hot topic. The topic attracts numerous retweeting, and affects a large number of users in the network. Therefore, we need to extract multiple features of topic dissemination, and create a model to predict potential hot topics.

First, we implement short text clustering for micro-blog posts. Posts with similar contents are put into a cluster, and each cluster corresponds to a topic. Compared with topics about biology or tourism, topics of social issues are always more popular. The more popular a topic is, the more users it would attract. Generally, most topics cannot get a mass of attention, and will gradually die out, while a few topics suddenly become hot after coherent interaction. Now we have a question how to define a hot topic. There are a variety of definitions. Erzhong Zhou [14] defined hot topics by the number of related posts, the number of replies, and user participation. Zhongfeng Zhang [15] described the popularity of a topic by the total number and time evolutionary frequency of hot words that appear in a document. Here, we do not concern more about the detailed definition of hot topics, but only use the total number of participants as the popularity of a topic. If the number of participants is greater than 1,000, the topic is considered as a hot topic.

Now, according to the data in the initial δ hours of each topic, we extract the features related to topic dissemination, including the attractiveness of topic content, and user behavior. The initial data of a topic is related to its popularity. In the topic content influence, we calculate the total number of posts in the first δ hours, the total number of replies, the total times of retweeting, the number of participants, the current number of other topics, and the average frequency of retweeting. In all, we have 6 content features. User behavior also affects the propagation of topics. We can extract 5 user features, that is, the number of friends for post author, the number of followers, the average times of being retweeted for post author, the total times of being retweeted for the users who has retweeted the post, and the current number of active users.

According to the previously mentioned 11 features that are related to the popularity of posts, we use support vector machines to combine these features to generate a predictive model. The support vector machine is an optimal margin classifier, which uses a linear mapping, to find a hyperplane as the boundary that can divide the data into two classes. The support vector machine makes the data points nearest from the hyperplane in both sides, get greater spacing [16].

The larger δ is, the higher recall and accuracy of predicted results we can get. Considering the need of network monitoring, we hope to identify potential hot topics as soon as possible, so as to provide early warning of network emergencies. In practice, we usually set δ at $\delta < 8$ h.

In order to improve the performance of the model, now we use the bagging method of ensemble learning to deal with the training data set. If the number of samples in training sets D is n, the bagging method divides D into m subsets D_i,

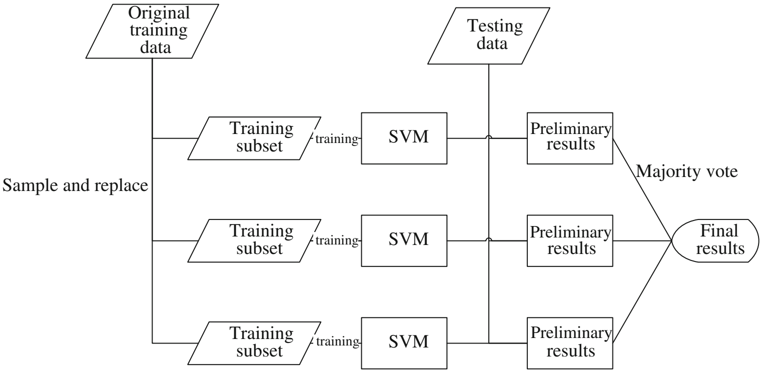


Fig. 3 Flow chart of ensemble learning process

and D_i includes n_i samples satisfying $n_i < n$ [17]. We uniformly sample data from D, and replace the elements randomly, to generate a subset D_i. We use each D_i respectively to train a model, and then use these trained support vector machines to predict new topic. Then we calculate the average or do majority vote of all the preliminary results to get our final prediction result. Using bagging algorithm to deal with the training data set can improve the robustness and the precision of the model, and avoid over fitting as well. The testing results indicate that the bagging method can, to some extent, improve the accuracy of the predicted results while keeping the recall almost unchanged. We sample data from the original training set. The original data set is divided into 3 groups, each of which is used to train a support vector machine separately. For a new topic, we use the three support vector machines to predict its popularity, and do majority vote for the 3 preliminary results. Figure 3 shows the ensemble learning process.

Conclusions

With the development of micro-blog and the explosive growth of information, the security of micro-blog content has attracted more and more attention. In recent years, the processing and mining of large amounts of data for micro-blog become a research hotspot. In this paper, we presented a micro-blog data perception and mining platform. It can automatically collect micro-blog data and analyze the data, detect hot topics and predict the diffusion trends of topics. Moreover, the platform shows a better performance on short text processing of micro-blog, and is able to predict hot topics effectively. Therefore, the system can provide early warning of online emergencies on micro-blog, and ensure the content security of micro-blog more forcefully.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Han R (2010) The influence of micro-blogging on personal public participation. In: Proceeding(s) of 2010 IEEE 2nd symposium on web society, pp 615–618
2. Kang S, Zhang C (2010) Complexity research of massively micro-blogging based on human behaviors. In: Proceeding(s) of 2nd international workshop on database technology and applications, pp 1–4
3. Wang R, Jin Y (2010) An empirical study on the relationship between the followers' number and influence of micro-blogging. In: Proceeding(s) of the international conference on e-business and e-Government, pp 2014–2017
4. Pouliquen B, Steinberger R et al (2004) Multilingual and cross-lingual news topic tracking. In: Proceeding(s) of the 20th international conference on computational linguistics, pp 23–27
5. Yang Y, Pierce T, Carbonell J (1998) A study on retrospective and on-line event detection. In: Proceeding(s) of the 21st annual international ACM SIGIR conference on research and development in information retrieval, pp 28–36
6. Jin H, Schwartz R, Wall F (1999) Topic tracking for radio, TV broadcast, and newswire. In: Proceeding(s) of the DARPA broadcast news workshop, pp 199–204
7. Pui G, Fung C, Yu JX, Lu H (2005) Parameter free bursty events detection in text streams. In: Proceeding(s) of the 31st international conference on very large data bases, pp 181–192
8. Zhu J, Xiong F, Piao D, Liu Y, Zhang Y (2011) Statistically modeling the effectiveness of disaster information in social media. In: Proceeding(s) of the IEEE global humanitarian technology conference, pp 431–436
9. Xiong F, Liu Y, Zhu J, Lian J, Zhang Y (2012) Hot post prediction in BBS forums based on multifactor fusion. J Convergence Inf Technol 7(12):129–137
10. http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
11. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(4–5):993–1022
12. Farrahi K, Gatica-Perez (2011) Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans Comput Logic 2(1):3
13. Can F, Ozkarahan EA (1990) Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. ACM Trans Database Syst 15:483–517
14. Zhou E, Zhong N, Li Y (2011) Hot topic detection in professional blogs. Active Media Technol 6890:141–152
15. Zhang Z, Li Q (2011) QuestionHolic: hot topic discovery and trend analysis in community question answering systems. Expert Syst Appl 38(6):6848–6855
16. Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300
17. http://en.wikipedia.org/wiki/Bootstrap_aggregating

A New Algorithm for Personalized Recommendations in Community Networks

Xin Zhou, XinXiang Xing and Yun Liu

Abstract In a graph theory model, clustering is the process of division of vertices in groups, with a higher density of edges in groups than among them. In this paper, we introduce a new clustering algorithm for detecting such groups; we use it to analyze some classic social networks. The new algorithm has two distinguished features: non-binary hierarchical tree and the feature of overlapping clustering. A non-binary hierarchical tree is much smaller than the binary-trees constructed by most traditional algorithms; it clearly highlights meaningful clusters which significantly reduce further manual efforts for cluster selections. The present algorithm is tested by several bench mark data sets for which the community structure was known in advance and the results indicate that it is a sensitive and accurate algorithm for extracting community structure from social networks.

Keywords Clustering · Graph theory · Hierarchical tree · Social network

X. Zhou
China Information Technology Security Evaluation Center, Beijing, China
e-mail: zhoux@itsec.gov.cn

X. Xing · Y. Liu (✉)
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: liuyun@bjtu.edu.cn

X. Xing
e-mail: 10111028@bjtu.edu.cn

X. Xing · Y. Liu
Key Laboratory of Communication and Information Systems, Beijing Municipal
Commission of Education, Beijing Jiaotong University, Beijing 100044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_86, © Springer Science+Business Media Dordrecht 2014

Introduction

Clustering is an important task for the discovery of community structures in networks. Its goal is to sort cases (people, things, events, etc.) into clusters so that the degree of association is relatively strong between members of the same cluster and relatively weak between members of different clusters. Webster [1] defines cluster analysis as "a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics". Various clustering algorithms have been proposed in the literature in many different scientific disciplines. Jain [2] broadly divided these algorithms into two groups: (1) hierarchical algorithm and (2) partitional algorithm. Hierarchical clustering algorithms recursively find nested clusters either in agglomerative mode or in divisive mode. The most well-known hierarchical algorithms are single-link and complete-link; in single-link hierarchical clustering, the two clusters whose two closest members have the smallest distance are merged in each step; in complete-link case, the two clusters whose merger has the smallest diameter are merged in each step. Compared to hierarchical clustering algorithms, partitional clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure. The most popular and the simplest partitional algorithm is K-means [3]. Berkhin [4] listed another six classifications besides the above two main groups: (3) Grid-Based Algorithms, (4) Algorithms Based on Co-Occurrence of Categorical Data, (5) Constraint-Based Clustering, (6) Clustering Algorithms used in Machine Learning, (7) Scalable Clustering Algorithms, (8) Algorithms for High Dimensional Data.

In recent years, a growing number of clustering algorithms for categorical data have been proposed based on various centrality measures. For instance, vertex betweenness has been studied by Freeman [5] as a measure of the centrality to detect communities in a network; Girvan and Newman [6] generalize vertex betweenness centrality to edge in order to discover community structures; Frey and Dueck [7] devised a algorithm called affinity propagation, which takes as input measures of similarity between pairs of data points. Newman [8] utilizes the eigenvectors of matrices to find community structure in networks; Rosvall and Bergstrom [9] use the probability flow of random walks on a network as a proxy for information flows in the real system and decompose the network into modules. Wu and Huberman [10] propose an approach for discovering the communities based on the property of resistor networks. For reviews see Refs. [11, 12].

A social network [13, 14] is a set of people or groups each of which has connections of some kind to some or all of the others. A cluster is a collection of individuals with dense relationship internally and sparse relationships externally. Based on this criterion, we introduce a new clustering algorithm for detecting community structures. In this algorithm, individuals and their relationships are denoted by weighted graphs, then the graph density we defined gives a better quantity depict of whole correlation among individuals in a community, so that a reasonable clustering output can be presented. Compared with other algorithms, this algorithm has two important features:

1. A much smaller hierarchical tree that clearly highlight meaningful clusters.
2. Overlapping clusters.

To evaluate the effectiveness of our algorithm, we applied it to analyse some classic bench mark data sets whose clusters are already known. These data sets include Karate Club, Davis southern club women, Dolphin, Books about US politics, American College Football. The accuracy of the outputs in those classical benchmark data sets is a supporting evidence of further applicability of the new algorithm.

The rest of the paper is organized as follows. In section A new clustering algorithm, we introduce the details of the new dense subgraph clustering algorithm. In section Applications of the algorithm in social networks, we apply it to some classic social networks and compare its results with that of known clusters. Finally, a summary and conclusions are given in section Conclusion.

A New Clustering Algorithm

A graph or network is one of the most commonly used models to present real-valued relationships of a set of input items. Let G = (V, E) be a graph with the vertex set V and the edge set E with weight w(e) on every edge e. Models with un-weighted graphs (the weight of every edge is set to 1) have been extensively studied in graph theory. In an un-weighted graph G, a subgraph H of G is defined as a clique if every pair of vertices of H is joined by one edge [15–17]. It is well-known that the search of maximum cliques in graphs is an NP-complete problem [18]. Therefore, it is not practical to define cliques as clusters. Furthermore, there is no appropriate definition for a clique in a weighted graph. However, in order to closely represent the nature and the real situation of the inputs in most applications (different degrees of similarity for clustering problems), we should use weighted graph models which are much more appropriate than un-weighted models. For simplification or other practical reasons, many designers of clustering algorithms may set a specific threshold, such as that any edge with weight below the threshold is deleted and the remaining ones have no associated weight. However, one may not be able to expect an accurate output since the cut-off (by threshold) may cause a loss of important information.

For a subgraphC(|V(C)| > 1), we define the density of C by

$$d(C) = \frac{2\sum_{e \in E(C)} w(e)}{|V(C)|(|V(C)| - 1)} \tag{1}$$

As seen above, if w(e) = 1 for every edge e in C and d(C) = 1, then the subgraph C induces a clique. For a weighted graph, a subgraph C is called a Δ-quasi-clique if d(C) ≥ Δ for some positive real number Δ.

Since clustering is a process that detects all dense subgraphs in G and construct a hierarchically nested system to illustrate their inclusion relation, a heuristic process is applied here for finding all quasi-cliques with density invarious levels. The core of the algorithm is deciding whether or not to add a vertex to an already selected dense subgraph C. For a vertex v $\notin$ V(C), we define the contribution of v to C by

$$c(v, C) = \frac{\sum_{u \in V(C)} w(uv)}{|V(C)|} \tag{2}$$

A vertex v is added into C if $c(v, C) > \alpha d(C)$ where α is a function of some user specified parameters.

Instance: G = (V, E) is a graph with edge weights w : $E(G) \mapsto R^+$.

Question: Detects Δ-quasi-cliques in G with various levels of Δ, and construct a hierarchically nested system to illustrate their inclusion relation.

Sub-Algorithm Growing(C, G):

(Grow a Community C in G)

```
while $V(G) - V(C) \neq \emptyset$
begin
pick $v \in V(G) - V(C)$ such that c(v, C) is a maximum
if $c(v, C) > \alpha_n d(C)$ then add v to C (where n = |V(C)|,
$\alpha_n = 1 - \frac{1}{2\lambda(n+t)}$, with $\lambda \geq 1$ and
t $\geq$ 1 as user specified parameters)
else return
end
```

Sub-Algorithm Decompose(G, w_0):

(decompose a graph G into communities using edges with weights at least w0).

```
Let $E_0 = \{e \in E(G) : w(e) \geq w_0\}$
For each $e = uv \in E_0$ in decreasing order of w(e)
begin
if either u or v is not in any community
then
begin
create a new empty community C and add u, v in it
  Growing(C, G)
end
end
```

Sub-Algorithm Merging(G):

```
For any two communities $C_i$ and $C_j$ in G, if

$|C_i \cap C_j| > \beta \min(|C_i|, |C_j|)$

then merge $C_i$ and $C_j$ into a new community $C = C_i \cup C_j$ (where $\beta$
is a userspecified parameter).
```

```
Contract each community to a vertex. The weight of an edge
is defined by
```

$$w(C_i, C_j) = \frac{\sum_{e \in E_{i,j}} w(e)}{|E_{i,j}|}$$

```
where the set of crossing edges
```

$$E_{i,j} = \{v_i v_j : v_i \in C_i, v_j \in C_j, v_i \neq v_j\}$$

Main-Algorithm

```
(generate hierarchic clustering tree or a graph G)
While $E(G) \neq \emptyset$
begin
Choose $w_0$ according to some criterion
Decompose(G, $w_0$)
Merging(G)
Store the resulted graph to G
end
Trace the movement of each vertex and generate the hierarchic tree.
```

Applications of the Algorithm in Social Networks

In this section, we present a number of applications of our algorithm to some classic social networks for which the community structure is already known and compare its results with that of preceded algorithms.

Zachary's Karate Club Study

The first social network is the well-known "karate club" of Zachary [19]. He observed 34 members of a karate club over 2 years. During the course of observation, the club members split into two groups because of the disagreement between the administrator of the club and the club's instructor, the members of one group left to start their own club. Zachary constructed a simple unweighted graph to show the friendships between two members of the club, each member in the club is represented by a node, and edge is drawn if the two members are friends outside the club activities.

Davis Southern Club Women

The data of the social participation of eighteen women in "Old City" was collected by Davis et al. [20]. The data (see Table 1) is a table with 18 rows—one for each woman—and 14 columns, one for each "event" (such as a club meeting, a church supper, a card party, etc.), held during the course of a year. For the simplicity, we use number 1–18 denote the 18 women, then a matrix A is generated to record their attendances of events: A(i, j) is 1 if woman i attended social event j, and 0 otherwise. The goal of the study was to determine the clique structure according to their records of attendances. The clique membership reported by Homans [21] is as follows. Group 1:1,2,7,8,14,15,16; Group 2: 11,12,13,17,18; women not clearly belonging to either groups: 3,4,5,6,9,10.

Dolphin's Network

The next social network was constructed from observations of a bottlenose dolphin community [22–24]. There are 62 nodes and 159 edges in this network: nodes represent the dolphins, edges between nodes represent associations between dolphin pairs occurring more often than expected by chance. This network is interest because, during the course of the study, the dolphin group split into two smaller subgroups following the departure of a key member of the population. The subgroups of the actual division in Newman [8] are represented by the shapes of the vertices (see Fig. 6), the squares and circles represent the actual division of the network observed when the dolphin community split into two as a result of the departure of a keystone individual. The individual who departed is represented by the yellow triangle. The dotted line denotes the division of the network into two equal-sized groups found by the standard spectral partitioning algorithm. The solid curve represents the division found by the algorithm based on the leading eigenvector of the modularity matrix in Newman [8]. Its result corresponds quite closely to the actually split—all but 3 of the 62 dolphins are misclassified.

We have also applied our algorithm to this network, the result corresponds perfectly with the actual division. The algorithms of Girvan and Newman [25] and Newman [12] also give precisely the same result.

An Example of Overlapping

Most of clustering algorithms generate non-overlapping groups, which are useful for graph drawing but are not as good for group analysis, since real social groups are usually more complex, involving different degrees of overlap among groups. In this section, we use an example in Santamaria and Theron [26] to illustrate the overlapping feature of our algorithm.

Conclusion

In this paper, we introduce a new clustering algorithm for detecting community structure in network and use it to analyze some classic social networks. Compared with the existing algorithms, the new algorithm has two distinguished features:

1. A much smaller hierarchical trees that clearly highlight meaningful clusters. It was pointed out in SAS/STAT User's Guide [27], "there are no completely satisfactory algorithms for determining the number of population clusters for any type of cluster analysis". Hence, a relatively small hierarchical tree in an output will significantly reduce the human involvement in the final selection of clusters. In Fig. 2 and Fig. 3, one may notice that each hierarchical tree (for karate club) with 34 leafs (inputs) has 33 internal nodes. By using the new algorithm, the hierarchical tree (see Fig. 4) contains only 22 internal vertices. Similarly, a hierarchical tree for women club with 18 leafs usually have 17 internal nodes, while the tree in Fig. 5 has only 8 internal vertices.
2. The feature of overlapping clustering does reflect the complexity of our real world. One may notice the overlapping clustering feature in Fig. 4: At the right end of the graph, a small group of club members {24, 30, 33, 34, 9, 31} hold multi-memberships in three internal clusters (internal nodes on the third level, right end). The overlapping clustering is a concept that has recently received increased attention in Palla et al. [28], Pereira-Leal et al. [29], Futschik and Carlisle [30], etc. One may also notice this feature in Fig. 14: The element n3 and n7 present in two groups and the node n4 presents in three groups. According to mathematical definition in graph theory, the "hierarchical trees" in Figs. 4 and 14 are not really trees–they are hierarchical networks in which the relations of clusters are hierarchically nested.

For further research, we will consider to develop an automated value selection algorithm for each parameter. Determine a function for each parameter in terms of some structural information of the input graph, so that graphs with different structures (density, connectivity, locally or globally) will be automatically assigned proper values.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Merriam-Webster Online Dictionary (2008) Cluster analysis. http://www.merriam-webster-online.com
2. Jain AK (2009) Data clustering: 50 years beyond K-means. http://dataclustering.cse.msu.edu/papers/JainDataClusteringPRL09.pdf

3. Steinhaus H (1956) Sur la division des corpmateriels en parties. Bull Acad Polon Sci C1. III IV:801–804
4. Berkhin P (2009) Survey of clustering data mining techniques, http://www.ee.ucr.edu/barth/EE242/
5. Freeman L (1977) A set of measures of centrality based upon betweenness. Sociometry 40:35–41
6. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826
7. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315:972–976
8. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104
9. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105(4):1118–1123
10. Wu F, Huberman BA (2004) Finding communities in linear time: A physics approach. Eur Phys J B 38:331–338
11. Danon L, Duch J, Diaz-Guilera A, Arenas A (2005) Comparing community structure identification. J Stat Mech 2005:P09008
12. Newman MEJ (2004a) Detecting community structure in networks. Eur Phys J B 38;321–330
13. Scott J (2000) Social network analysis: a handbook, 2nd ed. Sage Publications, London
14. Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge
15. Bondy JA, Murty USR (1976). Graph theory with applications. Macmillan, London
16. Diestel R (2005) Graph theory, Graduate texts in mathematics, vol 173, 3r edn. Springer, Heidelberg
17. West W (1996) Introduction to graph theory. Prentice Hall, Upper Saddle River,NJ
18. Gary MR, Johnson DS (1979) Computers and intractability. Freeman, NY
19. Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33(4):452–473
20. Davis A, Gardner BB, Gardner MR (1941) Deep South: a social anthropological study of caste and class. University of Chicago Press, Chicago
21. Homans GC (1950) The human group. Harcourt, Brace and World, New York
22. Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM (2003) The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? Behav Ecol Sociobiol 54:396–405
23. Lusseau D (2003) The emergent properties of a dolphin social network. Proc R Soc London B270:S186–S188
24. Lusseau D, Newman MEJ (2004) Identifying the role that individual animals play in their social network. Proc R Soc London B(Suppl.) 271:S377–S481
25. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
26. Santamaria R, Theron R (2008) Overlapping clustered graphs: coauthorship networks visualization. Lect Notes Comput Sci 5166:190–199
27. SAS Institute Inc (2003) Introduction to clustering procedures, Chapter 8 of SAS/STAT User's Guide.(SAS OnlineDocTM: Version 8) http://www.math.wpi.edu/saspdf/stat/pdfidx.htm
28. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
29. Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. PROTEINS: Struct Funct Bioinf 54:49–57
30. Futschik ME, Carlisle B (2005) Noise-robust soft clustering of gene expression timecourse. J Bioinf Comput Biol 3:965–988
31. Breiger RL (1974) The duality of persons and groups. Soc Forces 53(2):181–190

32. http://www-personal.umich.edu/mejn/netdata/
33. Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49:291–307
34. Krebs V (2009) http://www.orgnet.com/
35. Newman MEJ (2004b) Fast algorithm for detecting community structure in networks. Phys Rev E 69:066133

Design and Implementation of NGDC Geospatial Metadata Management System

Li Zhang

Abstract As an important part of the geospatial framework, National Geospatial Data Center (NGDC) is a website which can provide potential users with convenient geospatial metadata query services. It will improve the development of Digital City strongly. This paper firstly analyzes the content of the geospatial metadata. It puts forward the concept of the Integrated Geospatial Metadata based on the analysis of geospatial metadata query conditions entered in the NGDC website. The NGDC geospatial metadata management system is developed according to the Integrated Geospatial Metadata.

Keywords Geospatial metadata · National geospatial data center (NGDC) · Geospatial framework · Digital city · Internet · Geographical information system (GIS)

Introduction

With the deepening of information construction, the State Bureau of Surveying and mapping geographic information made the strategic decision of building a national public service platform of geographic information in order to implement the scientific development concept [1]. In 2006, the State Bureau of Surveying and Mapping took the Digital City Geospatial Framework Construction as one of "Eleventh Five-Year" key project [2, 3]. Geospatial framework includes the policies, regulations, standards, technology, facilities, mechanisms and human resources involved with the collection, processing, exchange of geospatial data. It is the basis of the national economic and social information supporting platform and the foundation which a variety of information resources spatially reference to [4, 5].

L. Zhang (✉)
School of Computer Engineering, Shenzhen Polytechnic, Shenzhen 518055 Guangdong, China
e-mail: zhangli@szpt.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_87, © Springer Science+Business Media Dordrecht 2014

In order to meet the growing demand for geospatial data, geospatial data production in China has begun to take shape in recent years. A platform is required, which provides users with the geospatial metadata query service according to coverage, layer features, price and other factors. National Geospatial Data Center (National Geospatial Data Center, hereinafter referred to as "NGDC") will be able to provide such a platform which can complete the transition of geospatial data from data producers to data users.

Geospatial metadata is the data foundation to achieve NGDC. The geospatial metadata management system in NGDC can effectively avoid the repetitive production of geospatial data, and can achieve the spatial information resource sharing between different agencies or departments.

Geospatial Metadata and NGDC

How to help potential data users to search the specified geospatial data set more effectively, more easily from such a wide variety of geospatial data products has become an unavoidable problem. It makes the study of geospatial metadata become so important.

Geospatial Metadata refers to the description of geospatial data and information resources. In fact, it is the generalization and extraction of the attributes and spatial characteristics of geospatial data sets or products. In short, the geospatial metadata will help data users know whether the content and quality of the specified geospatial data set meets the requirements of their actual GIS applications, whether the source of geospatial data set is reliable, whether the data format is compatible with other data formats before they get or buy the geospatial product.

Actually NGDC is a geospatial information service network platform. As a bridge between geospatial data producers and data users, NGDC can provide the data producers with a unified geospatial metadata standard and help them to deploy their geospatial metadata. On the other hand, NGDC can provide the efficient and effective query of geospatial metadata for data users in order to make it easy to get the existing geospatial data products of appropriate precision, which are updated and upgraded easily [6].

The Integrated Geospatial Metadata

One of the tasks of NGDC is to develop a set of practical geospatial metadata standard. This metadata standard is the foundation of NGDC geospatial metadata management system. In theory, geospatial metadata standard is divided into two levels. The first level is the directory information, which is mainly the macroscopic description of the data set. It is suitable for National Geospatial Data Center or global management and query spatial information. The second level usually includes the detailed or comprehensive description of the geospatial data set when they are deployed by geospatial data producers.

However, it does not need such detailed metadata fields for most data users to get geospatial information through NGDC web site. In other words, data users are usually asked to become a registered member of NGDC, or they can not login NGDC official website. After login, the user may input some query conditions (such as the scale of the digital map, coverage, etc.) in the web page of geospatial metadata query on NGDC web site. The geospatial metadata management system will filter out some candidates of geospatial data products according to the specified conditions. Then the data users will contact with data producers by telephone or visit them personally in order to consult related issues (such as technical parameters, view sample images, and so on) because the query result includes the contact information of data producers.

After analyzing the query conditions entered by the data user in the query page, we can extract a portion of geospatial metadata in order to form "the Integrated Geospatial Metadata"(hereinafter referred to as "IGM"), which includes the basic information of the data set, data publishers, geographical coverage, elevation range four aspects. The details of IGM are shown in Table 1.

Implementation of NGDC Geospatial Metadata Management System

As mentioned earlier, the main task of NGDC is to collect geospatial metadata provided by data producers, and to extract the relevant metadata information into the formation of IGM saved in NGDC database. At the same time, the official website of NGDC provides a hyperlink to geospatial metadata query page in order to provide geospatial metadata query services.

The NGDC geospatial metadata management system includes three components which are the user management function module, the function module of geospatial metadata maintenance, the function module of geospatial metadata query. The system functional block diagram is shown in Fig. 1.

The functions of user management module include user registration, user login, and user information maintenance and other functions. The types of NGDC user include administrators, data producers, data users, guests. Administrators can modify or delete the information for all users. Data users can query metadata after registering as a NGDC user. Guest can only browse NGDC site. The functions related to metadata maintenance and query will be available to someone only after he is a registered user in NGDC official web site.

The functions of geospatial metadata maintenance module consist primarily of functions of adding, modifying and deleting geospatial metadata. Geospatial metadata referred to here mainly refers to IGM, which is mainly used to support the geospatial metadata query function provided by NGDC through Internet. The

Table 1 The details of the integrated geospatial metadata

No.	Metadata item	Data item description	Remarks
1	Name	Name of data set or products	Basic information of data set
2	Version	Version of data set or products	
3	Scale	Scale of data set or products	
4	Abstract	Brief summary of the contents of the data set	
5	Objective	A brief description of the purpose to establish the data set	
6	Subject	The thematic name of data set, such as elevation, cadastral, geology, hydrology, vegetation, marsh, transportation, soil, etc.	
7	Key words	The feature words which can summarize the particular aspect of data sets	
8	Name of deployer	The name of individual or organization that are directly responsible for the contents of the data set	Information of data deployers
9	Province	The province, municipalities, district or Special Administrative Region of data deployer	
10	City	The city name of data deployer	
11	Post address	The detailed post address of data deployer	
12	Postal code	The postal code of data deployer	
13	Contact telephone	The telephone number of data deployer	
14	Fax	The fax phone number of data deployer	
15	E-mail	The E-mail address of data deployer	
16	Web site	The web site address of data deployer if exists	
17	Contact	The name of the contact person	
18	Longitude minimum	The minimum value of longitude for geospatial coverage	Geographical coverage of data set
19	Longitude maximum	The minimum value of longitude for geospatial coverage	
20	Latitude minimum	The minimum value of latitude for geospatial coverage	
21	Latitude maximum	The minimum value of latitude for geospatial coverage	
22	Minimum elevation	Minimum elevation value of data set (unit: m)	Elevation range
23	Maximum elevation	Maximum elevation value of data set (unit: m)	

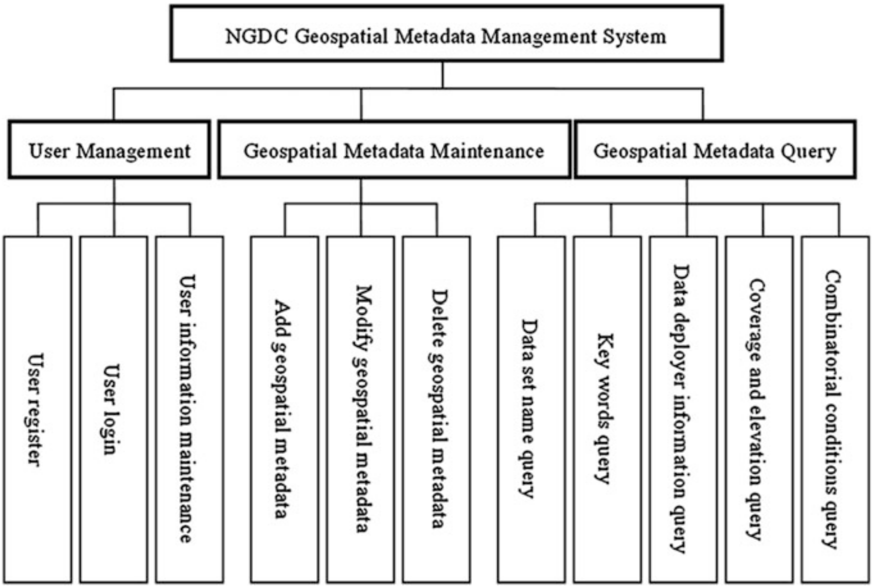
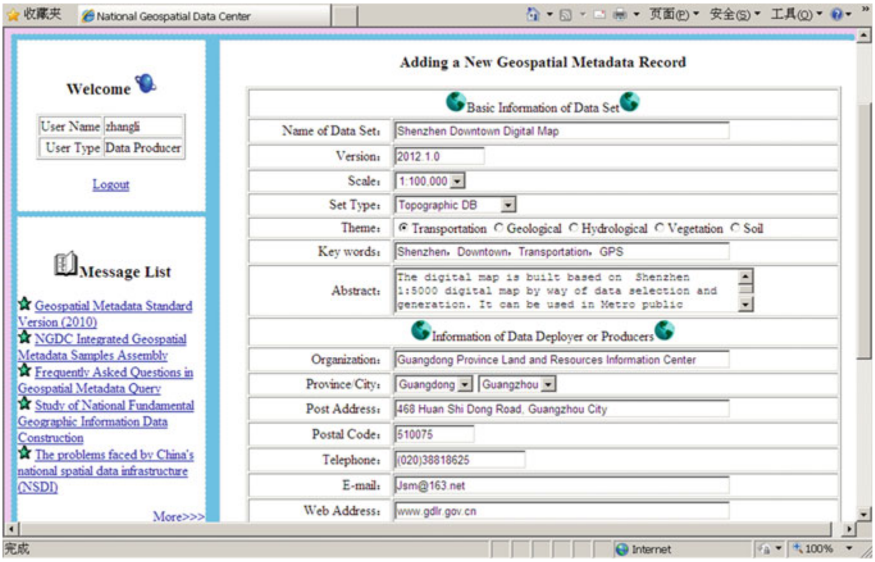

NGDC Geospatial Metadata Management System
User Management
Geospatial Metadata Maintenance
Geospatial Metadata Query
User register
User login
User information maintenance
Add geospatial metadata
Modify geospatial metadata
Delete geospatial metadata
Data set name query
Key words query
Data deployer information query
Coverage and elevation query
Combinatorial conditions query

Fig. 1 Functional block diagram of NGDC geospatial metadata management system



Welcome

User Name	zhangli
User Type	Data Producer

Logout

Message List

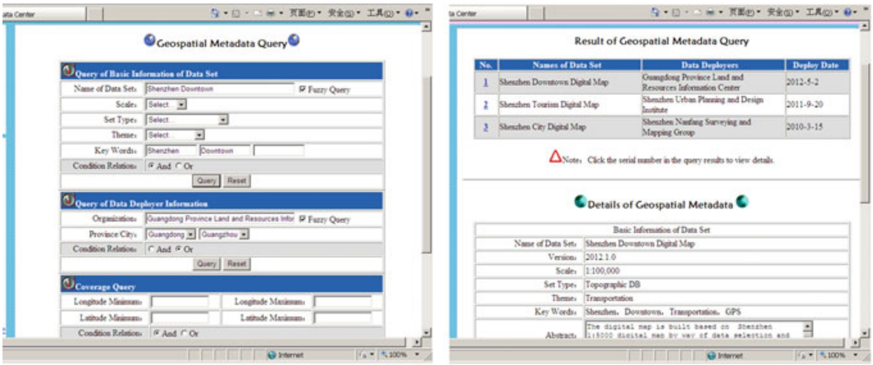
- Geospatial Metadata Standard Version (2010)
- NGDC Integrated Geospatial Metadata Samples Assembly
- Frequently Asked Questions in Geospatial Metadata Query
- Study of National Fundamental Geographic Information Data Construction
- The problems faced by China's national spatial data infrastructure (NSDI)

More>>>

Adding a New Geospatial Metadata Record

Basic Information of Data Set	
Name of Data Set:	Shenzhen Downtown Digital Map
Version:	2012.1.0
Scale:	1:100,000
Set Type:	Topographic DB
Theme:	● Transportation ○ Geological ○ Hydrological ○ Vegetation ○ Soil
Key words:	Shenzhen, Downtown, Transportation, GPS
Abstract:	The digital map is built based on Shenzhen 1:5000 digital map by way of data selection and generation. It can be used in Metro public
Information of Data Deployer or Producers	
Organization:	Guangdong Province Land and Resources Information Center
Province/City:	Guangdong Guangzhou
Post Address:	468 Huan Shi Dong Road, Guangzhou City
Postal Code:	510075
Telephone:	(020)38818625
E-mail:	Jsm@163.net
Web Address:	www.gdlr.gov.cn

Fig. 2 Web page of adding a new geospatial metadata record



Geospatial Metadata Query

Query of Basic Information of Data Set

Name of Data Set:	Shenzhen Downtown ☑ Fuzzy Query
Scale:	Select
Set Type:	Select
Theme:	Select
Key Words:	Shenzhen Downtown
Condition Relation:	● And ○ Or

Query of Data Deployer Information

Organization:	Guangdong Province Land and Resources Infor ☑ Fuzzy Query
Province City:	Guangdong Guangzhou
Condition Relation:	○ And ● Or

Coverage Query

Longitude Minimum:		Longitude Maximum:	
Latitude Minimum:		Latitude Maximum:	
Condition Relation:	● And ○ Or		

Result of Geospatial Metadata Query

No.	Names of Data Set	Data Deployers	Deploy Date
1	Shenzhen Downtown Digital Map	Guangdong Province Land and Resources Information Center	2012-5-2
2	Shenzhen Tourism Digital Map	Shenzhen Urban Planning and Design Institute	2011-9-20
3	Shenzhen City Digital Map	Shenzhen Nanfang Surveying and Mapping Group	2010-3-15

Note: Click the serial number in the query results to view details.

Details of Geospatial Metadata

Basic Information of Data Set	
Name of Data Set:	Shenzhen Downtown Digital Map
Version:	2012.1.0
Scale:	1:100,000
Set Type:	Topographic DB
Theme:	Transportation
Key Words:	Shenzhen, Downtown, Transportation, GPS
Abstract:	The digital map is built based on Shenzhen 1:5000 digital map by way of data selection and

Fig. 3 Web pages of geospatial metadata query and results

detailed geospatial metadata information can be maintained by data producers in accordance with the relevant metadata national standards. The adding Geospatial Metadata records page is shown in Fig. 2.

The functions of geospatial metadata query module include the queries of data set name, keywords, data producer information, coverage and elevation. Of course, combinatorial conditions query should be also supported. It is shown in Fig. 3.

Conclusion

In summary, as part of the Digital City Geospatial Framework the main task of NGDC is to help data users to search the data sets which meet their application needs from lots of data products. The distributed information technology based on Internet provides a strong guarantee for this task. The geospatial metadata management system is developed based on "the Integrated Geospatial Metadata" which is extracted from normal geospatial metadata. As a reference, local NGDC websites can expand the corresponding functions in this basic website according to their operation mode. For example, in order to allow data users get a direct feeling about the contents of the specified data set, local NGDC websites can provide local screenshots of digital products.

References

1. Liang J, Zhang Z, Cheng Y (2012) Research on framework data update system on public services platform. Bull Surveying Mapp 2012(12):79–83
2. Chen Z (2012) Research on urban planning information platform on geospatial framework. Bull Surveying Mapp 3:92–94

3. Xu Y, Jianbang H, WU P, Deng Y, Cao Y, MA L (2010) Research on geographic information sharing environment. Bull Surveying Mapp 6:20–22
4. Wang X, Chen H, Liu L, Li W (2012) Research and discussion on digital region geo-spatial framework construction-taking Hainan international tourism island digital geo-spatial framework for example. Bull Surveying Mapp 6:28–30
5. Li W (2011) On innovation of establishment of digital city geospatial framework. Bull Surveying Mapp 9:1–5
6. Feng X, Gu X, Fan W (2008) Discussion on the key problems of service oriented informationized mapping. Beijing Surveying Mapp 3:1–4

Apply Genetic Algorithm to Cloud Motion Wind

Jiang Han, Ling Li, Chengcheng Yang, Hui Tong, Longji Zeng and Tao Yang

Abstract Cloud Motion Wind (CMW) is a very important issue in the meteorology. In this paper, we firstly apply Genetic Algorithm (GA) to the CMW searching to reduce the computational complexity. We propose a novel CMW method, namely GA-CMW. Compared with the traditional Exhaustive CMW (E-CMW) algorithm, GA-CMW can obtain almost the same performance while with only 11 % of the computational complexity required. Generally speaking, the proposed GA-CMW method can obtain the wind vector picture in shorter time, which makes a lot of sense to the resource saving in the practical application.

Keywords Image matching · Cloud motion wind · Genetic algorithm · Cross-correlation coefficient · Computational complexity

Introduction

Nowadays, with the development of the economies, disaster warning and disaster control are becoming more and more important [1–3]. Hence, how to make the accurate and rapid forecast draws a lot of attentions to the governments.

J. Han (✉) · C. Yang
Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
e-mail: bjtu@bupt.edu.cn

L. Li
College of Mathematic and Information, China West Normal University, Sichuan, China

H. Tong
School of Science, Beijing University of Posts and Telecommunications, Beijing, China

L. Zeng · T. Yang
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_88, © Springer Science+Business Media Dordrecht 2014

Cloud Motion Wind (CMW) is a widely used method in meteorology, which aims to judge the speed and direction of the wind from the displacement of anchor cloud [1]. CMW has already been widely used in practical weather forecast and typhoon warning systems. In the traditional Exhaustive CMW (E-CMW) method, the meteorological satellite obtains the earth grey-scale maps on two different moments (commonly 30 min intervals). Then with image matching inside a specific searching window, the displacement of anchor cloud can be calculated and the speed and direction of wind is obtained.

Weather forecast and disaster warning have quite a high request to the instantaneity and computing speed. Algorithm with high speed can provide more authentic and valuable results. However, in the E-CMW algorithm, anchor cloud has to search the whole space with exhaustive method, which costs a lot of resource and time. Also, the matching algorithms are usually with high complexity. Numerous times of loops are required with the Sequence Similarity Detection Algorithm (SSDA algorithm) in [4] and Cross-correlation Coefficient (CC) method in [5]. Hence, searching for a method which can decrease the computational complexity of the traditional E-CMW algorithm is quite beneficial in both meteorology forecast and resource saving.

Genetic Algorithm (GA) is a classical intelligent algorithm to obtain the optimal or nearly optimal value, which has been widely used in mathematics [6], artificial intellegence [7, 8] and telecommunication engineering [9–11]. It is shown in [11] that with GA, a large percentage of computational complexity can be reduced. In this paper, we firstly apply GA to CMW searching and provide the details of the proposed GA-CMW algorithm. Simulations of both GA converging curve and the wind vector figure in a local area show that with little performance lost, only 11 % of the complexity is needed with the proposed GA-CMW.

The paper is organized as follows. Section Traditional E-CMW Algorithm describes the traditional E-CMW algorithm. Section GA based CMW Searching provides the details of the proposed GA-CMW algorithm, including the initialization of GA. Numerical analysis and simulation results on the performance and the computational complexity are shown in section Analysis and Numerical Simulation Results. Finally, conclusion is drawn in section Conclusion.

Traditional E-CMW Algorithm

The Principle of E-CMW

(1) *Model of E-CMW*: E-CWM is a method in meteorology which describes the parameters of wind vector, including its speed and direction. As the movement or displacement of a cloud depends on the wind around it, the parameters of wind vectors can be achieved by tracing the anchor cloud. As shown in Fig. 1, cloud *M* on time T is the anchor cloud, which is surrounded by small window

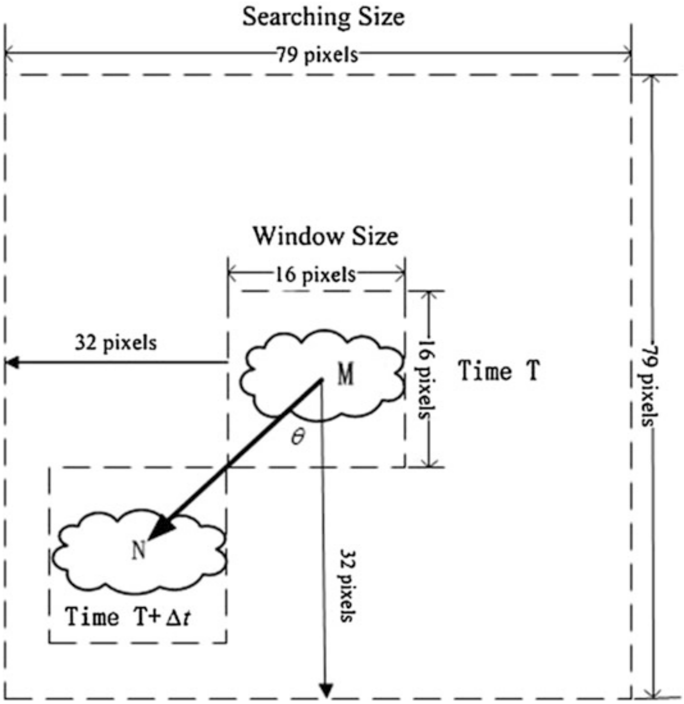


Fig. 1 The model of CMW matching process

> with 16 × 16 pixels. After Δ*t* interval, the anchor cloud moves to the new position *N*, by calculating the angle *θ* and distance *MN* of the movement, parameters of the wind vector are obtained.

Let's assume the coordinate of the anchor cloud on time T is $M(x_M, y_M)$, while its coordinate on time $T + \Delta t$ is $N(x_M, y_M)$. As shown in Fig. 1, considering *MN* is just a short distance on the sphere surface, vector $\overrightarrow{MN}$ is approximately equivalent to the arc $\overset{\frown}{MN}$. Then we can obtain the angle and speed of the movement as following:

$$\vec{v} = \frac{\sqrt{(y_N - y_M)^2 + (x_N - x_M)^2}}{\Delta t} \tag{1}$$

$$\cos\theta = \frac{|(y_N - y_M)|}{\sqrt{(y_N - y_M)^2 + (x_N - x_M)^2}} \tag{2}$$

(2) *Window Size and Searching Size*: The choice of window size and searching size can directly influence the matching performance and the searching complexity. In the E-CMW method, proper window size should cover most parts of the anchor cloud and the proper searching size should cover the refreshed position on $T + \Delta t$ moment.

With traditional E-CMW searching method, the searching times needed are:

$$\gamma_{CMW} = (S_{size} - W_{size} + 1)^2 \tag{3}$$

where S_{size} denotes the searching size and W_{size} denotes the window size.

In this paper, we apply the window size with 16 pixels and searching size with 79 pixels. Hence, with formula (3), 64 × 64 times of searching are totally required with the traditional E-CMW method. Further more, we will get numerous anchor clouds on the whole satellite cloud picture. Hence, the computational complexity will increase tremendously with the increasement of the predicted region.

Matching Algorithm

Above all, we have illustrated the principle of the traditional E-CMW algorithm. In this subsection, the commonly used matching algorithm CC is shown in (4).

$$CC(i,j,a,b) = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B}((S^{i,j}(a,b) - \bar{S}^{i,j}) \times (T(a,b) - \bar{T}))}{\sqrt{\sum_{a=1}^{A}\sum_{b=1}^{B}(S^{i,j}(a,b) - \overline{S}^{i,j})^2 \times \sum_{a=1}^{A}\sum_{b=1}^{B}(T(a,b) - \overline{T})^2}} \tag{4}$$

where *T(a,b)* denotes the original template, $S^{(i,j)}(a,b)$ denotes the matching template. *i* and *j* in (4) denote the coordinate of the matching template on $T + \Delta t$ moment, then we have $i \le 64$, $j \le 64$ in the whole searching space. A and B denote the template size, which are all equal to 16 in the paper. $\overline{T}$ and $\overline{S}^{i,j}$ denote the average value of the template T and S.

GA Based CMW Searching

Basic GA Based CMW Searching

In this subsection, details of the proposed GA-CMW are illustrated and some of the GA's initialization parameters are introduced.

(1) *Coding Scheme*: In the E-CMW searching, searching range is a window with $S_{size} \times S_{size}$ pixels. With exhaustive method in this paper, γ_{CMW} times of matching, which has been defined in (3), are needed with exhaustive method.

As CMW is a discrete searching problem, it is quite proper to apply the Gray Coding Scheme. Gray Coding has a quite good characteristic in suppressing the error propagation. With S_{size} equals to 79, W_{size} equals to 16, 64 × 64 times of searching are required. Hence, we can encode the coordinate of each template in the searching space with 8 bits.

(2) *Fitness Evaluation*: Each individual of the generation in GA has a fitness to indicate its weight, and the cost function varies from different problems. In the proposed GA-CMW method, the cost function is used to indicate the matching degree between the original template and the matching template, which can still be the CC matching formula in (4).

Individual with bigger value has the better fitness in GA-CMW matching.

(3) *Selection*: In the natural evolution, individuals die or live with natural selection, which has been indicated in details in the well-known Darwin's theory of evolution. In the GA algorithm, it also has the selection process to choose out those individuals with better fitness. With N_G nodes each of the generation, N_S nodes will survive through the selection, which as follows:

$$N_G = N_S \times P_S \tag{5}$$

where P_S is the selection probability in the algorithm. After the selection process, those $N_S - N_G$ nodes with the bottom fitness will be weeded out to release the space for new bloods.

(4) *Crossover*: The process crossover is used to produce the new individuals from those who have survived from the selection. By exchanging the genes randomly between two individuals, new individual is generated from its parents. P_C is used to indicate the probability of crossover in the N_S nodes and is often set to a high value.

(5) *Mutation*: Mutation is another process to produce the new bloods. Mutation also happens in the nature with a random change to the individual's chromosome which can thus generate unpredicted effect on the final results. In some situations, mutation can work to help the current generation jump out the local optimal areas. P_M is used to indicate the probability of mutation in the N_S nodes and is often close to 0.

(6) *Decision Making*: With the continuous iteration, new individuals keep on replacing those members with worse fitness in the previous generation. The iteration ends with the predefined iteration times or the optimization of objective function is achieved. Then node with the best fitness in previous generation is output to be the final decision.

Analysis and Numerical Simulation Results

In this section, we compare the performance and computational complexity between the traditional E-CMW method and GA-CMW method.

Performance Evaluation

Figure 2 shows the output matching value versus iteration time with different population sizes. The optimal matching value boundry with E-CMW method in Fig. 2 is 0.8085, and we assume the output value above 0.80 as converged. It is shown in Fig. 2 that with population size of 80, the iteration time to converged is 6. For population size of 40, 13 times of iteration are required. While with 20 individuals per generation, it takes 48 times to converged. In the simulation results shown in Fig. 2, if the iteration takes long enough time, the output matching value will anyway converged to the optimal matching value boundry, which means should be no performance lost between GA-CMW searching and the traditional E-CMW algorithm.

Figure 3 shows the traditional E-CMW searching simulation result. Figure 4 shows the simulation result of the proposed GA-CMW algorithm with predefined iteration times.

For E-CMW results shown in Fig. 3, The matching results are in a quite good order. For GA-CMW results shown in Fig. 4, it is shown that GA-CMW results may have several errors because of the suboptimal output of the GA searching. The yellow ellipses marked in Fig. 4 show the bad points of the GA-CMW results. But considering the overall trend of the wind, it is almost the same between Figs. 3 and 4. Hence, the performance of GA-CMW should anyway be accepted.

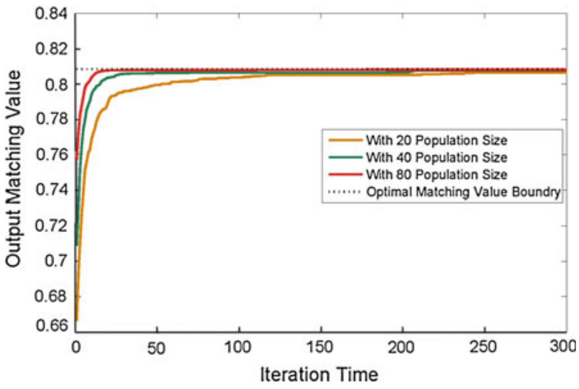


Fig. 2 Iteration time versus output matching value with different population size

Fig. 3 Matching result with traditional E-CMW algorithm

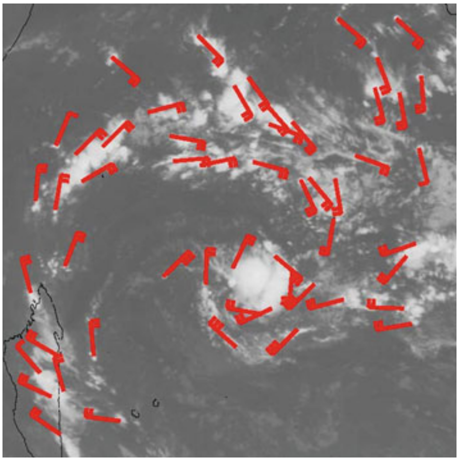
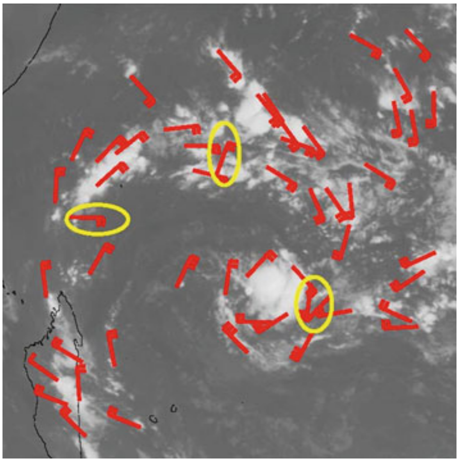


Fig. 4 Matching result with GA-CMW algorithm

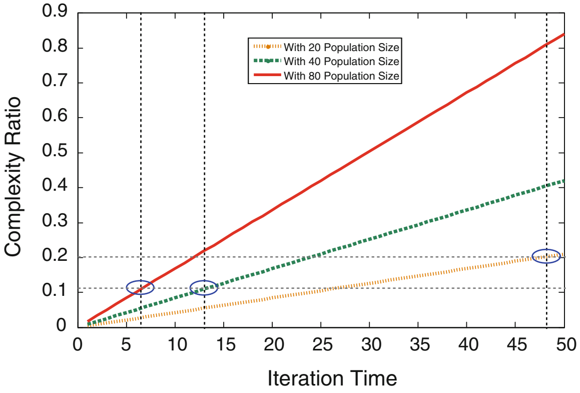


Complexity Evaluation

It is shown in Fig. 2 that with specific iteration time, the proposed GA-CMW has almost no performance lost compared with the traditional E-CMW method. In this subsection, we evaluate the computational complexity of the proposed algorithm. As the computational complexity cost by the selection, crossover, and mutation process is quite small compared with the fitness calculation. We use the running times of the matching function shown in (4) to indicate the computational complexity. Formula (3) and (6) indicate the complexity of traditional E-CMW and GA-CMW.

$$\gamma_{GA-CMW} = N_G \times Iter\_time \tag{6}$$

Fig. 5 Complexity comparison between GA-CMW and E-CMW



We define the complexity ratio as following:

$$\beta = \frac{\gamma_{GA-CMW}}{\gamma_{E-CMW}} \tag{7}$$

Figure 5 shows the complexity ratio between GA-CMW and E-CMW. It is shown that with the iteration time whose output value firstly beyond 0.80 as the converged time, only 11 % of the complexity is needed with GA-CMW. Even for only 20 individuals each generation, 16 % of the complexity is needed.

For the weather forecast, the satellite data is changing all the time and the CMW result needs to be refreshed quite often. Thus the decrease of the calculating time of CMW does making a lot of sense. By applying GA to the CMW searching, very small percentage of the computational resource will be taken compared with the traditional E-CMW. GA-CMW can save the computational resource and take quite less time to obtain the final CMW result.

Conclusion

In this paper, we firstly apply the GA algorithm to the CMW searching, which is a very important issue in the meteorology. We propose GA-CMW method to reduce the searching complexity in traditional E-CMW method. Simulation results show that with only 11 % of the computational complexity needed, there is almost no performance lost between GA-CMW and E-CMW. Generally speaking, the proposed GA-CMW method can obtain the wind vector picture with shorter time and quite acceptable performance, which makes a lot of sense to the resource saving and disaster control in the practical application.

Acknowledgments This paper is supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grand 2012ZX03001039-002, which is kindly acknowledged.

References

1. Purdom JF (1996) W. Detailed cloud motions from satellite imagery taken at thirty second one and three minute intervals. In: Proceeding to the 3rd international wind workshop in Ascona, Switzerland, 10–12 June 1996, pp 137–146
2. Wang ZH, Browning KA, Kelly GA (1997) Verification of the tracking technique used in an experimental cloud motion wind inferring system. JCMM Report. University of Reading, 1997
3. Wang Z, Zhou J (2000) A preliminary study of Fourier series analysis for cloud tracking with GOES high temporal resolution images. Acta Meteoro Sin 14(1):82–94
4. Leese JA, Novak CS, Clark BB (1972) An automated technique for obtaining cloud motion from geosynchronous satellite data using cross correlation. J Appl Meteor 10(1):118–132
5. Jianmin X, Qisong Z (1996) Calculation of cloud motion wind with GMS-5 images in China. In: Proceedings to the 3rd International Wind Workshop in Ascona Switzerland, pp 45–52, 10–12 June 1996
6. Revello TE, McCartney R (2002) Generating war game strategies using a genetic algorithm. In: Proceeding Congress Evolutionary Computation, 2002, vol. 2, pp 1086–1091
7. Campbell MS, Hoane AJ, Hus FH (2002) Deep blue. Artif Intell 134(1–2):57–83
8. Shibata T, Fukuda T, Tanie K (1997) Chapter 108: Synthesis of fuzzy, artificial intelligence, neural networks, and genetic algorithm for hierarchical intelligent control. CRC Press, Boca Raton, pp 1364–1368
9. Binelo MO, de Almeida ALF, Cavalcanti FRP (2011) MIMO array capacity optimization using a genetic algorithm. IEEE Trans Veh Technol 60(6):2471–2481
10. Mangoud MAA (2009) Optimization of channel capacity for indoor MIMO systems using genetic algorithm. Prog Electromagn Res C 7:137–150
11. Bashir S, Khan AA, Naeem M, Shah SI (2007) An application of GA for symbol detection in MIMO communication systems. In: Third International Conference on Natural Computation, ICNC 2007, Aug

A Topic Evolution Model Based on Microblog Network

Qingling Zhou, Genying Wang and Haiqiang Chen

Abstract Fission mode of information transmission in the Microblog network presents a challenge to the existing public opinion diffusion model. In order to depict the mechanism of the public opinion transmission in the Microblog network, this paper proposes a topic evolution model based on the complex networks and infectious disease dynamics theory. The model takes the directed scale-free network as the carrier, whi ch synthetically considers the characteristics of topic propagation and its influence factors. In addition, the research performs simulation analysis, of which result mirrors the topic evolution process from common topic into hot topic, and verifies its validity.

Keywords Microblog network · Topic evolution · Hot topic

Introduction

Microblog, namely micro blog, is a form of blog allowing users to update brief texts (usually less than 200 words) and can be published openly, which mainly provides posting, forwarding, comments, attention, etc. Fission mode of information transmission in the Microblog network presents a challenge to the existing

Q. Zhou · G. Wang
School of Electronic and Information Engineerring, Beijing Jiaotong University, Beijing 100044, China
e-mail: 11120232@bjtu.edu.cn

Q. Zhou · G. Wang (✉)
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China
e-mail: gywang@bjtu.edu.cn

H. Chen
China Information Technology Security Evaluation Center, Beijing 100085, China
e-mail: chenhq@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_89, © Springer Science+Business Media Dordrecht 2014

public opinion diffusion model. And Microblog walks into every aspect of people's life in form of "fragmentation" information, so that it gradually becomes the important position in public opinion. So, studying topic dissemination, modeling its evolution process and correspondingly making the quantitative analysis, have an important guiding significance for discovering and controlling the vital public sentiment.

Presently, most related works study information dissemination only from one perspective on interpersonal network, such as, information dissemination in the network topology [1], individual interaction rules [2], degree assortativity characteristic of collaborative network [3, 4], etc. But, actually, information dissemination and formation are a typical evolution process in the complex system, which needs synthetically consider individual interaction and network structure so as to describe topic evolution more accurately in the Microblog network. In this paper, we preliminarily establish the Microblog evolution simulation model based on topic spreading rules and network topology's influence on the spreading behavior. Above all, the result reflects the real situations, and verifies its validity.

Model

Evolution Mechanism

In the Microblog network, once one topic is posted, promptly, it will be discovered by bloggers and they will share, spread, or comment on it with corresponding probability. Meanwhile, if the user's friends found this topic, they would also determine whether to transmit or forward it depending on personal preferences. In this way, the topic is forwarded and propagated along with friend relationship, finally, it is judged whether the common topic evolves into a hot topic by the amount of forwarding and comments.

Based on topic transmission characteristics and whether it's a hot topic, we put the topics into two categories: common topic and hot topic. Common topic is a general topic, presently, on which attention is low, but it may become a hot topic over time. Hot topic is a topic that's paid attention quite high within a certain period of time, and also it's recovering into a common state at a certain speed. Now, we make the following sets:

(1) Common topic evolves into hot topic with the probability $\lambda 1$;
(2) Hot topic renews into a common topic with the speed v.

Topic Evolution Model Based on Microblog Network

For common topic, namely c (t); hot topic, namely h (t). Therefore, having the following topic evolution modes:

$$\begin{aligned}&(1)\,\mathrm{c}\ (t)\ \rightarrow\ \mathrm{h}\,(t);\\&(2)\,\mathrm{h}\,(t)\ \rightarrow\ \mathrm{c}\,(t).\end{aligned}\tag{1}$$

For the above formula, c (*t*) evolves into h (*t*) with the probability *λ1*. Meanwhile, h (*t*) renews into a common topic with the speed *v* over time. Then, we can obtain our model, namely common-hot-common (CHC) model, which is similar to the susceptible-infected-susceptible (SIS) model [5].

Microblog network is a directed scale-free network. Based on users' attention directions, the users' node degree can be divided into two categories, namely in-degree and out-degree [6]. As the amount of node in-degree is larger, users have the greater influence on topic evolution. Meanwhile, as the amount of node out-degree is larger, the information source is richer. This paper studies the propagation properties of user posting topics, so we only consider the situations of users' node in-degree.

Suppose one user with in-degree *k* posts a topic, which is in common state at time *t*. During the period of [*t*, *t* + *Δt*], $\mathrm{p_{cc}}$ denotes the probability of the topic remaining the same, and $\mathrm{p_{ch}}$ represents the probability of the topic evolution from common topic into hot topic, and $\mathrm{p_{ch}} = 1-\mathrm{p_{cc}}$. Where,

$$\mathrm{p_{cc}} = (1 - \Delta t\lambda 1)^{m},\tag{2}$$

Where, *m* = *m (t)* represents the topic forwarding amount. In addition, *m* is a random variable with the binomial distribution, as follow:

$$\mathrm{S}(m,\,t) = \binom{k}{m}[\alpha\beta \mathrm{q}(t)]^{m}\,[1 - \alpha\beta\mathrm{q}(t)]^{k-m},\tag{3}$$

Where, q(*t*) denotes the probability user's friends discover the topic at time *t*, α represents the users' interest factor [7] in the topic, and *β* is the sensitive topic factor. Q(*t*) is a function of time, obviously, of which process yields Poisson distribution, as below shown:

$$\mathrm{q}(\mathrm{T} = t) = \lambda^{t}\mathrm{e}^{-\lambda}/t!,\tag{4}$$

So, we can get user's average transition probability $\mathrm{p'_{cc}}$ with in-degree *k* in [*t*, *t* + Δ*t*], as follow:

$$\begin{aligned}\mathrm{p'cc} &= \sum\nolimits_{\mathrm{m=0}}^{\mathrm{k}}\binom{k}{m}(1 - \Delta t\lambda 1)^{\mathrm{m}}[\alpha\beta\mathrm{q(t)}]^{\mathrm{m}}[1 - \alpha\beta\mathrm{q(t)}]^{\mathrm{k-m}}\\&= [1 - \Delta t\,\lambda 1\alpha\beta\mathrm{q}(t)]^{k}\end{aligned},\tag{5}$$

This, yields the solution of (4) to (5),

$$\mathrm{p'_{cc}} = [1 - \Delta t\lambda 1\alpha\beta(\lambda^{t}\mathrm{e}^{-\lambda}/t!)]^{k},\tag{6}$$

Assuming N(*t*), C(*t*), H(*t*) denotes the amount of the total topic, common topic, hot topic, respectively, at time *t*. Then, we can obtain:

$$\mathrm{C}(t) + \mathrm{H}(t) = \mathrm{N}(t), \tag{7}$$

So, we can get the changes in the number of common topic in [t, $t + \Delta t$]. As below:

$$\begin{aligned} \mathrm{C}(t + \Delta t) &= \mathrm{C}(t) - \mathrm{C}(t)(1 - \mathrm{p}'_{\mathrm{cc}}) + \mathrm{H}(t) v \Delta t \\ &= \mathrm{C}(t) - \mathrm{C}(t)[1 - (1 - \Delta t \lambda 1 \alpha \beta (\lambda^{t} \mathrm{e}^{-\lambda}/t!))^{k}] + \mathrm{H}(t) v \Delta t, \end{aligned} \tag{8}$$

Similarly, we can get the changes in hot topic‘s amount in [t, $t + \Delta t$]. That is:

$$\begin{aligned} \mathrm{H}(t + \Delta t) &= \mathrm{H}(t) + \mathrm{C}(t)(1 - \mathrm{p}'_{\mathrm{cc}}) - \mathrm{H}(t) v \Delta t \\ &= \mathrm{H}(t) + \mathrm{C}(t)[1 - (1 - \Delta t \lambda 1 \alpha \beta (\lambda^{t} \mathrm{e}^{-\lambda}/t!))^{k}] - \mathrm{H}(t) \mathrm{v} \Delta \mathrm{t}, \end{aligned} \tag{9}$$

Yielding the solution of (7) to (8), then we can get:

$$\begin{aligned} \frac{\mathrm{C(t + \Delta t) - C(t)}}{\mathrm{N(t)\Delta t}} = &- \frac{\mathrm{C(t)}}{\mathrm{N(t)\Delta t}} \left[1 - (1 - \Delta \mathrm{t} \lambda 1 \alpha \beta (\lambda^{\mathrm{t}} \mathrm{e}^{-\lambda}/\mathrm{t}!))^{\mathrm{k}}\right] \\ &+ \frac{\mathrm{H(t)}}{\mathrm{N(t)\Delta t}} \mathrm{v} \Delta \mathrm{t} \end{aligned}, \tag{10}$$

When Δt converges to zero, we carry on the Taylor expansion for (9) at the right side. Then,

$$\frac{\mathrm{d}\rho_c(t)}{\mathrm{dt}} = -\mathrm{k}\lambda 1 \alpha \beta \rho_c(t) \left(\lambda^{\mathrm{t}} \mathrm{e}^{-\lambda}/\mathrm{t}!\right) + \rho_h(t) v, \tag{11}$$

Where, $\rho_{\mathrm{c}}(t)$ represents the density of common topic in the Microblog network at time t.

Similarly, from (9), we can obtain:

$$\frac{\mathrm{d}\rho_h(t)}{\mathrm{dt}} = \mathrm{k}\lambda 1 \alpha \beta \rho_c(t) \left(\lambda^{\mathrm{t}} \mathrm{e}^{-\lambda}/\mathrm{t}!\right) - \rho_h(t) \mathrm{v}, \tag{12}$$

Here, $\rho_{\mathrm{h}}(t)$ denotes the density of hot topic in the Microblog network at time t.

From Eqs. (11), (12), we can get the simultaneous equations of the dynamics of topic evolution, which is used to describe the changes of topic density with time. Meantime, topic evolution process is affected by users' node in-degree, sensitive topic factor and so on.

The Simulation and Results

Data Source

We use automatic search procedure to obtain a directed subnet of Sina Microblog, namely G = (V, E), where, node v ∊ V denotes one Microblog ID, directed edge (u, v) ∈ E represents user u is v's fan, namely u → v. Obtained subnet has the following features: (1) total nodes n = 59986; (2) average in-degree din = 312; (3) average out-degree dout = 208. As Fig. 1, of which shows the accumulated distribution of in-degree.

From Fig. 1, we know the distribution of users' node in-degree exhibits a power-law behavior in the Microblog network. Most of users' node in-degree is low, namely grass-root node, but only a few users' nodes have a larger in-degree, namely famous bloggers in the Microblog network. So the model constructs the microblog network structure consistently with the characteristics of real network. Based on this, we construct a topic evolution model, and perform the quantitative analysis.

Topic Density Distribution

1. At initial time, we assume that $\rho_c(0) = 1$, $\rho_h(0) = 0$. Then, we give the model parameters, as follows: $\lambda 1 = 1/20$, $\alpha = 1/10$, $\beta = 1/4$, v = 0.009, $\lambda = 6$. And the average in-degree din = 312, we can get the changes of $\rho_c(t)$, $\rho_h(t)$ with time, respectively. As the following Fig. 2:

From Fig. 2, during one week, we can see that $\rho_c(t)$, $\rho_h(t)$ is changing with time *t*. Once one user posts a topic, the amount of topic's forwarding or comments increases almost linearly during [0, 12 h]. Finally, common topic may evolve into a hot topic. And during this period, $\rho_h(t)$ grows with the approximate linearity, ultimately, $\rho_h(t)$ comes to its peak at some time point. Then, attention on this topic tends to go down. About one week later, generally, hot topic recovers back to the common situation again. At the same time, during this period, $\rho_h(t)$ is decreasing continually, and finally it tends to zero. This result is consistent to the characteristics of public opinion evolution we observed (As what $\rho_c(t)$-real, $\rho_h(t)$-real show in Fig. 2, which we get by the means of statistical average), when I worked

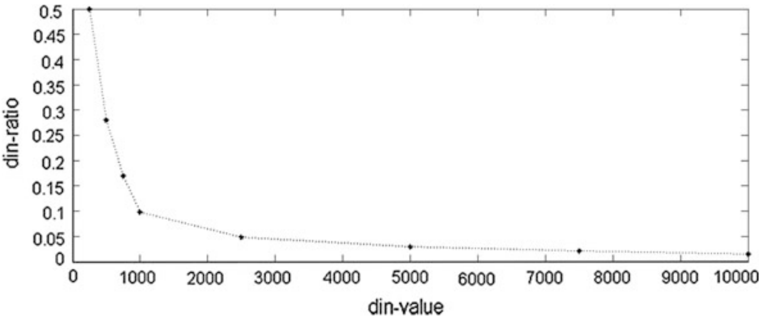


Fig. 1 Din-value, din-ratio represents the value of in-degree and its corresponding ratio, respectively

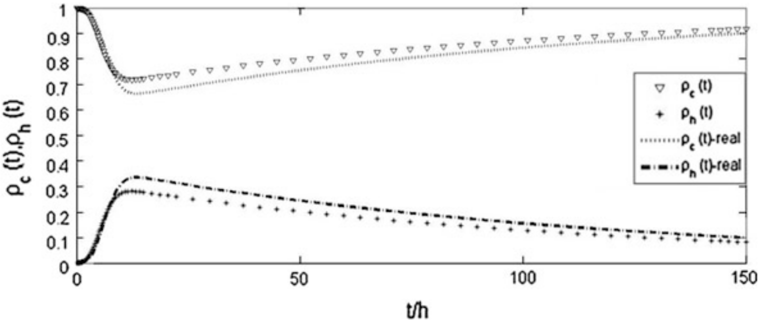


Fig. 2 $\rho_c(t)$, $\rho_h(t)$ changes along with time t, respectively

as an intern in a confidential company for monitoring the Microblog public opinion.

2. Under the condition (1), we discuss the influence on $\rho_h(t)$ for different β. Here, β takes 1/2, 1/8, 1/20, 1/50, respectively. Then we can get Fig. 3. As below:

By Fig. 3, with the increase of sensitive topic factor, topics are easily evolving from common topic into hot topic in a very short time, and then, this topic can be still discussed by everyone in a subsequent week, of which heat plays out slowly. As we know, the simulation results are accord with the real situations.

3. Under the condition (1), we discuss the influence on $\rho_h(t)$ for different k. Here, k takes 50, 300, 2000, 6000, respectively. Then we can obtain Fig. 4. As follow:

From Fig. 4, with the increase of the users' node in-degree, if one user posts a topic, this topic easily evolves from common topic into a hot topic in an extremely short time. What's more, if the users' node in-degree is large enough, the common topic will definitely turn into a hot topic, so we even can ignore the other factors. The results reflect great influence of famous bloggers on topic evolution, playing

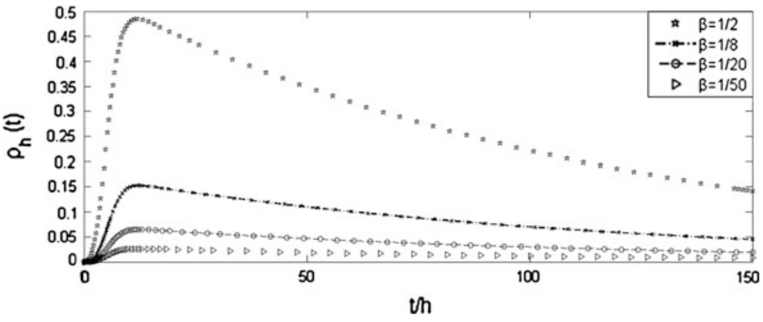


Fig. 3 When β takes different values, $\rho_h(t)$ changes with time t

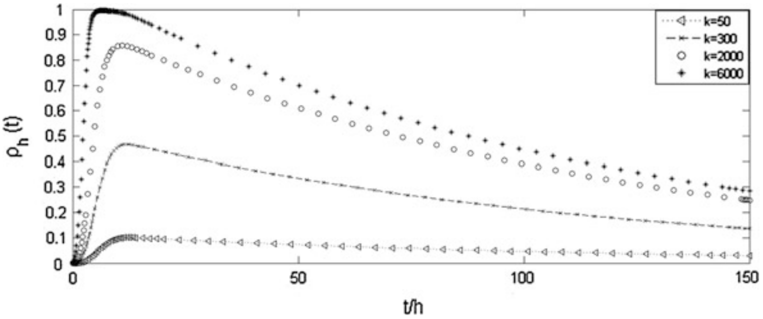


Fig. 4 When k takes different values, $\rho_{\mathrm{h}}(t)$ changes with time t

the roles of opinion leaders [8]. And also it can create conditions for "water army" spreading rumors. Even one topic is not a real event, after several paid posters with large in-degree jacking up, it may also completely be hyped into a real hot one.

Conclusion

This paper studies the characteristics of topic propagation in the microblog network, and proposes a topic evolution model, namely CHC model. The model combines complex network with infectious disease dynamics theory, further to establish the differential equations, which can better describe the topic evolution characteristics. The simulation results show that: all the topic evolution procedure is transforming from common topic into hot topic, and then the hot topic slowly recovers to the common state. In addition, the more sensitive the topic, the more easily common topic evolving into hot topic in extremely short time, and the topic heat keeps high for quite so long time. The larger the user's node in-degree, the more easily common topic evolving into hot topic to play the roles of opinion leaders. Meantime, it can create conditions for spreading rumors in the microblog network. Even one topic is not a real event, after several "water army" with large in-degree jacking up, it also may completely be hyped into a real hot topic. This study will help us understand the topic propagation characteristics deeply, and be meaningful to discover and conduct the crucial public sentiment.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Jaewon Y, Leskovec J (2010) Modeling information diffusion in implicit networks. IEEE 10th international conference on data mining. Berlin, Germany, IEEE, 2010, pp 599–608
2. Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in twitter: the million follower fallacy. In: Proceedings of international AAAI conference on weblogs and social. Washington, DC, ICWSM, 2010, pp 97–105
3. Hu H, Wang X (2009) Evolution of a large online social network. Phys Lett A 373:1105–1110
4. Alojairi A, Safayeni F (2012) The dynamics of inter-node coordination in social networks: a theoretical perspective and empirical evidence. Int J Project Manage 30:15–26
5. Zhou J, Liu Z (2008) Epidemic spreading in complex networks. Front Phys China 3(3):331–348
6. Liu Z, Lai Y (2002) Connectivity distribution and attack tolerance of general networks with both preferential and random attachments. Phys Lett A 303:337–344
7. Yan Q, Yi L (2012) Human dynamic model co-driven by interest and social identity in the Microblog community. Phys A 391:1540–1545
8. Iñiguez G et al (2011) Modelling opinion formation driven communities in social networks. Comput Phys Commun 182:1866–1869

A Platform for Massive Railway Information Data Storage

Xu Shan, Genying Wang and Lin Liu

Abstract With the development of national large-scale railway construction, massive railway information data emerge rapidly, and then how to store and manage these data effectively becomes very significant. This paper puts forward a method based on distributed computing technology to store and manage massive railway information data, builds massive railway information data storage platform by using the Linux cluster technology. This system consists of three levels including data access layer, data management layer, application interface layer, enjoying safety and reliability, low operation cost, fast processing speed, easy expansibility characteristics, which shall satisfy the massive railway information data storage requirement.

Keywords Massive railway information data storage · Hadoop distributed technology · Cluster system

X. Shan · G. Wang
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: 11120075@bjtu.edu.cn

X. Shan · G. Wang (✉)
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China
e-mail: gywang@bjtu.edu.cn

L. Liu
China Information Technology Security Evaluation Center, Beijing, China
e-mail: liul@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_90, © Springer Science+Business Media Dordrecht 2014

Introduction

The Medium and Long-Term Railway Network Plan has established the goal to finish railway construction in 2020, when the high standard and large-scale railway construction will appear in full swing and will generate huge amounts of railway information data. These data, being vast and complex, diverse, heterogeneous, dynamic, relate to various aspects such as railway geographic information, railway construction, railway operation and maintenance, and railway dispatching. However, the current situation is lack of the unified collection and storage criterion or standard, leading to the data island phenomenon. How to store and manage massive railway information data and how to make more efficient use of these data, become one of the key even the bottleneck projects in railway department, and that's what this paper is about.

Traditional methods to deal with massive data mostly use distributed high performance computing and grid computing technology [1], which consume expensive computing resources, need tedious programming to realize effective segmentation of massive data and reasonable distribution of computing tasks. Fortunately, the new development of Hadoop distributed technology can solve these problems better [2].

Basing on Linux cluster technology and using the Hadoop distributed technology, this platform effectively processes the massive amounts of railway information data and stores them in the distributed database, which designs and implements an easily extended and effective massive data storage management system. In the section Platform Architecture of this article, the author gives an introduction to the massive data storage platform architecture. The section The Key Technologies of the Platform Development focuses on the key technologies of system implementation needed. The section The Platform Test and Result Analysis illustrates the implementation and performance characteristics of the system. Finally, the section Conclusion presents the conclusions.

Platform Architecture

Platform Overall Framework

According to the actual needs, we employ the MVC three-tier framework for system design. The platform is divided into data resource layer, business logic layer and presentation layer. Each layer has a clear division of responsibilities, high cohesion and low coupling, making the structure more clear and easier to maintain. The platform overall Framework is shown in Fig. 1.

Data resource layer, composed of several storage nodes, storing and managing massive railway information data, is the foundation of the whole platform.

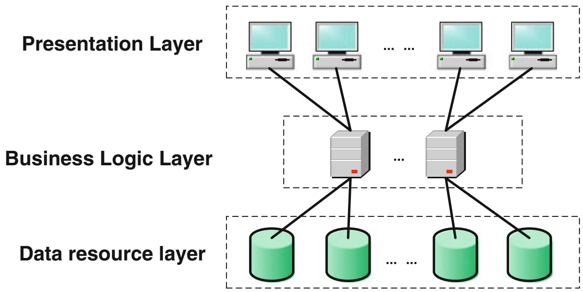


Fig. 1 Platform overall framework

Business logic layer, provides the parallel loading storage of the massive railway information data and the management and support services to ensure the normal operation of the system, is the core of the whole platform.

Presentation layer, provide users with user-friendly interface, convenient for users to query the railway information data and extend system, which meets users directly.

Network Topology of the Platform

The platform adopts the distributed, hierarchical structure in the design, which stores the resources in multiple nodes in different parts of the cluster server. The main service node schedules and manages the distributed server cluster nodes in an unified way. With the data volume increasing and the complex application requirements changing, this platform can be easily extended, and the existing relational database can also be integrated into the platform [3]. And through the de-isomerization process, the platform and the existing relational database can jointly provide storage services for users, which serves the users transparently massive railway information data by storage and management functions (Fig. 2).

Overall Functional Design of the Platform

According to the system function, the system can be divided into three layers, including data access layer, data management layer and application interface layer, as is shown in Fig. 3.

Data Access Layer, with the support of the upper data management, connects the storage nodes distributed in different parts through the local area network (LAN) and the Internet to form a distributed cluster system, which provides shielding for different sources of various databases, and provides database access by service function, only to provide a convenient management and deployment chance.

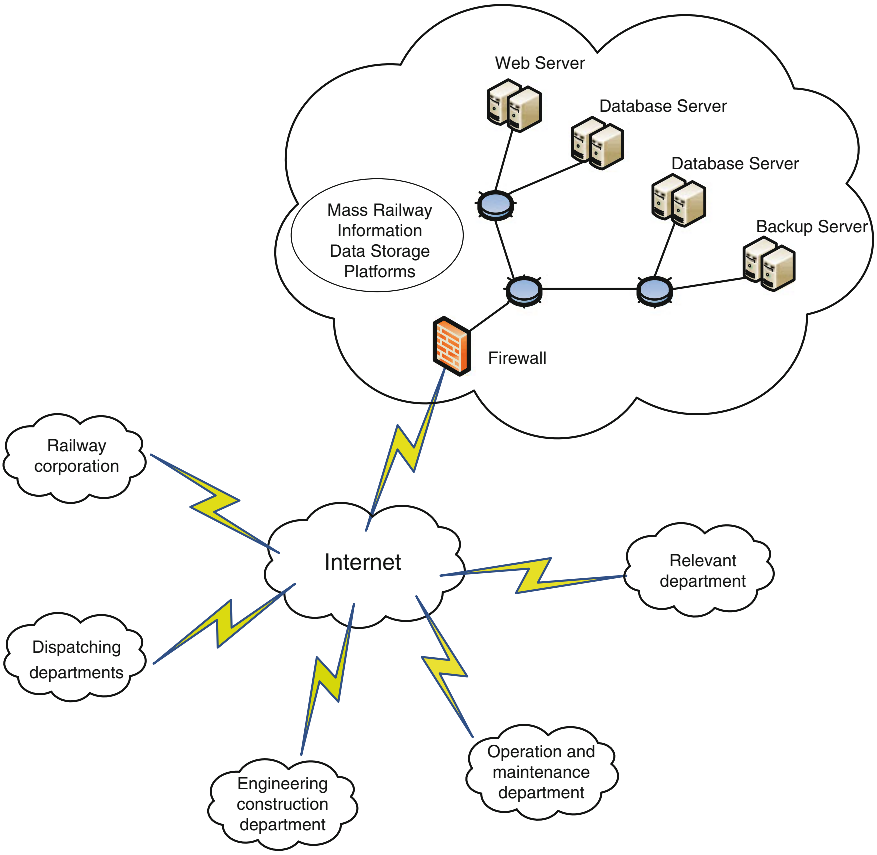


Fig. 2 Network topology of the platform

Data Management Layer, after the huge amounts of data parallel processing, stores the processed data in the distributed database of this system, while it provides management support service to guarantee the system run normally. It is formed of six function modules, including system management module, log management module, parallel load module, load balancing module, storage module, parallel query module, backup recovery module. System management module is used for software management, including running state monitoring, remote system deployment and autonomous running and maintenance, etc. Log management module is used for software operation log management, including the system trajectories, key events and state records, etc. Load balancing serves load balancing and fault tolerance of management of storage node. Parallel loading storage module provides parallel loading and storage for huge amounts of data. Parallel query module provides parallel query, user-defined function such as transaction processing. Backup recovery module provides system stored data backup management, backup storage, backup restore, etc. [3].

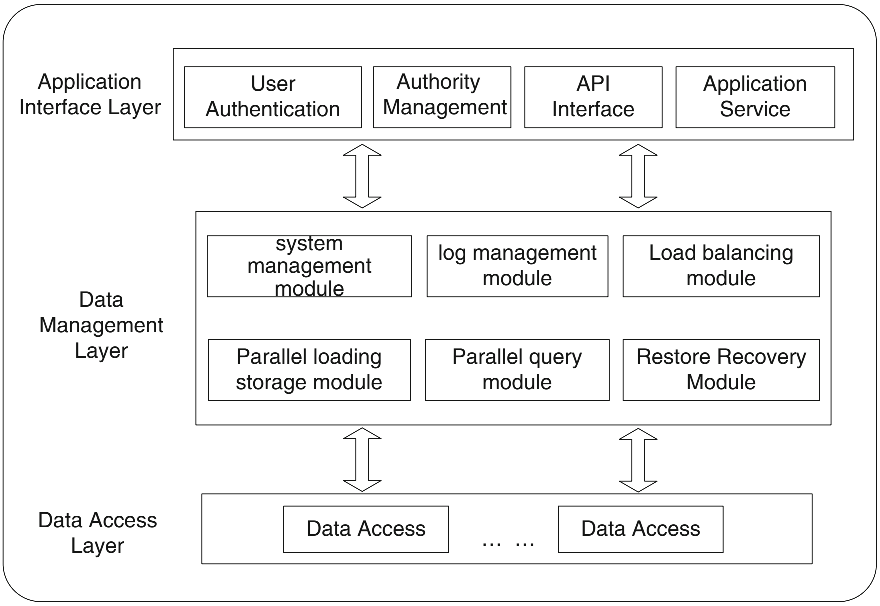


Fig. 3 Overall functional design of the platform

Application Interface Layer, according to the actual business types, provides different application service interface, and provides various application service on the basis of user authentication to meet the needs of different users.

The Key Technologies of the Platform Development

According to the front introduction of Hadoop and functional design of the platform, the most important part is the data management layer. When to manage it, parallel loading storage module of the data become the core of the entire platform. The Hadoop distributed technology provides models and methods of data storage and data processing for the platform. We use the Hadoop distributed file system to store a huge number of source data, and use MapReduce distributed computing model to deal with these data, then use HBase distributed database to store the processed data, in order to realize the storage and management of massive railway information data. Hadoop storage architecture is shown in the Fig. 4 [4].

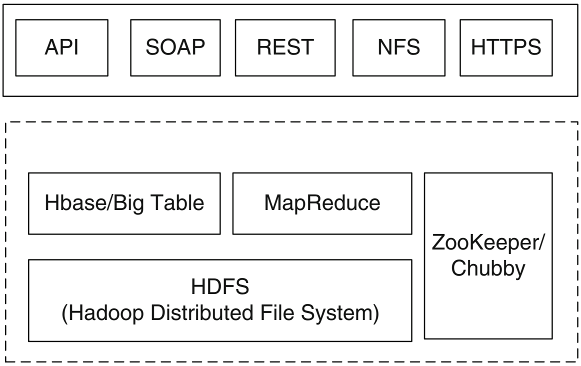


Fig. 4 Hadoop storage system architecture diagram

The Hadoop Distributed File System

HDFS is the storage basis of distributed computing, which has high fault tolerance and can be deployed on cheap hardware equipment, to store massive data [4].

HDFS uses a structural model of master–slave (Master/Slave). A HDFS cluster consists of a NameNode and several DataNodes. NameNode plays as the primary server, which manages the file system namespace and the client operation access to the file. DataNode manages stored data. From internal point, the file is divided into a plurality of data blocks, and the plurality of data blocks are stored in a set of DataNode. The NameNode execute file system namespace operations, such as opening, closing, renaming file or directory. It is also responsible for mapping data block to the specific DataNode. DataNode is responsible for handling the file system client's file read and write requests, and carries on the data block create, delete, and copy job under the unified dispatching of the NameNode. The HDFS architecture is shown in Fig. 5.

HDFS is of high throughput of data reading and writing, provides a basis to massive storage for railway information data, stored as an unprocessed set of source data in the Hadoop distributed file system. In our platform, we use HDFS to store these large amounts of source data.

The MapReduce Distributed Computing Model

MapReduce is a summary of task decomposition and results. Map is to break the task down into multiple tasks, and Reduce is to sum the results of the breakdown multitasking together to get the final result. Calculation process can be concluded as the Map (in_key in_value)— > list (inter_key inter_value) and Reduce (inter_key, list (inter_value))— > list (out_value). In this platform, we will firstly read vast railway information data from the HDFS and divide them into M pieces to operate in parallel Map, and secondly form the state intermediate pair <k, value

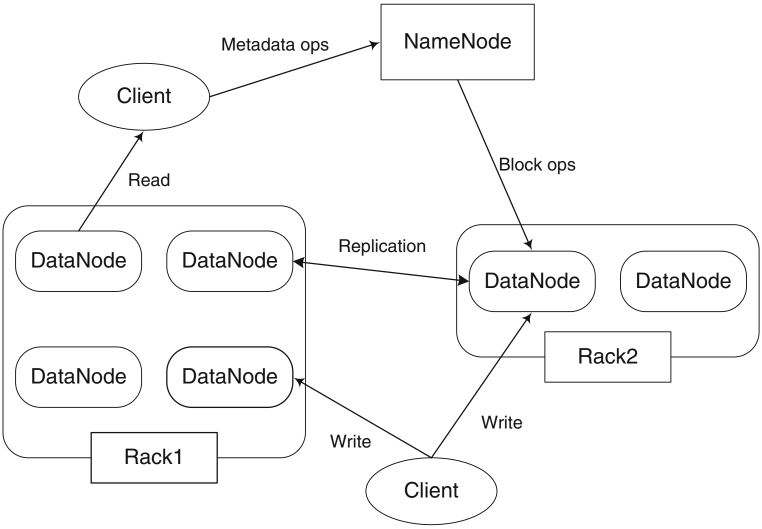
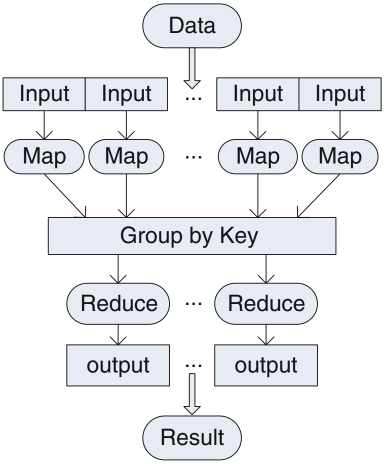


Fig. 5 HDFS architecture diagram

Fig. 6 MapReduce computing model



> , and then operate in Group on k value, just form new <k list (value) > tuple, and then break the tuples into R segments to operate in parallel Reduce, and finally, the processed results shall be stored in the distributed database. Calculation model is shown in (Fig. 6).

The implementation of the MapReduce computing model in this platform is composed of JobTracker running on the primary node singly and TaskTracker running on nodes of each cluster [5]. The primary node is responsible for all the tasks scheduling, which constitutes the whole work, and these tasks are distributed in different nodes. The master node monitors their execution, and reruns the failing tasks; sub-node is responsible for the tasks assigned by the master node. When a Job is submitted, the JobTracker receives operation and configuration information, it will distribute the configuration information to the sub-nodes, and schedule the task and monitor TaskTracker's execution. Our platform use this way to achieve the massive railway information data processing.

HBase Distributed Database

HBase is of high-reliability, high performance, oriented column, scalable distributed storage system [6]. The data line includes three basic types, Row Key, Timestamp and Column Family. Each line includes a sortable line keywords that uniquely identifies the data row in a table. An optional time stamp, each data operation has an associated timestamp. One or more column clusters, each column clusters can consist of any number of columns, they can have data or not. Vast amounts of railway information data after the MapReduce computation can use the value of k as a line keyword for distributed storage, implements massive data storage and management functions. Railway information data storage sample as shown in Table 1.

Line keywords represent railway geographic, construction, operation and maintenance, and dispatching information. Timestamp means the time it cost to operate the data. As is shown in the table, at time t3 Jinan to Qingdao direction at 73 km happens Roadbed damages, and in the moment of t6 shows the dispatching information that the G88 train starts from Shanghai to Beijing at 15:00.

The Platform Test and Result Analysis

Platform Performance Test

When testing the system, the data files are divided into different order of magnitude to get rule numeration, and time consuming of the single machine and of Hadoop cluster shall get comparative analysis. Test results is shown in (Fig. 7).

Table 1 Railway information data storage sample

Row Key	Timestamp	Column family	
		Location	Value
Geographic	t1		
Construction	t3	Jinan	Jinan-Qingdao 73 km roadbed damage
	t2		
Maintenance	t4	Beijing	
Dispatching	t5	Shanghai	Shanghai-Beijing G18 15:00

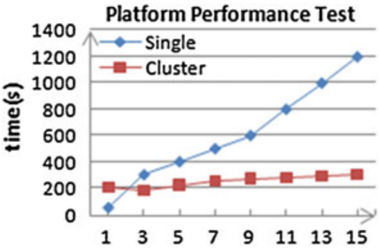


Fig. 7 Clustering performance test results

We can see from the diagram that when the system deals with 1 GB data, the elapsed time the cluster takes is about 4 time of single machine, which is because the distributed architecture of cluster costs some time when the system initialization and intermediate files generated and passed. When the data quantity is small, the Hadoop cluster cannot play out the advantages of distributed computing. As the amount of input file data, Hadoop cluster advantages of distributed parallel computing plays out gradually. When the amount of entering data increases to 15 GB from 5 GB, single machine processing time increases significantly, while processing time of cluster system increases in a tiny amplitude. When data volumes get close to 20 GB, cluster system takes about a quarter time of single one. Data test shows that with the amount of data increasing, the cluster saves more time than single machine, which embodies the advantage of the Hadoop cluster on the large amount of data processing speed.

The Result Analysis

Through performance tests, this platform not only can efficiently store and manage massive data, but also has the following features:

(1) High safety and reliability. System will save file in different server in the form of multiple copies to ensure the security and integrity of data.
(2) Data processing speed is fast. System makes documents distribution to different local compute nodes to process the data, which reduces data transfer amount and improves the speed of data processing through the MapReduce model.
(3) Operation cost is low. Using distributed computing architecture, server performance requirements are lower, just to reduce the cost.
(4) God extensibility. System adopts parallel expansion method, which can extend cluster scale and storage capacity at any time according to need.

Conclusion

This paper bases on the Linux cluster technology, uses Hadoop distributed technology, employs the HDFS distributed file system, Map/Reduce distributed computing model and HBase distributed database technology to deal with huge amounts of data, and designs and develops the vast railway information data storage platform based on Hadoop. By doing a lot of ordinary test on cheap computers, it meets the requirements of railway information efficient storage and management of massive data. This platform has the characteristics of high safety and reliability, fast data processing speed, low running cost, good scalability, which will provide certain reference value to railway department for data storage.

Acknowledgments This research is supported by National Natural Science Foundation of China under Grant 61071076, the National High-tech Research And Development Plans (863Program) under Grant 2011AA010104-2, the Beijing Municipal Natural Science Foundation under Grant 4132057.

References

1. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(01):107–113
2. Apache Hadoop, http://hadoop.apache.org/core/
3. Papadimitriou S, Sun J (2008) DisCo: distributed co-clustering with Map-Reduce. IEEE ICDM'08, 2008, pp 512–521
4. Yang HC, Dasdan A, Hsiao RL, Parker DS (2007) Mapreduce-merge: simplified relational data processing on large clusters. SCMD'07, 2007, pp 1029–1040
5. Li Y (2010) Research on parallelization of clustering algorithm based on MapReduce. Zhongshan University 2010, pp 30–33
6. Xue-song D, Jing Z, Qiang G (2010) A massive data management system based on the hadoop. Microcomput Inf 26(05-I):202–204

Hybrid Data Fusion Method Using Bayesian Estimation and Fuzzy Cluster Analysis for WSN

Huilei Fu, Yun Liu, Zhenjiang Zhang and Shenghua Dai

Abstract Data fusion is the process of combining data from multiple sensors in order to minimize the amount of data and get an accurate estimation of the true value. The uncertainties in data fusion are mainly caused by two aspects, device noise and spurious measurement. This paper proposes a new fusion method considering these two aspects. This method consists of two steps. First, using fuzzy cluster analysis, the spurious data can be detected and separated from fusion automatically. Second, using Bayesian estimation, the fusion result is got. The superiorities of this method are the accuracy of the fusion result and the adaptability for occasions.

Keywords Data fusion · Fuzzy cluster analysis · Bayesian estimation · Spurious data

H. Fu · Y. Liu (✉) · Z. Zhang · S. Dai
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: liuyun@bjtu.edu.cn

H. Fu
e-mail: 12120065@bjtu.edu.cn

Z. Zhang
e-mail: zhjzhang1@bjtu.edu.cn

S. Dai
e-mail: shdai@bjtu.edu.cn

H. Fu · Y. Liu · Z. Zhang
Key Laboratory of Communication and Information Systems, Beijing Municipal
Commission of Education, Beijing Jiaotong University, Beijing 100044, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_91, © Springer Science+Business Media Dordrecht 2014

Introduction

Sensor networks usually have a large number of sensor nodes to observe the interest parameter in the environment. This often results in data redundancy and repetition. As a consequence of that, the energy of the network is wasted. It is important to minimize the amount of data transmission so that the lifetime can be extended and the bandwidth can be saved. Data fusion is the process of combining data from multiple sensors in order to minimize the amount of data [1]. Many algorithms about data fusion have been proposed in literatures and books include arithmetic mean, Kalman filter, Bayesian estimation, Entropy theory, Dempster-Shafer theory, fuzzy logic and neural network [2–8]. As the energy of the network is limited, the data fusion algorithm for WSN is required to be easy. The objective of multisensor data fusion is to obtain an accurate, consistent and meaningful information. This information cannot be achieved by any single sensor in the network because of the uncertainties in the network. These uncertainties are mainly caused by two aspects, device noise and spurious measurement. Device noise here includes device inaccuracy and the noise in the environment. Spurious measurement is caused by sensor failure or even security attack. If any of the aspects is not considered during data fusion, it might lead to an inaccurate or erroneous result. Hence, a good data fusion algorithm should consider both of the aspects.

Considering the device noise and the spurious measurement, this paper proposes a new method for data fusion. The two aspects have different performance on the observation. The observation of each sensor is a combination of the true value of the interest parameter and the device noise. The aspect of device noise is reflected in the variance of the measured data. The quantification of the device noise can be obviously got by the variance of the measured data. The aspect of spurious measurement is mainly reflected in the mean value of the measured data. The quantification of the spurious measurement can be got by comparing the value of the observation between sensors. This method uses fuzzy cluster analysis to deal with the spurious data and then uses the Bayesian Estimation or random weighting method in [7] to deal with the device noise.

Related Work

A random weighting method for multisensory data fusion is proposed in [7]. This method achieves the fusion result in least mean square error. However, this method does not consider the probability that the sensor will provide spurious measurement and it will fail under that condition. Dealing with the spurious measurement, there are several methods reported in the literature. A Bayesian method is proposed in [8], it is effective in sensor validation and identification of inconsistent data. However, the spurious data is not eliminated totally and still has some impacts on the fusion result. Also, a constant value b_k is unsolved, different value will lead to different results. Fuzzy cluster analysis is another one of the most effective methods to recognize the inconsistency. A comparison of all the method in fuzzy cluster analysis is made in [9]. This paper finds out the best method in fuzzy cluster analysis. The conclusion of [9] is used in this paper. And the methods in [7, 8] are used as comparisons during simulation.

Hybrid Data Fusion Method

General Distribution Under Gaussian White Noise

The observation of each sensor is a combination of the true value of the interest parameter and the noise. Assume that x is the true value of the interest parameter, w is the value of device noise, y is the measured value. Then the equation of the measured data can be written as:

$$y = x + w \tag{1}$$

Assume that the device noise of each sensor is Gaussian white noise. σ_w is the standard deviation of Gaussian white noise and the noise has a mean value 0.

The distribution of the measured data is also Gaussian, and it can be written as

$$P(Y = y|X = x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}} \tag{2}$$

σ is the standard deviation of the measured data and it is equal to the standard deviation of Gaussian white noise. x is the true value of the interest parameter and it is also the mean value of the observation

$$\mathrm{Var}[y] = E(y - E[y])^2 = E(y - x)^2 = E(w)^2 = \sigma_w^2 = \sigma^2 \tag{3}$$

Spurious Measurement Elimination

In fuzzy math, the fuzzy cluster analysis can give a quantitative determination of sample relationships using mathematical methods. According to incomplete statistics, there are 13 methods in fuzzy cluster analysis now. Such as Hamming distance, Euclidean distance, Chebyshev distance, absolute reciprocal… In the previous research [9], the author has given an analysis and comparison of 13 methods. It puts forward three principles to compare the 13 methods. The result shows that only the absolute reciprocal meet the requirement of the three principles. The formula of the absolute reciprocal is

$$r_{ij} = \frac{c}{\sum_{k=1}^{m} \left| y_{ik} - y_{jk} \right|} \; i \neq j \tag{4}$$

r_{ij} is the similarity between sample y_i and y_j. c is a constant value. k is the round of measurement.

$$r_i = \sum_{j \neq i} r_{ij} \tag{5}$$

r_i is the overall similarity between sensor i and other sensors, or the support degree of sensor i from other sensors.

From the formula we know that if a measurement has a clear difference from other sensors, r_{ij} will be decreased. And if the sensor continues to provide spurious data, as the number of sensors and the amount of measurement increase, the decrement will become larger. This means if the sensor continues to provide spurious data, the support degree will decreases significantly. But if the sensor only provides a few spurious data, the support degree will still decrease but not very obvious and then this sensor will not be eliminated. So, if a sensor measurement has a clear difference from the other sensors and the other sensors' measurements are in agreement, the support degree of the sensor is low. Here, the support degree consists of k rounds measurements, because the measurement of one round has too much randomness, and the effect of fuzzy cluster analysis is not good.

For the purpose of spurious measurement elimination, a threshold k is needed to separate the spurious measurement.

In order to eliminate the effect of the constant c and give an appropriate threshold for all range of value, r_i needs to be normalized.

$$v_i = \frac{r_i}{\sum_i r_i} \tag{6}$$

Hence, if $v_i < k$, that means the support degree of sensor i from other sensors is low, the measurement of sensor i is considered as spurious measurement. The measurement should be eliminated. If $v_i > k$ the measurement of sensor i is considered in data fusion.

Threshold Selection

In Gaussian distribution, the probability that the point fall in $[x - \sigma, x + \sigma]$ is 0.6826. And the probability that the point fall in $[x - 3\sigma, x + 3\sigma]$ is 0.9974. Hence, we consider $x, x + \sigma, x + 3\sigma$ for setting the threshold k.

$$k = v_{x+3\sigma} = \frac{\frac{c}{3\sigma} + \frac{c}{2\sigma}}{\left(\frac{c}{3\sigma} + \frac{c}{2\sigma}\right) + \left(\frac{c}{\sigma} + \frac{c}{2\sigma}\right) + \left(\frac{c}{3\sigma} + \frac{c}{\sigma}\right)} = 0.2273 \tag{7}$$

Bayesian Estimation for Data Fusion

According to the Bayesian formula in Statistics,

$$\mathrm{P}(\mathrm{x}|\mathrm{y}) = \frac{\mathrm{P}(\mathrm{x})\mathrm{P}(\mathrm{y}|\mathrm{x})}{\sum \mathrm{P}(\mathrm{x})\mathrm{P}(\mathrm{y}|\mathrm{x})} = \frac{\mathrm{P}(\mathrm{x})\mathrm{P}(\mathrm{y}|\mathrm{x})}{\mathrm{P}(\mathrm{y})} \tag{8}$$

x is the true value of the interest parameter, y is the measured value. Since the denominator depends only on the measurement (the summation is carried out over all possible values of state), an intuitive estimation can be made by maximizing this posterior distribution, by maximizing the numerator. This is called maximum a posteriori (or MAP) estimate [8], and the fusion formula is given by

$$\mathrm{x}_{\mathrm{MAP}} = (\sigma')^2 \left(\frac{\mathrm{z}_1}{(\sigma_1)^2} + \frac{\mathrm{z}_2}{(\sigma_2)^2} + \ldots + \frac{\mathrm{z}_\mathrm{n}}{(\sigma_\mathrm{n})^2} \right) \tag{9}$$

$$(\sigma')^2 = \left[(\sigma_1)^{-2} + (\sigma_2)^{-2} + \ldots + (\sigma_\mathrm{n})^{-2} \right]^{-1} \tag{10}$$

From above, we know that the standard deviation of the fused distribution is smaller than any of the individual sensor measurement distribution. Hence, the fused data is more accurate with less uncertainty.

Process of Hybrid Data Fusion Method

Step one: Get sensor measurement samples.
Step two: For each sensor i, calculate v_i. If $\mathrm{v}_\mathrm{i} < k$, sensor i is eliminated from data fusion. If $\mathrm{v}_\mathrm{i} > k$, sensor i is considered in data fusion.
Step three: For each sensor i considered in data fusion, calculate the σ_i.
Step Four: Use MAP estimate or random weighting in [7] to get fusion result.

Simulation Results

A simulation is made to show the performance of the proposed method. A comparison is made to show the improvement of the method. Four methods are compared together. They are arithmetic mean, random weighting proposed in [7], method proposed in [8] and the method proposed in this paper.

In order to simplify, assume there are 3 sensors in the sensor network. In simulation, the variance of the measurement is assumed. It is feasible because, in particular, the measurement variance can be calculated from the sample. The curves in the figures are not the standard Gaussian distribution because the samples are randomly selected. And according to the samples, the probability density is recalculated. This is in order to match the reality. Assume that each sample contains 100 points. In all figures, x label represents the value of the parameter, y label represents the probability of the fused data and the measured data. So the curves mean the probabilities of the measured and fused value at that point around the true value. Let x be the true value (Figs 1, 2, 3).

From the above simulations, it is obviously that the proposed new method can give an accurate result. In situation one, there is no spurious measurement. It is shown in the figure that the new method can give a result which is almost the same with the random weighting. In situation two and three, there is one sensor that provides spurious measurement. Although the ranges of the value are different, the new method can distinguish and eliminate the spurious measurement automatically. The new method can give a result which is more accurate than the random weighting and arithmetic mean. The random weighting method gives more weight to the sensor which has a smaller variance. This method only applies to the situation that none of the sensor provides spurious measurement. And when the variance of the spurious measurement is the smallest, it performs even worse than the normal arithmetic mean. Also the result is better than the method proposed in [8] and it does not need to set the constant value. Hence, the new method can be applied to all the situations.

However, there are still some weaknesses of the proposed method. If two of the three sensors provide spurious measurement, this method will fail. For extension, if the amount of the failed sensor is larger than the normal, this method will lead to inaccurate estimation. It is also the weakness of the fuzzy cluster analysis.

Fig. 1 Situation one. Sensor A: $\sigma_A = 2, x = 20$, Sensor B: $\sigma_B = 2.5, x = 20$, Sensor C: $\sigma_C = 3, x = 20$

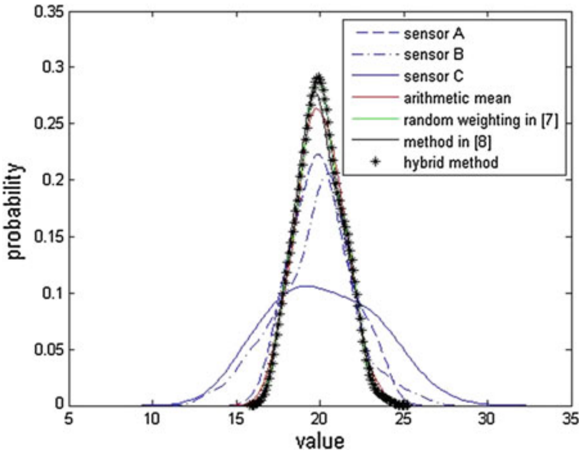


Fig. 2 Situation two. Sensor A: $\sigma_A = 2,\ x = 20$, Sensor B: $\sigma_B = 2.5,\ x = 20$, Sensor C: $\sigma_C = 3,\ x = 40$

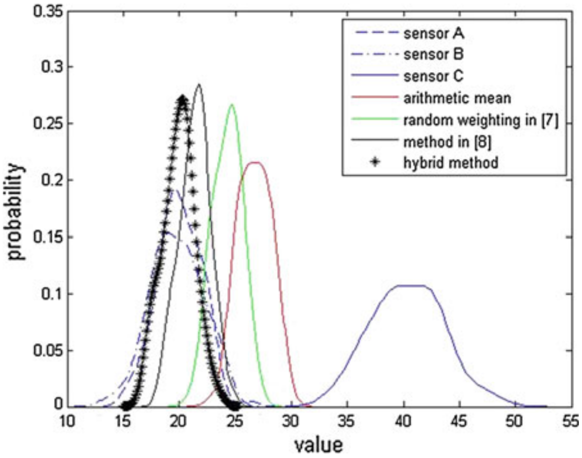
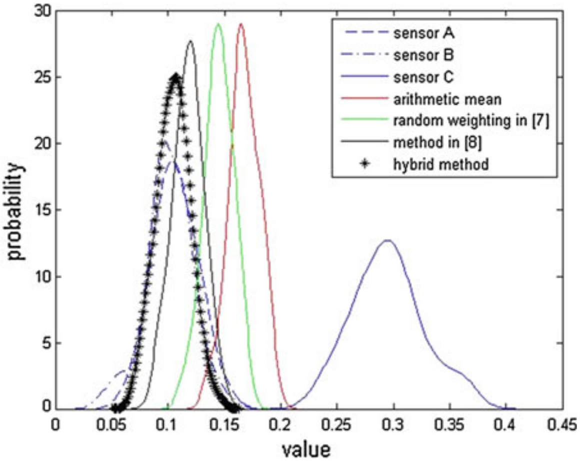


Fig. 3 Situation three. Sensor A: $\sigma_A = 0.02$, x = 0.1, Sensor B: $\sigma_B = 0.025$, x = 0.1, Sensor C: $\sigma_C = 0.03$, x = 0.3



Conclusion

This paper gives a new method for multisensory data fusion. This method can distinguish and eliminate the spurious measurement automatically, regardless of the range of the value. And whether there is spurious measurement, the method can give an accurate result.

Acknowledgments This research is supported by National Natural Science Foundation of China under Grant 61071076, the National High-tech Research And Development Plans (863 Program) under Grant 2011AA010104-2, the Beijing Municipal Natural Science Foundation under Grant 4132057.

References

1. Ozdemir S, Xiao Y (2009) Secure data aggregation in wireless sensor networks: a comprehensive overview. Comput Netw 53:2022–2037
2. Bahador K, Khamis A, Karry FO, Razavi SN (2013) Multisensor data fusion: a review of the state-of-the-art. Inf Fusion
3. Welch G, Bishop G (1995) An Introduction to the Kalman Filter. Department of Computer Science, University of North Carolina, North Carolina
4. Yen J (1990) Generalizing the Dempster–Shafer theory to fuzzy sets. IEEE Trans SMC 20:559–570
5. Maskell S (2008) A Bayesian approach to fusing uncertain, imprecise and conflicting information. Inf Fusion 9:259–277
6. Zhu H, Basir O (2006) A novel fuzzy evidential reasoning paradigm for data fusion with applications in image processing. Soft Comput J—A Fusion of Foundations, Methodologies and Applications, 2006
7. Gao S, Zhong Y, Li W (2011) Random weighting method for multisensor data fusion. IEEE Sens J 11:1955–1961
8. Kumar M, Grag DP, Zachery RA (2007) A method for judicious fusion of inconsistent multiple sensor data. IEEE Sens J 7:723–733
9. Xinzhou W, Haichi S (2003) Construction of Fuzzy Similar Matrix. J Jishou Univ (Nat Sci edn)

Enhancements of Authenticated Differentiated Pre-distribution Key Methodology Based on GPSR

Lin Sun and Zhen-Jiang Zhang

Abstract Wireless sensor network consists of plentiful energy and computing power-constrained tiny sensor nodes. On the basis of the protection of security, the main goal of our study is to maintain the network life time at a maximum with the appropriate routing protocol. In this article, we will propose a new routing algorithm on the basis of the original routing algorithm. Considering the security of the wireless sensor network, the residual energy of nodes as well as issues such as physical distance, we try to extend the maximum lifetime of a wireless sensor network with all the combination of these factors for routing. This energy-balance routing algorithm takes into account the number of pre-distributed shared keys between the sending node and receiving node, the residual energy of the receiving node and the physical distance between two nodes in order to protect the security at the same time, to maintain the network life time at a maximum.

Keywords WSN · Energy saving · Life time · RKP

Introduction

With the rapid development of wireless sensor network, it has too many crucial applications in many aspects, i.e. military applications, environmental applications, health applications, home applications and numerous commercial applications [1].

L. Sun · Z.-J. Zhang
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: 11120222@bjtu.edu.cn

L. Sun · Z.-J. Zhang (✉)
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China
e-mail: zhjzhang1@bjtu.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_92, © Springer Science+Business Media Dordrecht 2014

Compared with traditional networks, wireless sensor network has many advantages, but also because of its features of battery-powered, large number of nodes, mobility, and low processing capacity defect, so that there are a lot of restrictions on its applications. This requires complete the maximum amount of data transfer with minimal energy cost under the resource-constrained condition such as computing power and storage capacity.

At the same time, wireless sensor nodes are often deployed in public, untrusted or even hostile environments, which often bring a large quantity of security problems. The security problems will be more important in wireless sensor network than in other networks. In all safety problems, the security of data transmission is particularly significant. The main target in the design of the security agreement is to provide a safe and reliable security data communication between sensor nodes. The first key agreement protocol was introduced by L. Eschenauer and V. D. Gligor called random key pre-distribution (RKP) which has been very popular during these years.

Random key pre-distribution (RKP) is the one of the key distribution that is developed for the sake of safe communication of WSN. The main characteristic of RKP [2] is that each sensor is pre-distributed with k distinct keys randomly chosen from a key pool with the size of K keys before deployment. Then two neighboring sensors attempt to establish a pair-wise key with the pre-distributed keys, for secure communications between themselves [3].

Our contributions: In this paper, we show some improvements in their routing protocol and the differentiated key pre-distribution algorithm. We demonstrate that the next hop will be chosen not only considering the residual energy of the receiving node but also the physical distance between the two nodes. Moreover, we propose a simple modification to the original protocol to protect the security while taking the problem of network lifetime into account.

Paper Organization: The remainder of the paper is organized as follows: section Background and Related Work focuses on the background of RKP Scheme and routing protocol GPSR and section Review of the Methodology reviews the differentiated key pre-distribution methodology. In the next section the weaknesses of the former methodology are discussed while the our revised protocol are proposed in section Our Revised Version of the Protocol. Final section concludes the paper.

Background and Related Work

Something About RKP

In this section, we present a basic introduction of the RKP algorithm., including an extension of the concrete steps and a variety of RKP algorithm.

Basic RKP Scheme.

RKP is one of the key distribution that is developed for the sake of safe communication of WSN. RKP specific algorithm can be divided into two steps.

The first step is the key pre-distribution and the second is the pairkey establishment. In the first step the process of the key pre-distribution, each node randomly selects the same number of k keys from key pool with the size of K keys as its pre-distribution keys. After the first step, we carried out the second step pairkey establishment. Nodes communicate with its neighbors within their communication range in order to find the specific neighbor which has the same pre-distributed keys with them, then nodes, with their neighbors, establish pairkeys which can be used to encrypt/decrypt communications between neighboring sensors in each hop.

Variants of RKP Scheme.

In [4], it demonstrates a probabilistic unbalanced distribution of keys throughout the network, which distributes different number of keys to nodes according to the function of the specific each node. In the area of network security, the cluster head nodes are responsible for key distribution, updating the keys, the expulsion of nodes and seeking common features such as recovery. Taking into account that the cluster head node is in the important role of wireless sensor network, the cluster head node needs to ensure that the information is not procure, through encryption and key management.

In [5], considering in homogeneous sensor networks, where all sensor nodes have identical capabilities in terms of with limited computation, communication, energy supply and storage capability, nodes are divided into the H nodes and L nodes. H sensors are more powerful nodes and have more computation, communication, energy supply and storage capability, while L nodes have limited computing power, communication capabilities, energy supply and storage capability.

In [3], three random key pre-allocation algorithm, the q-composite random key pre-distribution scheme, the multi-path key reinforcement scheme and the random-pairwise keys scheme are presented, proposing the idea of multi-key path to increase the flexibility of the link between the two points, which can better able to protect the security of data transmission under attack.

In [6], each node with a simple calculation ability based on a symmetric key system only has a small number of pre-shared keys regardless of the size of the entire sensor network. While the multipath key reinforcement scheme in [7] uses multiple physically disjoint paths between two sensor nodes and decides the number of pre-distributed keys according to the specific security level.

Something About GPSR

This article will present some improvements to the original routing protocol GPSR. In this well known location centric routing protocol, the node uses greedy algorithm to make the choice of the next hop from its neighbour nodes. If the distance from all the neighbors to the destination are further than the current node, then there will be the existence of a “hole”. The node will follow the right-hand rule to forward along the “hole” in the border until it finds a distance closer to the destination node, and then forward packets using greedy algorithm.

Energy Saving in WSN

In the design of network routing protocols, the node energy consumption and the network energy balance problems will be considered to avoid frequent use of a path or a few nodes. The energy of these nodes will soon run out [8], which can not only result in the incomplete coverage of the network to obtain monitoring data, but also seriously affecting the entire wireless sensor network security.

We will do a brief introduction to a few formulas in this article concerning the energy-efficient wireless sensor networks. In [9], based on the preliminary report on this work in [10] and comparing with [11], the methodology gives a weight value to each of its neighbor based on the initial energy level of node energy and the current energy level, used to indicate the node energy consumption of the neighbor how much, so that in the routing. Then in [12], node, in the different pre-distribution key methodology, are divided into c classes, so the difference is to abandon the node graded idea in this paper.

Review of the Methodology

In this section, we depict the intuition and technical details of the differentiated key pre-distribution methodology and the GPSR routing protocol.

The core idea of the original text methodology is to pre-distribute different number of keys to different nodes. The quantity of N nodes is divided into c classes, and each class of each node is assigned with k_i ($1 \leq i \leq c$) keys. In the key pre-distribution phase, each node of the first class randomly selected k1 keys from the key pool with the size of K. Then the first class k nodes are divided into two parts: A part consists of $n_1 - (k_i - \lfloor k_i/n_1 \rfloor \cdot n_1)$ nodes, well B part possess $(k_i - \lfloor k_i/n_1 \rfloor \cdot n_1)$ nodes. In order to increase the probability of the key share between nodes, other class of nodes apart from class 1 nodes, follow the semi-random key pre-distribution process. The process is as follows: each node in class i ($1 \leq i \leq c$) randomly chooses $\lfloor k_i/n_1 \rfloor$ keys from each of the A part nodes, and at the same time chooses $\lceil k_i/n_1 \rceil$ keys from each of the B part ones. If some of the chosen keys are the same, repeat the above steps until all k_i keys are distinct. We define $\lceil x \rceil$ as the smallest integer no less than x, while $\lfloor x \rfloor$ is the largest integer no more x.

After the key pre-distribution phase, the node begins to communicate with its neighbors within its communication range γ and gets the ID of its neighbors as well as their pre-distributed keys. Node i establishes a one-hop or two-hop secure communication link by sending request to its secure neighbors which have the same pre-distributed key with node i. We assume that there are p two-hop links between node i and node j, and define $s_1 (1 \leq l \leq p)$ as the proxy of each link,and k(i, j) as the number of the shared keys between node i and node j. Then the protection keys can be obtained as follows,

$$\mathrm{key}\,(\mathrm{i},\mathrm{j}) = \mathrm{k}\,(\mathrm{i},\mathrm{j}) + \sum\nolimits_{\mathrm{l}=1}^{\mathrm{p}} \min(\mathrm{k}\,(\mathrm{i},\mathrm{s_l}),\mathrm{k}(\mathrm{s_l},\mathrm{j}))$$

Routing

The original text presents some improvements to GPSR in order to better protect the security of point to point communication. We give weight values to secure neighbors which are closer to the sink than the node itself. U(i) is defined as the set of the secure neighbors which are closer to the sink than the node itself and key(i, j) as the number of the protection keys between node i and node j. So the weight value can be obtained as,

$$\omega = \frac{key(i,j)^{\alpha}}{\sum_{\mathrm{m}\in \mathrm{U(i)}} \mathrm{key}(\mathrm{i},\mathrm{m})^{\alpha}}$$

ω_{j} is the probability that node j is chosen as the proxy of node i. When α is zero, all nodes have the same probability of being selected independent of the resilient of the link. While when α is positive, the bigger the resilient of the link, the higher the probability that the node will be selected. Only the highest resilient link can be chosen when α trends toward infinity. Mean while the selection of α also has something with the attention of the system to the network security and the lifetime.

Weaknesses of the Differentiated Key Pre-distribution Methodology

In the original text, the weight value only pays attention to the number of the protection keys on the link, as a result the energy consuming will be much higher than the links with lower resilient and energy of nodes on the high resilient link drain out rapidly. The disappearing of the high resilient links brings a more negative impact on the security of the whole network.

Moreover the α in the formula takes the balance of the network lifetime and the security into consider. When α takes to a relatively high value, it means the link with high resilient is given more priority and the wireless sensor network considers more about the security, as a result the energy of the nodes on these links drain out rapidly, which, to a very great extent, contraries to our starting point for trying to improve network security.

Our Revised Version of the Protocol

Data transmission in the wireless sensor network needs the help of the organic collaboration between nodes. While sensor nodes powered by batteries, have limited computing, communications, energy and other resources. Considering that, an important problem at the same time ensuring the security of data transmission we also needs to solve is how to save energy, in order to prolong the life cycle of wireless sensor network.

In this paper, we will preprocessing the data collected by the nodes before the data transmission, and define a new weight value. The new weight value W_{ij} not only takes the protection of security on the basis of saving energy into account, but also able to balance the residual energy. Moreover the W_{ij} adjusts the routing protocol according to the residual energy of the node itself, which makes a contribution to entend the lifetime of the node. The weight value W_{ij} can be obtained as,

$$W_{ij} = \theta^{\alpha}\omega^{\beta}$$

In this formula α, β are two parameters that control the relative weight of security and network lifetime. When α is zero, the weight value considers more about the number of the protection keys on the link which means the security of the link. On the contrary, when β is zero, the residual energy of the receiving node and the distance between the transmitting node and the receiving node will be considered more, which pay more attention to the problem of energy. When α tends to infinity, the situation is almost the same when β is zero. While when α, β are positive, the weight value makes a combination of the security and the network lifetime. An appropriate α, β can be used to make a good trade-off in the energy and life and security, on the basis of the protection of security, and at the same time as far as possible to extend the life of the nodes to reduce energy consumption.

This weight value in this paper uses parameter ω as the considering of the security and the ω is defined in the original text.

θ is introduced in this algorithm as the consider of the residual energy of the receiving node and distance of the two nodes when choosing the next hop routing in order to make a better trade off between the security and the energy of the node. θ can be defined as follows,

$$\theta = \frac{\eta(i,j)^{-\alpha_2}}{d(i,j)^{\beta_2}}$$

In order to protect the energy balance in the data transmission, this algorithm is dedicated to the residual energy of the receiving node to calculate the weight values, and find the node with highest weight value to transfer data. Improvements on the basis of this algorithm to get rid of the assumptions on the grade if nodes, but only consider the physical distance between the nodes and the residual energy of the receiving node, have been made. Suppose that node j is the receiving node of node i, and η(i, j) is the energy consumption of the secure neighbor of node i, d(i, j) is the distance between node i and node j. In the formula, α and β represent the rest energy factor and the distance factor. α_2,β_2 are the parameters make a perfect trade off between the residual energy of the receiving node and the distance between the two nodes. The system will only consider the physical distance between two nodes on the condition that when α_2 is zero and in the other case when β_2 is zero, the residual energy will be the only consideration. Moreover when α_2 is positive, the node with more residual energy will get a bigger weigh value than the one with the closer distance, and at the same time will be chosen preferentially when routing. When α_2 tends to infinity, only the node with the most residual energy can be chosen. The value of parameter β_2 is roughly the same with α_2. During the routing, node i will take the residual energy of the receiving node as well as the distance between two nodes into consideration because of the the existence of α_2 and β_2. On the basis of the original algorithm, taking into account the residual energy issues as well as physical distance, we can ensure the nodes with less residual energy while on the high resilient link bear less communication traffic, but the surplus energy nodes on the relatively lower resilient link take part in the data communication tasks to make a better balance of energy. In this way, it may extend the network life time.

The weight coefficient $\eta\,(\mathrm{i},\mathrm{j})$ of all the nodes j in communication radius of i is defined as the following,

$$\eta\,(\mathrm{i},\mathrm{j}) = \frac{(\mathrm{I}-\mathrm{e_j})^{-1}}{\sum_{\mathrm{n}\in \mathrm{U}_{(\mathrm{i})}}(\mathrm{I}-\mathrm{e_j})^{-1}}$$

In this formula, η(i, j) is a weight value about the condition of the energy consumption of the secure neighbor of node i. I can be defined as the initial energy level and with the passage of time, the energy consumed gradually. Then $\mathrm{e_j}$ is the current energy level of the receiving node j. To simplify the problem, we set a proper time T as the time period and after time T, the node would broadcast to its neighbors their energy levels, so that its neighbors will use the value of its residual energy during routing as a reference. At last choosing $\mathrm{U}_{(\mathrm{i})}$ as the set of the secure neighbor of node i. Above all, the node with lower residual energy is less likely to be chosen as the next hop than the node with the higher residual energy level. $\eta(\mathrm{i},\mathrm{j})$ can be a crucial reference as the measure of the secure neighbor of node I when routing.

Description of the Algorithm:

Step 0. Initial the network and then divide the nodes into c classes.

Step 1. Pre-distribute the keys to each node and then establish the pairwise keys between secure neighbors.

Step 2. Collect the information from its secure neighbors including the residual energy level and the distance between them.

Step 3. Get the weight value according to the formula 6 to choose the perfect next hop.

Step 4. Circulate step 2 until reach the destination sink node.

This algorithm is dedicated to finding a routing algorithm, not only in the protection of security on the basis of energy conservation, but also to make a perfect energy balance. More importantly, the algorithm in this paper is to adjust the routing protocol according to the residual energy level of the node itself, so we can extend the lifetime of the node to the maximum.

Conclusion

In this paper we show the weakness of the differentiated key pre-distribution protocol which only takes the resilience of links into consideration. First it is overly dependent on the link resilience to choose the next hop during routing, as a result the energy of nodes on the high resilient links will drain out rapidly. Then the security of the whole network will seriously affected. In general, the weaknesses of protocols in the original text are arisen from the only consideration of resilience of link during routing.

We also provide a revised version of this protocol which adds two crucial parameters into the routing weight value which are the residual energy level and the distance between the two nodes. The studied protocol, on the basis of ensuring the security of the network, makes a wonderful energy balance in the whole network.

So designing a perfect secure routing protocol has been taken into account as our future work.

Acknowledgments This research is supported by National Natural Science Foundation of China under Grant 61071076, the National High-tech Research and Development Plans (863 Program) under Grant 2011AA010104-2, the Beijing Municipal Natural Science Foundation under Grant 4132057.

References

1. Du W et al (2003) A pairwise key pre-distribution scheme for wireless sensor networks. In: Proceedings 10th ACM conference computer and communication security, Oct 2003, pp 42–51
2. Eschenauer L, Gligor VD (2002) A key-management scheme for distributed sensor networks. In: Proceedings 9th ACM conference computer and communication security. Security, Nov. 2002, pp 41–47
3. Chan H, Perrig A, Song D (2003) Random key predistribution schemes for sensor networks. Department of Electrical and Computer Engineering, Paper 20
4. Patrick Traynor, Heesook Choi, Guohong Cao, Sencun Zhu and Tom La Porta (2006) Establishing pair-wise keys in heterogeneous sensor networks. In: INFOCOM 2006. Proceedings 25th IEEE international conference on computer communications
5. Poornima AS, Amberker BB (2008) Tree-based key management scheme for heterogeneous sensor networks. In: Networks, 2008. ICON 2008. 16th IEEE international conference on 2008
6. Zhu S, Xu S, Setia S, Jajodia S (2003) Establishing pairwise keys for secure communication in ad hoc networks: a probabilistic approach. In: The 11th IEEE international conference on network protocols, IEEE, 2003
7. Chan H, Perrig A, Song D (2003) Random key predistribution schemes for sensor networks. In: Proceedings of the IEEE security and privacy symposium 2003, May 2003
8. Wu CX, Liu Y (2012) WSN on-demand multipath routing protocol based on energy-aware. In: Comput Eng 38(9)
9. Gu W, Dutta N, Chellappan S, Bai X (2011) Providing end-to-end secure communications in wireless sensor networks. IEEE Trans Netw Serv Manag 8(3):205–218

10. Camilo T, Carreto C, Silva J, Boavida F (2006) An energy-efficient ant base routing algorithm for wireless sensor networks. In: ANTS 2006—Fifth international workshop on ant colony optimization and swarm intelligence, 4150, pp 49–59
11. Okdem S, Karaboga D (2006) Routing in wireless sensor networks using ant colony optimization. In: First NASA/ESA conference on adaptive hardware and systems—AHS, pp 401–404
12. Xue J, Qi X, Wang C (2011) An energy-balance routing algorithm based on node classification for wireless sensor networks. J Comput Inf Syst 7:2277–2284

Research on Kernel Function of Support Vector Machine

Lijuan Liu, Bo Shen and Xing Wang

Abstract Support Vector Machine is a kind of algorithm used for classifying linear and nonlinear data, which not only has a solid theoretical foundation, but is more accurate than other sorting algorithms in many areas of applications, especially in dealing with high-dimensional data. It is not necessary for us to get the specific mapping function in solving quadratic optimization problem of SVM, and the only thing we need to do is to use kernel function to replace the complicated calculation of the dot product of the data set, reducing the number of dimension calculation. This paper introduces the theoretical basis of support vector machine, summarizes the research status and analyses the research direction and development prospects of kernel function.

Keywords Support vector machine · High-dimension data · Kernel function · Quadratic optimization

Introduction

Support vector machine (SVM) was introduced into the field of machine learning and its related area in 1992 [1], having received widespread attention of researchers in later time and has made great progress in many fields. It uses a

L. Liu · B. Shen
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: 11120126@bjtu.edu.cn

L. Liu · B. Shen (✉)
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China
e-mail: bshen@bjtu.edu.cn

X. Wang
China Information Technology Security Evaluation Center, Bejing, China
e-mail: wangx@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_93, © Springer Science+Business Media Dordrecht 2014

nonlinear mapping to map original training data into high-dimensional data space in order to find the optimal classification hyper plane separating those data belonging to different categories. Support vector machine is based on SLT (statistical learning theory) [2, 3] VC dimension theory and structural risk minimization principle. Compared with traditional neural networks, support vector machine gains great enhancement in generalization ability and overcomes some problems existing in feed-forward neural networks, such as local minimum and the curse of dimensionality [4]. The introduction of kernel function greatly simplifies the complexity of dot product operation in support vector machine for nonlinear data classification, and it can be used to distinguish and enlarge the useful features, and support vector machine based on kernel function is playing a powerful role in the field of data mining.

Support Vector Machine

If the training data set is linear separable, the given data set is $\mathrm{D}:\ (X_1, y_1)$, $(X_2, y_2), \ldots, \left(X_{|D|}, y_{|D|}\right)$, among which X_i is training data with class label y_i. The scope of each y_i is $+1$ or -1, namely $y_i \in \{+1, -1\}$. In dealing with classification problem, the optimal classification hyper plane can be denoted as follows:

$$W \cdot X + b = 0. \tag{1}$$

W is a weight vector, that is to say, $W = \{w_1, w_2, \ldots, w_n\}$, where w_i is the weight of X_i; n is the number of attributes; parameter b is a scalar, and is often referred to as the bias. The formula $W \cdot X$ stands for the dot product of W and X. From geometric point of view, the entire input space can be divided by hyper plane into two parts: one part is positive value data set; another part is the negative one. Hyper plane is a line in two-dimensional space, a surface in three-dimensional space. The biggest edge distance between two types of training data set is $\frac{2}{\|W\|}$. Support vector machine discovers the optimum classification hyper plane by means of support vectors and the edges between them [5] and gets the maximum edge distance of two classes of data sets at the same time.

Research on Kernel Function

For linear separable data, support vector machine can directly classify the data set into two categories in the input space; for those nonlinear separable data, SVM has to map the original input data X with nonlinear mapping $(\Phi : X \to F)$ into another high-dimensional space where we can solve the maximum interval of classification, and this new high-dimensional space is the feature space. A dot product operation can be directly substituted by kernel function in feature space, and we needn't know the concrete eigenvector and mapping function, which is also known as kernel trick. Frequently-used kernel functions include the following ones:

$$\text{Linear kernel function: } K(X_i,\ X_j) = X_i \cdot X_j. \tag{2}$$

$$\text{Polynomial kernel function: } K(X_i,\ X_j) = (X_i \cdot X_j + 1)^h. \tag{3}$$

$$\text{Gaussian radial basis function kernel function: } K(X_i,\ X_j) = \ell^{-\left\|X_i - X_j\right\|^2/2\sigma^2}. \tag{4}$$

$$\text{Sigmoid kernel function : } K(X_i, X_j) = \tanh(\kappa X_i \cdot X_j - \delta). \tag{5}$$

The performance of support vector machine mainly depends on model selection, including the selection of the kernel function type and the kernel parameter selection [6]. In the study of kernel function selection, the kernel alignment is a good method [7]. Kernel alignment method is based on a hypothesis: the kernel matrix of a good kernel function should be as similar as possible to the calibration matrix. Under normal circumstances, the first consideration of choosing kernel function is the Gaussian radial basis function kernel function (RBF), which is because RBF has fewer parameters to select, and what's more, for some parameters, RBF has similar performance to the Sigmoid kernel function.

The development and application of support vector machine have been greatly promoted since the introduction of kernel function, and its application area has extended from hand-written numeral recognition, reference time series prediction test and other traditional application area to new areas such as information image processing [8], industrial process control, etc. The following content in this paper will center on the discussion of kernel function in support vector machine and put forward its future research directions.

Kernel Clustering

Clustering analysis divides data objects into different subsets and the data objects in the same subset are similar to each other, while those located in different subsets have different properties. Kernel clustering combines kernel function and clustering together, which is based on the characteristics of clustering [9]. In the first stepwise, kernel clustering clusters the training data and test data, and then constructs the kernel function on the basis of clustering results. Chapell et al. [10] proposed an overall framework of constructing kernel clustering, using different conversion functions to change the eigenvalue decomposed by kernel matrix. Jason Weston et al. came up with the bagged clustering kernel [10] to overcome the time complexity problem existing in Chapell's clustering kernel. Although the bagged clustering kernel shortened the kernel clustering time, there still is much room for improvement in terms of classification accuracy.

Fuzzy C-means (FCM) clustering algorithm [11] introduces fuzzy set theory to the process of clustering. Fuzzy kernel clustering algorithm firstly maps the data of input space into high-dimensional feature space to enlarge pattern differences between classes, and then carry through fuzzy clustering in the feature space [12, 13]. Fuzzy kernel clustering algorithm is able to highlight the differences of different sample characteristics, increasing clustering accuracy and speed [14]. It has always been an important research problem to expand SVM classifier to multi-class classification [15, 16], so Zhao et al. [17] applied fuzzy kernel clustering to multi-class classification method, solving the serious problem of fuzzy overlapping, but didn't have a large increase in classification speed.

The objective function of the fuzzy kernel clustering algorithm is as follows:

$$J_n(U,\ v) = \sum_{i=1}^{c}\sum_{j=1}^{m} u_{ij}^{n} \left\| \Phi(x_j) - \Phi(v_i) \right\|^2. \tag{6}$$

In the above formula, parameter c is the clustering number initially set; v_i is the initialized clustering center; u_{ij} is the membership function of sample j belonging to i category; $U = \{u_{ij}\}, v = \{v_1, v_2, \ldots, v_c\}$, parameter n is weighted index, and $n > 1$. The criterion of fuzzy clustering algorithm is for the minimum value of the above objective function until each membership value stabilized.

Kernel clustering support vector machine has been successfully utilized in biomedical, text classification [18] and many other application fields, undoubtedly, it will involve a wider range of application fields in later time.

Super-Kernel Function

Support vector machine for data classification often brings about two dilemmas using one single kernel function: one is unable to complete effective nonlinear mapping; the other is over-fitting or under-fitting [19]. The extrapolation ability of Gaussian radial basis function kernel function weakens along with the increase of σ parameter [20], so it has strong locality. Polynomial kernel function regulates different mapping dimensions through adjusting h parameter, and computation grows with h parameter, thus having strong global property and poor locality.

In order to adapt to the increase on data set and high efficiency requirements, many algorithms have been developed, such as large data set training [21] algorithm, super kernel learning [22] algorithm, fast convergence algorithm [23], etc. Combining several kernel functions of different categories with a polynomial composition can give full play to the excellent characteristics of different kernel functions in dealing with data classification, and overcomes the shortage of single kernel function while maintaining translational invariance and rotation invariance, which provides a fresh effective way to studying the construction of kernel function in support vector machine. A simple form of super-kernel function is like the following formula:

$$K(X_i, X_j) = \beta_1 \ell^{-\left\| X_i - X_j \right\|^2 / 2\sigma^2} + \beta_2 (X_i \cdot X_j + 1)^h. \tag{7}$$

It is the linear combination of Gaussian radial basis function kernel function and h polynomial kernel function. Super-kernel function parameters were regulated in the form of parameter vector, that is to say, super-kernel function adjusts all parameters at the same time, and those several more parameters just add to the length of parameter vector but not affect the determination time of parameters.

Kernel Parameter Selection

Selection of kernel parameter has a direct impact on the performance of the SVM classifier. A commonly used parameter selection method is based on the Generalization Error estimates [24]. The Generalization Error estimates predicate and forecast the generalization capability of classification decision criteria by means of training data set [25]. Vapnik et al. [26] estimate generalization ability by span of support vectors on the basis of Generalization Error estimates, which has an advantage of higher accuracy but has more complex computation. A method of utilizing data set to evaluate the optimal kernel parameter method [27] is put forward on the basis of Vapnik's method, determining the optimal kernel parameter choice from a geometric point of view.

Qi et al. [28] proposed a kernel parameter selection method to solve LOO (leave-one-out) upper bound minimum point based on genetic algorithm, where we can choose the reproduction operator and combine genetic algorithm with steepest descent method, improving the accuracy of forecast but without leading to local optimal solution. Chen et al. [29, 30] adopted different generalization ability estimates as the fitness function of genetic algorithm, and it not only reduced the computation time to choose parameters but also the dependence on the initial value.

Parameter selection method based on kernel matrix similarity measure starts from research on kernel matrix in order to search for the optimal kernel parameter and learning model, and this method improved the calculation speed of SVM. Liu Xiangdong et al. [31] eventually found the optimal kernel parameters and the kernel matrix by means of experimenting on UCI standard data set and FERET standard faces library. Parameter selection method based on kernel matrix similarity measure can serve as a feasible method to choose the optimal SVM model, and it also has certain reference value for choosing other kernel parameters.

Conclusion

The study of kernel function of SVM is an important data mining research, so choosing the appropriate kernel function and its parameters can give full play to the performance of SVM and even has remarkable significance in promoting the popularization and application of data mining. This paper does research on the kernel function of SVM and does some summary comments on the kernel clustering, super kernel function and the selection of kernel parameters. Judging from the current study, the author believes that the study of kernel function in the following areas is to be further developed:

1. Finishing data mapping efficiently and reliably in the environment of big data. "Big data" has features of giant, high growth rate and diversification, and it needs new processing mode to excavate useful information from the massive data and get an insight in them. In this case, the conventional transformation of kernel function will face with new bottleneck in processing speed and processing quality. On the basis of existing research, it remains further study to extend the scale of the expansion of kernel function processing data and select the appropriate kernel parameters and further improve the quality and speed of processing data.
2. Giving full play to the advantages of different kernel functions in super-kernel functions. On the perspective of present research on the selection of kernel parameters, Gaussian radial basis function (RBF) and its parameter selection have been more detailed studied in the field by virtue of its favorable advantages in computer vision. How to expand the application area of polynomial kernel function and Sigmoid kernel function, especially how to give full play to the advantage of each kernel function in the super-kernel function still needs deeper exploration. It is worth studying that selecting and optimizing super-kernel function parameters and applying the concept of constructing super-kernel function to support vector machines.
3. Selecting appropriate kernel function of support vector machines for specific applications. The scope of data mining processing data is developing from structured data to the direction of semi-structured and unstructured data. As one part of the data mining classification algorithms, the application fields of SVM continue to expand, thus it appears particularly important to select the appropriate kernel functions of specific domain. The choice of kernel function is closely related to the data field [32], in the meantime, the performance of kernel function depends largely upon the selection of parameters. Future studies are required to determine the kernel function and its parameters according to different application areas of the support vector machine for the sake of reducing the consuming of storage space and computing time of computers.

Big data processing is the future research tendency, and with the arrival of the cloud era, big data is attracting more and more attention. In future work, the author will mainly focus on the study of achieving efficient and accurate classification with support vector machine in the environment of big data.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, and Beijing Natural Science Foundation under Grant 4112045, and the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Boser B, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers [C]. In: Proceedings of the 5th annual ACM conference on computational learning theory, Pittsburgh, pp 144–152
2. Vapnik VN (2000) The nature of statistical learning theory [M]. Translated by Zhang Xuegong (trans: Zhang X). Tsinghua University Press, Beijing
3. Vapnik VN (2004) Statistical learning theory [M]. Translated by Xu Jianhua, Zhang Xuegong (trans: Xu J, Zhang X). Publishing House of Electronics Industry, Beijing
4. Vapnik V (1995) The nature of statistical learning theory [M]. Springer, New York
5. Ju C, Guo F (2010) A distributed data mining model based on support vector machines DSVM [J]. 30(10):1855–1863
6. Zhu S, Zhang R (2008) Research for selection of kernel function used in support vector machine [J]. Sci Technol Eng 8(16):4513–4517
7. Cristianini, N, Taylor Shawe J, Kandola J et al. (2002) On kernel target alignment. In: Proceedings neural information processing systems. MIT Press, Cambridge, pp 367–373
8. Yang Z (2008) Kernel-based support vector machines [J]. Comput Eng Appl 44(33):1–6
9. Li T, Wang X (2013) A semi-supervised support vector machine classification method based on cluster kernel [J]. Appl Res Comput 30(1):42–45
10. Tison C, Nicolas JM, Tupin F et al. (2004) A new statistical model for Markovian classification of urban areas in high-resolution SAR images [J]. IEEE Trans Geosci Remote Sens 42(10):2046–2057
11. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
12. Wu Z, Gao X, Xie W (2004) A study of a new fuzzy clustering algorithm based on the kernel method. Journal of Xi ‘an university of electronic science and technology magazine, 31(4):533–537
13. Zhang N, Zhang Y (2010) Support vector machine ensemble model based on KFCM and its application [J]. J Comput Appl 30(1):175–177
14. Cao W, Zhao Y, Gao S (2010) Multi-class support vector machine based on fuzzy kernel clustering [J]. CIESC Journal 61(2):420–424
15. Angulo C, Parra X (2003) K-SVCR Andreu Catala. A support vector machine for multi-class classification [J]. Neurocomputing 55(9):55–77
16. Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification [J]. Adv Neural Inf Proc Syst 12(3):547–553
17. Zhao H, Rong L (2006) SVM multi-class classification based on fuzzy kernel clustering [J]. Syst Eng Electron 28(5):770–774
18. Yang Z (2008) Research progress of the kernel function support vector machine [J]. Sci Technol Inf 19:209–210
19. Jia L, Liao S (2008) Support vector machines with hyper-kernel functions [J]. Comput Sci 35(12):148–150
20. Guo L, Sun S, Duan X (2008) Research for support vector machine and kernel function [J]. Sci Technol Eng 8(2):487–489
21. Collobert R, Bengio S (2001) SVM torch: support vector machines for large-scale regression problems. J Mach Learn Res 1:143–160
22. Cheng SO, Smola AJ, Williamson RC (2005) Learning the kernel with hyper-kernel. J Mach Learn Res 6:1043–1071
23. Platt J, Burges CJC (1998) Fast training of support vector machines sequential minimal optimization. In: Sholkpof B, Smola AJ (eds) MIT Press, Cambridge
24. Tao W (2003) Kernels’ properties, tricks and its application on obstacle detection [J]. National University of Defense Technology, Changsha
25. Y Fu, D Ren (2010) Kernel function and its parameters selection of support vector machines [J]. Sci Technol Innov Herald 9:6–7

26. Chapelle O, Vapnik V, Bousquet O et al (2002) Choosing multiple parameters for support vector machines [J]. Mach Learn 46(1):131–159
27. Men C, Wang W (2006) Kernel parameter selection method based on estimation of convex [J]. Comput Eng Des 27(11):1961–1963
28. Qi Z, Tian Y, Xu Z (2005) Kernel-parameter selection problem in support vector machine [J]. Control Eng China 12(44):379–381
29. Chen PW, Wang JY, Lee HM (2004) Mode selection of SVNs using GA approach [C]. Proceedings of 2004 IEEE international joint conference on neural networks. IEEE Press, Piscataway, pp 2035–2040
30. Zheng CH et al. (2004) Automatic parameters selection for SVM based on GA [C]. Proceedings of the 5th World congress on intelligent control and automation, IEEE Press, Piscataway, pp 1869–1872
31. Liu X, Luo B, Qian Z (2005) Optimal model selection for support vector machines [J]. J Comput Res Dev 42(4):576–581
32. Wang T, Chen J (2012) Survey of research on kernel selection [J]. Comput Eng Des 33(3):1181–1186

Improved Multi-dimensional Top-k Query Processing Based on Data Prediction in Wireless Sensor Networks

Zhen-Jiang Zhang, Jun-Ren Jie and Yun Liu

Abstract Since the scale of wireless sensor networks is expanding and one single node can sense a variety of data, selecting the data of interest to users from a tremendous data stream has become an important topic. With further development in the field of WSN query, extensive research is being conducted to solve different kinds of query issues. Skyline is a typical query for multi-criteria decision making, and many applications have been developed for it. Studies of multi-dimensional top-k query processing have proven it to be more efficient than traditional centralized scheme. In some cases, variations of observed conditions, such as temperature and humidity, are related to time. Thus, we used a data- prediction method to establish the bi-boundary filter rule, which helps filter the data that may be dropped by the final result set. The bi-boundary filter rules determine whether the received or generated data will be transmitted. We analyzed the simulation results and concluded that the bi-boundary filter rules can be more energy-efficient in situations in which temporal correlation exists.

Keywords WSNs · Top-k · Data filter

Z.-J. Zhang · J.-R. Jie (✉) · Y. Liu
Department of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Jiaotong University, Beijing, China
e-mail: 10120101@bjtu.edu.cn

Z.-J. Zhang
e-mail: zhjzhang1@bjtu.edu.cn

Y. Liu
e-mail: liuyun@bjtu.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_94, © Springer Science+Business Media Dordrecht 2014

Introduction

The preliminary ideas of promoting the Internet for the use of the general population without restrictions will come to fruition, and this gradual process has validated the concept of ubiquitous computing that was first proposed by Mark Weiser [9]. In the future, people can access the Internet, acquire information, and conduct official and private business seamlessly through the processors that currently are and that will be available in almost any location. The wireless sensor network (WSN) is an expansion of the ubiquitous computing concept, and it consists of many sensors that allow us to meet the requirements of exploring, utilizing, and managing the physical world. The database theory associated with WSNs, proposed by Berkeley in 2002, heightens the prospects extensive applications of WSNs and brings the research related to WSNs one step closer to the vision of ubiquitous computing [3].

The database theory of WSNs attempts to transplant the mature database technology of the Internet to the data stream environment of WSNs. The concept is that users propose queries to the WSNs in the form of structured query language, after which the gateway server analyzes the requests and sends the queries to the sink node. The sink node broadcasts the queries to wireless sensor networks and begins the aggregation of the data. Finally, the results are generated, and the server classifies them for corresponding users.

Top-k and skyline are two popular optimal data query approaches that have common features, and both return the representative or special results to users. In many situations, we need to combine the two querying technologies to satisfy the user's preferences, [8] describe an algorithm to achieve top-k skyline query, [4] gave a novel types of skyline queries called the scored k-dominant skyline query to solve the problem in high dimensions. A framework based on the filter rules with dominant graph has been improved to solve the problem associated with multi-dimensional top-k query processing problem [1, 2]. Actually, this problem is more of an issue for the skyline approach than the top-k approach. The simulation results showed that the schemes are more effective than the traditional schemes. Considering that WSNs are task-based networks, they may be useful and effective for applications in which the observed object varies regularly or remains stable. Normally our workplace or the production workshop in a factory requires a relatively stable environment; otherwise, anomalous situation could occur and would have to be dealt with in some way. For example, the temperature of our lab for one day approximates a sine curve. In this case, we can use the predicate method to estimate future values.

The use of time series for making predictions has been researched extensively in many fields, including financial analysis, stock trend forecasting, and environmental monitoring. Reference [5] addressed the modeling of complex environmental data. Reference [6] was focused on one task associated with financial time-series analysis, i.e., forecasting future stock prices based on historical data. Reference [7] proposed prediction rules based on a hybrid model.

All of these prior research efforts inspired us to pursue the research reported in this paper.

In this paper, we used historical data to conduct predictions by time–space relativity and to optimize data retrieval and processing on the basis of [1] with keeping dominate graph locally, and attempt to reduce more energy consumption. We conducted a multi-dimensional top-k query under regular variation environmental conditions. The contributions of this paper were related to the proposals summarized below:

(1) We propose a data-dominated prediction probability to describe the possibility that the future data dominate history data.
(2) We propose a bi-boundary filter to help limit the transmission of data that match the result condition momentarily but that may not be accepted by the final result set.
(3) We simulate these bi-boundary filter algorithms and compare them to the original skyline algorithms. The simulation results will be analyzed thoroughly.

The rest of the paper is organized as follows. Second section provides background information and defines the multi-layer skyline problem briefly. Section Design of Algorithms is devoted to the multi-layer skyline query processing algorithms with bi-boundary filter rules. The simulations are discussed in fourth section, and our conclusions are presented in fifth section.

Background Information

The k-Layers Extraction Algorithm

Given two data points p = (p.x_1, p.x_2…p.x_n) and q = (q.x_1, q.x_2…q.x_n), if p has no dimensions worse than q but has at least one dimension that is better than q, then we say p dominates q, which is written as $p \succ q$ [11]. For convenience, in this paper, we used the term "better" to mean "a greater value", and this condition will not affect the correction of our algorithm or its results.

In the following graph, {P_1, P_2, P_3, P_4} is a subset that contains all of the data that cannot be dominated by other data, which is the traditional skyline scheme. The skyline scheme identifies the data that are of special concern to us. But in WSNs, we may need more information that is provided by just one skyline dataset. Then, we have the following definition:

Given a data set S and a data point p ∊ S, if p is dominated by k other data points in S, p belongs to the $(k + 1)$ layer [1]. The $(k + 1)$-layer of S is the subset that contains all of the data that are dominated by the other k data in S. (Fig. 1).

In the following picture, the 1-layer skyline is {P_1, P_2, P_3, P_4}, the 2-layer skyline is {P_7}, the 3-layer skyline is {P_5, P_6, P_9} and the 4-layer skyline is {P_8}.

We give the k-layers extraction algorithm as Algorithm 1 which will help us get the sets of top-k skyline result.

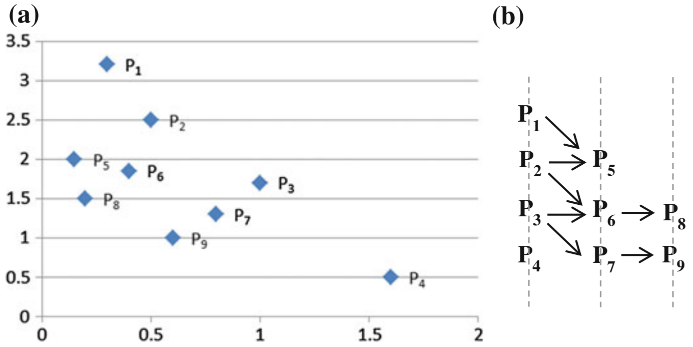


Fig. 1 An example of skyline

Algorithm 1 ***k*-layers Extraction Algorithm**

```
Require: Input:a dataset S and k
      Output: k skylines of dataset S : {L1...Li....Lk}
All data points in S are put into $Q_{candidate}$.
set iteration i= 1; $L_1, L_{2...}L_k = \emptyset$;
while $Q_{candidate} \neq \emptyset$ and $i \leq k$ do
  for evey data $d$ in $Q_{candidate}$ do
    if $d$ has less than k predecessors in DG then
      $L_i = L_i \cup \{d\}$;
      $Q_{candidate} = Q_{candidate}\ /\{d\}$;
    end if
    i = i+1;
  end for
end while
set $Q_{candidate} = \emptyset$;
return {L1...Li....Lk}
```

The results that were retrieved were in the form of strata instead of collections because it was important for us to generate the details associated with the bi-boundary.

The Problem

Users send a query to the WSN and want the k-layer skyline data during some future time T. The server must know: (1) the dimensions of the user's interest; (2) the k; and (3) the query time, T. The server allows the sink node to broadcast this query to every sensor belongs to the WSNs.

The nodes generate data when they are awake, and the data di will be packed in the structure, (di.x_1, di.x_2...di.x_m, di.Id, di.t, di.P). In this structure, di.x_1, di.x_2...di.x_m means the data in each of the dimensions, di.Id is the unique number for every node, di.t is the timestamp of these data, and di.E is the dominate expectation of the next sample, which we will introduce in detail in the next section. The data are always effective during the query time. T; this is unlike continuous query, in which the lifetime of the data is limited.

Assume that the data set D contains all the data generated during the query time, T, by the entire WSN, but the the user needs only the subset D' that contains the top-k skyline layer, $\{\mathrm{L}_1 \ldots \mathrm{L}_\mathrm{i} \ldots \mathrm{L}_k\}$. So, our target is to reduce the communication of data that are not included in the results.

The problem description is simplified and omits the situations of time sliding and query continuing. This is done just to stress the new filter rules, and the new rules can be transplanted directly to the original frame.

Design of Algorithms

Our improved algorithms stress the design of the bi-boundary filter based on the data prediction rules. We assumed that the routing tree already has been constructed and that all the sensor nodes that participated in query processing are awake.

Prediction Probability Algorithms

We use ε to designate prediction coefficients in the range of 0–1 to show the strength of the relationship between historical data and future predictions. This is a simple statistic or empirical value. This is a simple statistical or empirical value. We can test a node in a specific place for an m sampling period. If the result has n inflection points, we have $\varepsilon = (\mathrm{m} - \mathrm{n})/\mathrm{m}$, usually ε is affected by sensor quality and environmental conditions. For example, in a manufacturing shop, we can divide the environment into different areas, such as the production line, offices, storage, and laboratory. Environmental factors vary between these locations, but they are relatively stable within a given location.

In the one dimension situation, a data point d1 is generated by sensor Si during the first sampling period. In the next sampling period we assume that d2 have the half probability better than d1. We use DOM*i* (n) to describe the dominate state of dn to dn − 1, if dn dominates dn − 1, then DOMi (n) = 1; if dn is dominated by dn − 1, then DOM*i* (n) = −1, else DOM*i* (n) = 0.

$$\mathrm{P} = 1 * \frac{1}{2} + \left(\frac{1}{2}\right)\left[\frac{\varepsilon}{2} DOM\ (n)\right] + \left(\frac{1}{2^2}\right)\left[\frac{\varepsilon^2}{2} DOM\ (\mathrm{n}-1)\right] + \left(\frac{1}{2^3}\right)\left[\frac{\varepsilon^3}{2} DOM\ (\mathrm{n}-2)\right]. \tag{1}$$

After the merger:

$$\mathrm{P} = \frac{1}{2} + \frac{1}{2}\sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon}{2}\right)^i DOM(\mathrm{n}-i+1) \tag{2}$$

It is apparent that the probability of dominate status in the next time increases when the sensor Si keeps generating a series of data that dominate the old data. If ε is close to 1, that means the relationship of historical and future values is very strong. If ε is close to 1, the extreme value, so the probability will also approach 1 that predicts the new data will always dominate the old data. Note that the probability is not actually the accurate dominated probability in the real world; rather, it is a predicted probability that we designed to serve as a rough estimate.

In the 2-d situation, if sensor S_i gets two data points, d_1 and d_2, and d_2 dominates d_1, d_2 is not worse than d_1 in any two of the dimensions and gets at least one dimension better than d_1. We used x and y to note the two dimensions temporarily, so the relationship constants are εx and εy. DOMx (n) denotes the dominant status between dn and dn − 1 in dimension x, as was defined before, so as the DOMy (n). We have the probability that the dn + 1 can dominate dn as follows:

$$\mathrm{P}_x = \frac{1}{2} + \frac{\varepsilon_x}{2^2}(DOM_x(\mathrm{n})) + \frac{\varepsilon_x^2}{2^3}(DOM_x(\mathrm{n}-1)) + \frac{\varepsilon_x^3}{2^4}(DOM_x(\mathrm{n}-2)). \tag{3}$$

Similarly, we have the probability in the y dimension as:

$$\mathrm{P}_y = \frac{1}{2} + \frac{\varepsilon_y}{2^2}(DOM_y(\mathrm{n})) + \frac{\varepsilon_y^2}{2^3}(DOM_y(\mathrm{n}-1)) + \frac{\varepsilon_y^3}{2^4}(DOM_y(\mathrm{n}-2)). \tag{4}$$

So, it is obvious that the probability that d_{n+1} dominates d_n is PxPy, as shown below:

$$\mathrm{P}_x\mathrm{P}_y = \frac{1}{4} + \frac{1}{2}\sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon_x}{2}\right)^i DOM_x(\mathrm{n}-i+1) + \frac{1}{2}\sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon_y}{2}\right)^i DOM_y(\mathrm{n}-i+1) + \left[\sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon_x}{2}\right)^i DOM_x(\mathrm{n}-i+1)\right]\left[\sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon_y}{2}\right)^i DOM_y(\mathrm{n}-i+1)\right] \tag{5}$$

The formula may be somewhat complicated when conducting multi-dimensional computations. So, some simplifying modifications are discussed below.

Still using the 2-dimensional case as an example, when d_n dominates $\mathrm{d}_{\mathrm{n}-1}$, we have DOM (n) = 1. We want to use DOM (n) instead of DOM_x (n) and DOM_y (n) for simplicity. We note $\mathrm{P_xP_y}'$ as the simplified calculation. $\mathrm{P_xP_y}' \leq \mathrm{P_xP_y}$ because

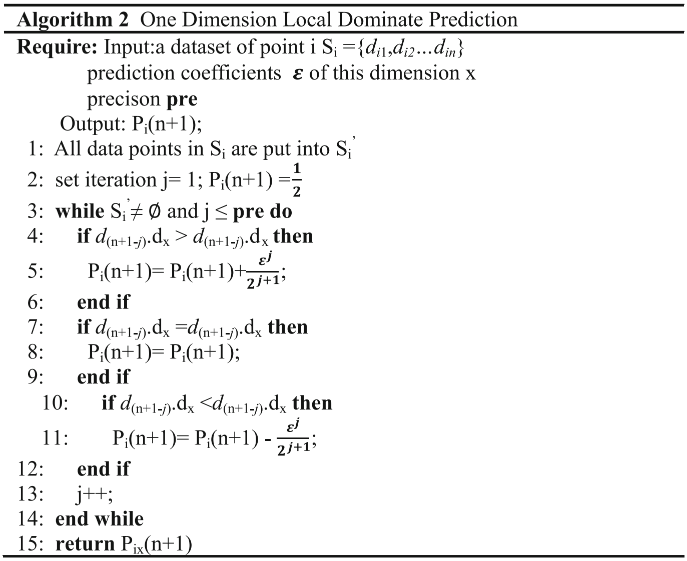
DOM (n) = 1 is an unnecessary and sufficient condition for $\mathrm{DOM_x}$ (n) = 1 or $\mathrm{DOM_y}$ (n) = 1. Then we have $\mathrm{P_xP_y}'$ as shown below:

$$\mathrm{P}_x\mathrm{P}_y' = \frac{1}{4} + \frac{1}{2}\sum\nolimits_{i=1}^{n}\left[\left(\frac{\varepsilon_x}{2}\right)^i + \left(\frac{\varepsilon_y}{2}\right)^i\right] DOM(\mathrm{n} - i + 1) + \left[\sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon_x}{2}\right)^i\right]\left[\sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon_y}{2}\right)^i\right] DOM(\mathrm{n} - i + 1)^2 \qquad (6)$$

In addition, if we think that ε_x is equal to ε_y, the formula can be simplified further. If we use ε instead of ε_x and ε_y, the formula becomes:
$\varepsilon_x + \varepsilon_y \geq 2\sqrt{\varepsilon_x\varepsilon_y}$. We set $\varepsilon = 2\sqrt{\varepsilon_x\varepsilon_y}$, so:

$$\mathrm{P}_x\mathrm{P}_y'' = \frac{1}{4} + \sum\nolimits_{i=1}^{n}\left(\frac{\varepsilon}{2}\right)^i DOM(\mathrm{n} - i + 1) + \left[\frac{\varepsilon^2\left(1 - \left(\frac{\varepsilon}{2}\right)^n\right)^2}{(2 - \varepsilon)^2}\right] DOM(\mathrm{n} - i + 1)^2 \qquad (7)$$

In three dimensions, the polynomial will be a little more complex, but when the time relationship constant ε is small, terms with high time power can be ignored. In addition, uncertainties increase when the number of dimensions increases. The predicted probability will be very small when there are more than four dimensions, and the algorithm will be inefficient. So, another problem to be solved is making the use of these concepts more efficient and effective.



Algorithm 2 One Dimension Local Dominate Prediction

```
Require: Input:a dataset of point i $S_i = \{d_{i1}, d_{i2} \ldots d_{in}\}$
        prediction coefficients $\varepsilon$ of this dimension x
        precison pre
    Output: $P_i(n+1)$;
All data points in $S_i$ are put into $S_i'$
set iteration j= 1; $P_i(n+1) = \frac{1}{2}$
while $S_i' \neq \emptyset$ and $j \leq$ pre do
    if $d_{(n+1-j)}.d_x > d_{(n+1-j)}.d_x$ then
        $P_i(n+1) = P_i(n+1) + \frac{\varepsilon^j}{2^{j+1}}$;
    end if
    if $d_{(n+1-j)}.d_x = d_{(n+1-j)}.d_x$ then
        $P_i(n+1) = P_i(n+1)$;
    end if
    if $d_{(n+1-j)}.d_x < d_{(n+1-j)}.d_x$ then
        $P_i(n+1) = P_i(n+1) - \frac{\varepsilon^j}{2^{j+1}}$;
    end if
    j++;
end while
return $P_{ix}(n+1)$
```

Bi-boundary Filter Algorithms

We used two boundaries to filter the local data. One boundary is the real boundary, which is calculated by all the data points sink node has already received so far. The other boundary is the predicted boundary, which is calculated by the exception of recently received data points at sink node.

Use bi-boundary filter will delay the data not filter by the real boundary but have a strong possibility not to be adopted by the final result set. Following is the detailed description of the bi-boundary filter calculation.

(1) *current information table kept at sink node*

We regard the sink node as unlimited with respect to energy and as providing sufficient storage. In addition to the essential data-record table, we keep a global dominant information table at the sink node. This table has four labels:

- node ID *i*;
- Exception about layer increase of node i's data last received at sink: this value is calculated by algorithm 3 when predicting filter generation.
- Correction factor η: this fixed value helps us acquire a predicted value that is closer to the actual value. It is described in more detail later.
- The count for invalid data: used to directionally update the local filter; detailed information is not provided here, but it is available in Ref. [1] (Table 1).

(2) *correction factor* η

The predicted dominant possibility is a conjecture, not the real possibility, w, as we have emphasized before. When we use this possibility to generate the predicted filter, we could perhaps encounter the situation in which the data's layer is different with our prediction, so we use η to fix it.

$$\eta_i = \frac{\Delta Layer}{E_i} \tag{8}$$

This algorithm is runing at sink node, which always has a strong energy supply and limitless memory. So this algorithm can be used efficently and quickly at sink node with its dominant graph. The generated bi-boundary consist of the fixed prediction line and the actual line. The sink node only has to send these packages to the designated nodes.

Table 1 Current information table

ID	E	η	Count
1	0.71	1.63	0
2	0.21	0.29	3
3	0.56	0.92	2
…	…	…	…

Here, we have to set a redundance value because the estimations are always decimals. We use rounded data instead of the approximated values.

(3) *Generation of the Bi-boundary Filter*

As we know, if a data point cannot be included in the result at any timestamp, this data point will never be accepted by the sink node. So the data that are filtered by the true boundary have no chance to be sent during the query time. Algorithm 3 is used to count the predicted boundary. The sink node will count $\{L_k, L'_k\}$ for every time period and respond to the nodes' requests for filter updates.

Algorithm 3 Predicted k-layer skyline Boundary

```
Require: Input:a dataset $S_{sink}$ at (n-1), $S_{sink}'$ at n
         Time correlation $\boldsymbol{\varepsilon}$ of dimension $x_1,x_2...$
         $\eta$ and k
      Output: predicte k-layer $\{L_1', L_2'... L_k'\}$
$S_{sink} = S_{sink} + S_{sink}'$
update DG in sink
caculate $\{L_1...L_i....L_k\}$ with Algorithm1
$\{L_1', L_2'... L_k'\}=\{L_1...L_i....L_k\}$
for every data $d_i \in L_j$ (j=1,2...k) do
  if j = 1then
    $E = E_i$
  else use DG count $E_{max}$;E= $E_{max}$
  if $[\eta E] \geqslant 1$then
      $L_1 = L_1 / \{d_i\}$
      if $[\eta E] \leqslant k$ then
        $L_{[\eta E]+1} = L_{[\eta E]+1} \cup \{d_i\}$
      end if
   end if
end for
return $\{L_1', L_2'... L_k'\}$
```

Simulation

We used the nodes in our laboratory to collect the temperature and humidity data from 20 different places, including the working space, the server cabinet, a meeting room, and outspace. These environments were relatively stable, but the data set was not enough for simulation. So, we generated an extensive amount of simulated data with existing data, and the maximum size of the data dimensionality was five. We classified the data to three degrees, i.e., (1) range A: $\varepsilon \sim [0.8 - 1]$, stable; (2) range B: $\varepsilon \sim [0.5 - 0.8]$, medium; (3) Range C: $\varepsilon \sim (0 - 0.5)$ frequent change. We used three methods to conduct this simulation. The first was the traditional, centralized, exact method in which all data were sent to the sink node [10]. The second method was the original scheme proposed in [1]. The third method was the improved scheme proposed in this paper.

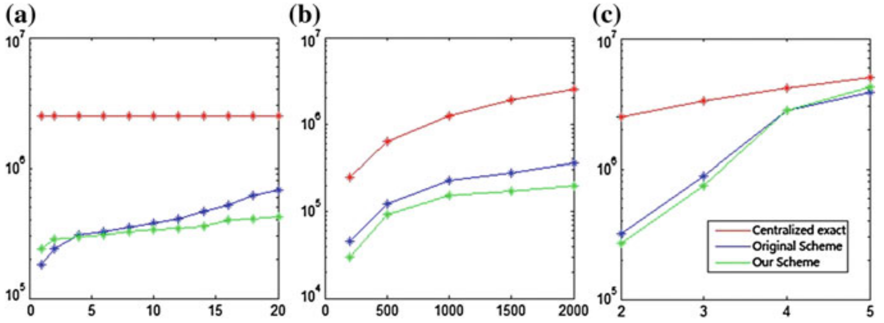
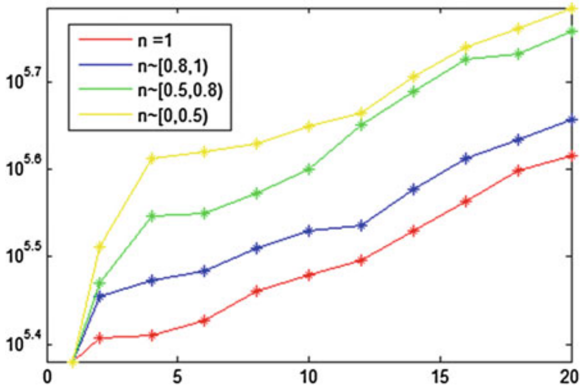


Fig. 2 The simulation about the 3 algorithms. **a** N = 2,000, D = 2, ε ∼ A; **b** D = 2, ε ∼ A k = 10; **c** N = 2,000, k = 10, ε ∼ A)

We assumed that each packet contained a single, double-precision, floating-point number (8 bytes). Accordingly, a 2-dimensional data point takes three packets to transmit, i.e., two for di.x1, di.x2, one for di.E, and one for the time-stamp di.t and ID di.id. The query time was set to one hour. We conducted the simulation in MATLAB 2009.

All of the data used in Fig. 2 came from rank A, i.e., ε ∼ [0.8,1]. In Fig. 2a, the size of the nodes is 2,000, and there are two dimensionalities. Our improved scheme cost more energy when k is less than 3, because, when k is too small, the two boundaries are not much different and the packets have one more byte than the original scheme. When k increases, the effect of the bi-boundary filter is apparent. The efficiency of our scheme increases as the value of k increases. In Fig. 2b, there are two dimensionalities, and the value of k is 10. Our scheme is approximately 20 % better than the original scheme. In Fig. 2c, the size of the nodes is 2,000, and the value of k is 10. It is apparent that the meaning of the prediction decreases as the number of dimensions increases. When the number of dimensions is more than four, the communications exceed those of the traditional scheme and the original

Fig. 3 The simulation on different η



scheme. This occurred because the results of the predicted exception algorithm become invalid as the uncertainty increases.

The simulation results in Fig. 3 show that the relationship between historical and predicted data is very strong when ε becomes closer to 1, so our bi-boundary will be more effective. That means that our scheme is better suited than the other schemes for a relatively stable monitoring environment, and it can save energy cost by reducing communications.

Conclusion

In the research reported in this paper, we applied the original scheme proposed in [1] to a stable environment. We designed a dominant prediction algorithm of node data and used it to improve the filter rules. According to the simulation results, our scheme can save energy costs by reducing communications when there is a large quantity of data, the time relativity of the data is strong, and the dimensions do not exceed four.

Since time and funds are limited at the present time, we expect to conduct additional research in this area in the future focused on (1) making the prediction more accurate and (2) taking into account other WSN data that also are related to the distribution area.

References

1. Jiang H, Cheng J, Wang D, Wang C, Tan G (2011) Continuous multi-dimensional top-k query processing in sensor networks. In: Proceedings of IEEE INFOCOM
2. Zou L, Chen L (2008) Dominant graph: an efficient indexing structure to answer top-k queries. In: Proceedings of IEEE ICDE
3. Madden S, Franklin MJ (2002) Fjording the stream: an architecture for queries over streaming sensor data. In: Proceedings of IEEE ICDE
4. Kim YS, Jung HR, Sung MK, Chung YD (2011) On processing scored k-dominant skyline queries. In: Proceedings of IEEE conference
5. Peralta J, Gutierrez G, Sanchis A (2009) Shuffle design to improve time series forecasting accuracy. In: IEEE congress on digital object identifier, pp 741–748
6. Huang Y, Wang H, McClean S (2009) Neighborhood counting for financial time series forecasting. Evolutionary computation. CEC '09. IEEE Congress, pp 815–821. doi: 10.1109/CEC.2009.4983029
7. Wang JJ, Wang JZ, Zhang ZG, Guo SP (2012) Stock index forecasting based on a hybrid model. Omega Int J Manage Sci 40:758–766
8. Sun YQ, Li Q, Chen ZY (2009) The top-k skyline query in pervasive computing environments. Pervasive Computing (JCPC), 335–338
9. Weiser M (2002) The computer for the 21st century. Pervasive Computing, IEEE
10. Madden SR, Franklin MJ, Hellerstein JM, Hong W (2005) TinyDB: an acquisitional query processing system for sensor networks. ACM Trans Database Syst 30(1):122–173
11. Wu M, Xu J, Tang X, Lee W-C (2007) Top-k monitoring in wireless sensor networks. IEEE Trans Knowl Data Eng 19(7):962–976

An Improved LMAP^{++} Protocol Combined with Low-Cost and Privacy Protection

Fei Zeng, Haibing Mu and Xiaojun Wen

Abstract With the fast development of the Radio Frequency Identification (RFID) technology, the cost and privacy of RFID systems have become an obstacle that prevents it from being deployed in a larger scale. In this paper, we introduce the attacks which the LMAP^{++} protocol suffers from briefly. We modify the LMAP^{++} protocol by importing hash function on the parameters A and B as well as setting the lowest bit of ID as "1". Analysis shows that the protocol can improve the security for the traceable and desynchronized problems with reasonable overhead.

Keywords RFID · LMAP^{++} · Privacy · Traceability · Desynchronization

Introduction

The Internet of things is called the third revolution in the information industry after computers and the Internet. And RFID is the most important technology of the Internet of things. It provides instant identification functions for objects. Despite of a good prospect, the cost and privacy problems constrain the development of the Internet of things. Hence, a perfect combination between low cost and high privacy protection has been the target of the academic circles.

F. Zeng · H. Mu (✉)
Beijing Key Laboratory of Communication and Information Systems, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: hbmu@bjtu.edu.cn

F. Zeng
e-mail: 12125056@bjtu.edu.cn

X. Wen
School of Computer Engineering, Shenzhen Polytechnic, Shenzhen 518055, China
e-mail: szwxjun@sina.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_95, © Springer Science+Business Media Dordrecht 2014

In this paper we proposed a protocol which is improved from LMAP^{++}. The hash function is used to confuse the parameters and the exchanged information which will prevent attackers from getting useful information. It is low-cost and can avoid from traceability attack and desynchronizaion attack.

The rest of the paper is organized as follows: Section Related Work is the related work about RFID protocols which focus on the effective combination between cost and privacy. Section LAMP^{++}, Traceability Attack and Desynchronization Attack is the description of traceability attack and desynchronizaion attack against LAMP^{++} protocol. Section The Improved Protocol is the detailed process of my protocol. We analyze the protocol we propose in Section Analysis.

Related Work

There are large quantities of related studies which focus on the balance of cost and privacy of the RFID systems.

Peris et al. [9] introduced a protocol which was lightweight. But it had privacy protection problems. Because of the plaintext transmission of the tag, it can't guarantee the users' privacy, such as the location.

The protocol produced by Juels et al. [10] was traceable because users that used their blocker tags would need to acquire an opt-out mechanism to protect their privacy.

Mubarak et al. [11] introduced a new idea that combined security, trust and privacy to protect the users' privacy. They applied Trusted Platform Module (TPM) to provide trusted computing. Through the interaction between the reader, tag and the TPM module, the protocol had higher security.

In [1], Peris et al. put forward a Lightweight Mutual Authentication Protocol named LMAP. Besides, they proposed an extension of this protocol—LMAP^{+}. These protocols were indeed lightweight and used only simple bitwise operations. However, it was discovered before long that the claimed security was not achieved. Later, following the LMAP designing strategy, Li [2] proposed a new lightweight protocol. Li [2] also called the proposed scheme LMAP^{+}. However, to avoid confusion with the extension of LMAP proposed by Peris et al. in [1], we call Li's scheme LMAP^{++} protocol in the rest of this paper. The LMAP^{++} protocol can be seen as a modified version of SLMAP protocol [3] which has been analyzed in [4, 5].

In [6], Masoumeh et al. used the model for traceability proposed by Jules and Weis in [7] to successfully attack LMAP^{++} and presented a desynchronization attack which has the success probability of 1/16 on each run of the protocol.

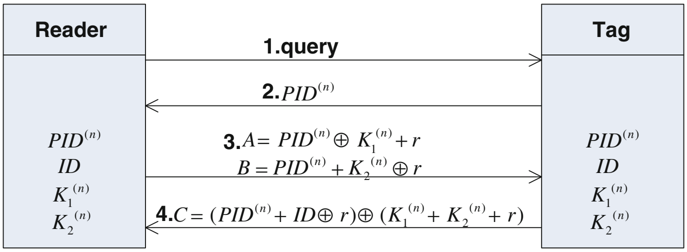

Reader
Tag
1.query
2.$PID^{(n)}$
3.$A = PID^{(n)} \oplus K_1^{(n)} + r$
$B = PID^{(n)} + K_2^{(n)} \oplus r$
4.$C = (PID^{(n)} + ID \oplus r) \oplus (K_1^{(n)} + K_2^{(n)} + r)$
$PID^{(n)}$
ID
$K_1^{(n)}$
$K_2^{(n)}$
$PID^{(n)}$
ID
$K_1^{(n)}$
$K_2^{(n)}$

Fig. 1 The process in a single run of LMAP++ protocol

LAMP++, Traceability Attack and Desynchronization Attack

LMAP++

LMAP++ is a lightweight mutual authentication protocol. It assumed the channel between the database and reader is secure. The process of the protocol is described in Fig. 1.

Notation:

$PID^{(n)}$	The pseudonym of the tag at the nth run
ID	The tag's identifier
$K_1^{(n)}$, $K_2^{(n)}$	Two secret keys at the nth run
$\oplus$	XOR operation
+	Addition mode 2^m
r	A random number
$\|\|$	Concatenation operator

The process that LMAP++ protocol runs is described as below:

1. The reader sends a query message to the tag.
2. The tag responses with a $PID^{(n)}$.

The reader looks for $PID^{(n)}$ from the table. If there is a matching option, calculate A, B as follows:

$A = PID^{(n)} \oplus K_1^{(n)} + r$

$B = PID^{(n)} + K_2^{(n)} \oplus r$

Next, passes $A \| B$ to the tag.

3. The tag extracts r from A and B:

$$r_1 = A - \left(PID^{(n)} \oplus K_1^{(n)}\right)$$

$r_2 = \left(B - PID^{(n)}\right) \oplus K_2^{(n)}$
If $r_1 = r_2$, the tag authenticates the reader. Else, the protocol will be terminated. The tag computes $C = \left(PID^{(n)} + ID \oplus r\right) \oplus \left(K_1^{(n)} + K_2^{(n)} + r\right)$ and passes C to the reader. And the tag updates $PID^{(n)}$, $K_1^{(n)}$, $K_2^{(n)}$ as follows:
$PID^{(n+1)} =^{(} PID^{(n)} + K_1^{(n))} \oplus r + (ID + K_2^{(n))} \oplus r$
$K_1^{(n+1)} = K_1^{(n)} \oplus r + \left(PID^{(n+1)} + K_2^{(n)} + ID\right)$
$K_2^{(n+1)} = K_2^{(n)} \oplus r + \left(PID^{(n+1)} + K_1^{(n)} + ID\right)$

4. The reader calculates $C' = \left(PID^{(n)} + ID \oplus r\right) \oplus \left(K_1^{(n)} + K_2^{(n)} + r\right)$

   If $C = C'$, the reader authenticates the tag. At last, the reader updates $PID^{(n)}$, $K_1^{(n)}$, $K_2^{(n)}$ as above.

Traceability Attack and Desynchronization Attack

In [6], Masoumeh et al. used the traceability model proposed by Jules and Weis in [8]. The attacking process can be summarized as this. They assumed that there were two tags which the lowest bits were respectively "0" and "1". If they could distinguish the tags after the attacker eavesdropped a whole run of the protocol, they declared LMAP^{++} had suffered from traceability attack.

In addition, they analyzed that under the probability of 1/16, attackers could desynchronize the updating of the keys through tampering with the data which the reader transferred to the tag.

The Improved Protocol

The improved protocol proposed is as below:

1. Set the lowest bit of ID as "1".
2. The reader sends a query message to the tag.
3. The tag responses with a $PID^{(n)}$.

   The reader looks for $PID^{(n)}$ from the table. If there is a matching option, calculate A, B using PID, K_1 and K_2. And then, hash A||B and pass it with r to the tag. Else, the protocol will be terminated.

4. In the tag, use PID, K_1, K_2, r to compute A_1, B_1. If H (A||B) = H (A1||B1), the tag authenticates the reader.

   In the tag, compute $C = \left(PID^{(n)} + ID \oplus r\right) \oplus \left(K_1^{(n)} + K_2^{(n)} + r\right)$ and

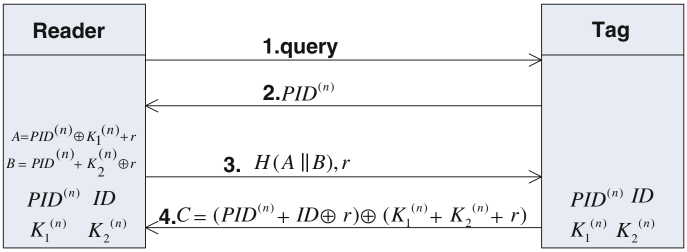


Fig. 2 The process in a single run of the improved protocol

passes C to the reader. And the tag updates $\mathrm{PID}^{(n)}$, $\mathrm{K}_1^{(n)}$, $\mathrm{K}_2^{(n)}$ as follows:

$$\mathrm{PID}^{(n+1)} = (\mathrm{PID}^{(n)} + \mathrm{K}_1(n)) \oplus \mathrm{r} + \left(\mathrm{ID} + \mathrm{K}_2^{(n)}\right) \oplus \mathrm{r}$$

$$\mathrm{K}_1^{(n+1)} = \mathrm{K}_1^{(n)} \oplus \mathrm{r} + \left(\mathrm{PID}^{(n+1)} + \mathrm{K}_2^{(n)} + \mathrm{ID}\right)$$

$$\mathrm{K}_2^{(n+1)} = \mathrm{K}_2^{(n)} \oplus \mathrm{r} + \left(\mathrm{PID}^{(n+1)} + \mathrm{K}_1^{(n)} + \mathrm{ID}\right)$$

5. The reader calculates $\mathrm{C}' = \left(\mathrm{PID}^{(n)} + \mathrm{ID} \oplus \mathrm{r}\right) \oplus \left(\mathrm{K}_1^{(n)} + \mathrm{K}_2^{(n)} + \mathrm{r}\right)$

If $\mathrm{C} = \mathrm{C}'$, the reader authenticates the tag. Then, updating $\mathrm{PID}^{(n)}$, $\mathrm{K}_1^{(n)}$, $\mathrm{K}_2^{(n)}$ as above.

The process of the improved protocol is shown in Fig. 2.

Analysis

On one hand, the desynchronization attack has a premise that $(\mathrm{PID})_0 = (\mathrm{K}_1)_0 = (\mathrm{K}_2)_0 = (\mathrm{ID})_0 = 0$ while our protocol sets $(\mathrm{ID})_0 = 1$ at first. Hence, our protocol is immune to this attack.

On the other hand, our protocol can resist from the traceability attack because the reader passes H (A||B) to the tag instead of A||B in my protocol. Under this circumstance, although the attacker can intercept A||B, he has no way to get a knowledge of the lowest bit of A or B. Furthermore, we transmit the random number r with H (A||B) at the same time. R will be generated from the random number generator every time the protocol successfully runs. Though attackers eavesdrop the channel between readers and tags, there is no approach to guess the effective information. Besides, the secret keys and the pseudonym stored in the reader and the tag will be updated synchronously, too. Therefore, attackers aren't able to distinguish the two tags which the lowest bit is separately "0" and "1".

Compared with LMAP++, my protocol adds hash operation modules to enhance security. But hash operation modules are not so complicated that we can apply it to the design of the RFID system. So, our protocol still costs little. Moreover, we solve the traceable problems that LMAP^{++} is suffering from.

As mentioned above, my protocol not only remains the low-cost advantages but also possess a more secure performance.

Conclusion

In this paper we propose an improved protocol that is based on LMAP^{++}. To fix the vulnerability, we import hash operations and encrypt the exchanging information. It fulfills the aim that finishing a good RFID protocol is a balance of low-cost and privacy preserving. However, this is not enough to put our protocol into large-scale applications. The next step of our work will to be design a better RFID protocol in the future.

Acknowledgments This work is supported by National Natural Science Foundation of China under Grant 61201159, and Fundamental Research Funds for the Central Universities under Grant 2012JBM016. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Peris-Lopez P, Hernandez-Castro JC, Estevez-Tapiador JM, Ribagorda A (2006) LMAP: a real lightweight mutual authentication protocol for low-cost RFID tags. In: Proceedings of RFIDSec06 workshop on RFID security, Graz, Austria, pp 12–14
2. Li T (2008) Employing lightweight primitives on low-cost RFID tags for authentication. In: Proceedings of vehicular technology conference fall, pp 1–5
3. Li T, Wang G (2007) SLMAP—a secure ultra-lightweight RFID mutual authentication protocol. In: Proceedings of Chinacrypt, vol 07, pp 19–22
4. Hernandez-Castro JC, Estevez-Tapiador JM, Peris-Lopez P, Clark JA, Talbi E-G (2009) Metaheuristic traceability attack against SLAMP, an RFID lightweight authentication protocol. In: Proceedings of 23rd IEEE international symposium on parallel and distributed processing (23rd IPDPS'09), workshop on nature inspired distributed computing, pp 1–5
5. Hernandez-Castro JC, Estevez-Tapiador JM, Peris-Lopez P, Clark JA, Talbi E-G Metaheuristic traceability attack against SLAMP, an RFID lightweight authentication protocol. Int J Found Comput Sci
6. Safkhani M, Bagheri N, Naderi M (2011) Security analysis of LMAP^{++}, an RFID authentication protocol. In: Proceedings of 6th international conference on internet technology and secured transactions, 11–44 Dec 2011, Abu Dhabi, United Arab Emirates
7. Juels A, Weis SA (2007) Defining strong privacy for RFID. In: Proceedings of IEEE computer society, PerCom workshops, pp 342–347
8. Juels A, Weis SA (2007) Defining strong privacy for RFID. In: Proceedings of IEEE PerCom' 07, pp 342–347
9. Peris-Lopez P, Lee LT, Li T (2008) Providing stronger authentication at a low-cost to RFID tags operating under the EPCglobal framework. In: Proceedings of IEEE/IFIP international

symposium on trust, security and privacy for pervasive applications—TSP'08, Shanghai, China, pp 159–166

10. Juels A, Rivest R, Szydlo M (2003) The blocker tag: selective blocking of RFID tags for consumer privacy. In: Proceedings of conference on computer and communications security—ACM CCS, USA, pp 103–111
11. Mubarak MF, Manan J, Yahya S (2011) A critical review on RFID system towards security, trust, and privacy (STP). In: Proceedings of 2011 IEEE 7th international colloquium on signal processing and its applications

Empirical Analysis of User Life Span in Microblog

WeiGuo Yuan and Yun Liu

Abstract The aim of this work is to study two kinds of user life spans and their connection to the distribution of followers, friends and statuses in microblog. Both the user activity spans and user age approximately follow a two-part exponential distribution. Moreover, the users' average number of followers and statuses increases linearly with the active span, but the average number of friends does not change significantly during the life span. We plot the distribution of users, followers, friends, and statuses as a cumulative sum to obtain a strict power-law form, which indicates an allometric growth phenomenon. These new findings show that the users' production capacity is consistent and with self-similar growth ability in different user activity spans. We argue that the scale effect of user count development is the reason for the allometric growth phenomenon in microblog.

Keywords Life spans · Power-law · Allometric growth · Microblog

Introduction

Online social networks (OSN) have changed how we communicate. Research on OSNs is very popular and highly interdisciplinary. Some researchers focus on the network topological structure [1, 2]. Most OSNs are scale-free, which means that

W. Yuan · Y. Liu (✉)
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing, China
e-mail: liuyun@bjtu.edu.cn

W. Yuan
e-mail: ywg@cnnic.cn

W. Yuan
Computer Network Information Center, Chinese Academy of Sciences, Beijing, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_96, © Springer Science+Business Media Dordrecht 2014

the node degrees follow a power-law distribution. OSNs provide useful information for the study of user behavior and human dynamic. Recently, there have been an increasing number of studies in this active field [3, 4]. In particular, empirical analyses of user behavior in a variety of OSNs have received extensive attention [5, 6].

Microblog, a new kind of OSN applications in the age of Web 2.0, has attracted much attention from researchers studying user behavior characteristics. Kwak et al. [7] studied the correlation between the number of users' friends, followers, and statuses. Yan et al. [8] found that the time interval of two consecutive publishing statuses followed a power-law distribution. These studies are important to understand the structure and evolution of OSNs.

Allometric growth is widespread in many natural and social phenomena [9]. For example, growth in urban populations and areas, volume, and energy consumption present a power-law relationship in the growth process. Recently, Bettencourt et al. [10] found an allometric growth relationship between consumption and population size, indicating that the emergence of a city makes people get together in order to consume less energy and create more output. Similarly, Wu et al. [11] argued that the total human online activity grows faster than the active population in OSNs, indicating an allometric growth phenomena.

Sina Weibo, the largest microblog website in China, attracted over 500 million registered users as of December 2012. Our paper focuses on the relations between user life span and user characteristics in Sina Weibo. This paper offers rich empirical materials to improve understanding of user behavior and human dynamic features in OSNs.

The paper is organized as follows. Dataset and Measurement Methodology describes the data sets and methods. User Life Span Data Analysis presents two kinds of user life span statistics: user activity span and user age. User Profile Features and Growth with Activity Spans analyzes the distribution of some user features with the same user activity spans and examines the allometric growth phenomenon with different user activity spans. Finally, we provide conclusions with a brief look at future works in Conclusions.

Dataset and Measurement Methodology

In this paper, we collected a dataset based on the Sina Open API interface to enhance accuracy and efficiency. By using the snowball crawling strategy, we selected the author as a starting point and then collected his detailed information as well as the profile information of his friends and followers. We obtained 881,146 user profiles from February 9, 2012 to February 26, 2012. The basic attributes of users include user ID, registration time, and the number of followers, friends, and statuses. At the same time, we collected users' latest status information including the publishing time.

In order to study the user life span distribution, we draw from complex networks theory and statistical theory. To facilitate discussion, we introduce the complementary cumulative distribution function (CCDF). The power-law distribution CCDF can be defined as follows:

$$F_x(x) = P(X > x) \sim x^{-\alpha} \tag{1}$$

User Life Span Data Analysis

In this section, we analyze two kinds of users life span statistics collected in our dataset:

- Activity span $T_{Activity}$, which is the time between the creation of a registered user t_{reg} and the latest time of the user published statues t_{status}, namely

$$T_{Activity} = t_{status} - t_{reg}. \tag{2}$$

- Age T_{Age}, which is the time between the creation of a registered user t_{reg} and when it is collected $t_{collect}$, namely,

$$T_{Age} = t_{collect} - t_{reg}. \tag{3}$$

User activity span means how long the user has been active, while user age means how long the user has existed since the user registration. As Sina Weibo was launched in August 2009, and the earliest user registration time collected is August 14, 2009. The dataset was collected in February 2012; the user latest status time collected was on February 26, 2012. The longest user activity span and user age is about 30 months. More detailed information on the user time of the dataset is shown in Table 1.

Table 1 User time satistics results

User time type	Min (in days)	Max (in days)
$t_{collect}$ (user collected time)	09-Feb-2012	26-Feb-2012
t_{reg} (user registration time)	14-Aug-2009	26-Feb-2012
t_{status} (user lastest status time)	27-Sep-2011	26-Feb-2012
$T_{Activity}$ (user activity span)	0	914
T_{Age} (user age)	0	926

User Activity Span Distribution

The user activity span, one of the most important statistics, describes the length of time that the user account has been active. As shown in Fig. 1, the CCDF distribution of user activity spans (in months) follows an exponential distribution, with more than 70 % of users who have been active less than 15 months. The distribution can be divided into two sections: the first part with activity spans less than 15 months, which obeys a linear exponential distribution ($R^2 = 0.98$), and the second part with activity spans of more than 15 months, which follows a parabola curve exponential distribution ($R^2 = 0.99$) in log-scale. Ribeiro et al. [12] found a similar result with Myspace, which is one of the largest OSNs in the world.

User Age Distribution

User age describes the growth of the number of users. Figure 2 shows the distribution of user ages (in months). Similar to the activity spans, the distribution of user ages is approximately an exponential distribution. Less than 20 % of the total number of accounts are older than 18 months. They were created during Sina Weibo's early years and experienced quick exponential growth. The remaining 80 % of accounts are newer than 18 months and have been on a slower exponential growth. Huberman [13] believes that the WWW sites' exponential growth can be explained by the number of pages on its website to produce power-law distributions. Thus, these two modes of growth in the user age may be the reason behind the double power-law distribution [14] of the followers count in Sina Weibo accounts.

Fig. 1 The CCDF distribution of user activity spans ($T_{Activity}$)

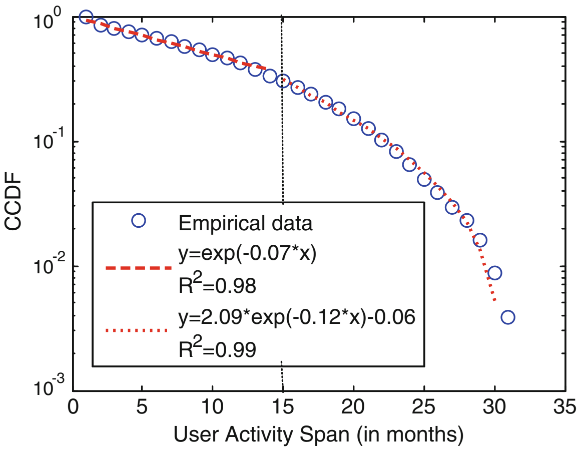
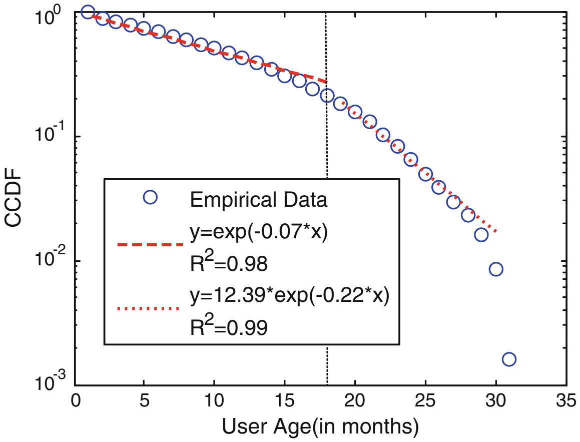


Fig. 2 The CCDF distribution of user age (T_{Age})



User Profile Features and Growth with Activity Spans

User Profile Features Distribution with Activity Spans

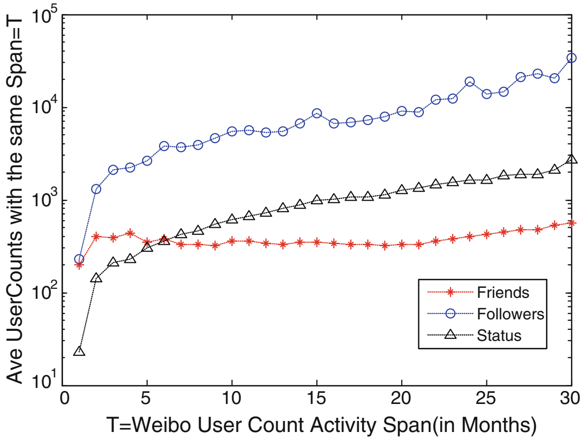
A user profile on Sina Weibo displays the user name, the number of followers and friends, and the number of statuses. In this section, we focus on the impact of user activity spans on the profile features distribution, which includes followers count, friends count, and statuses count. While the user profile features distribution of OSNs has been extensively studied in the literature [7, 14], we show some statistical properties that have been excluded from previous works.

In Fig. 3, we plot the average counts of three user profile features with the same activity spans. The number of statuses and friends increases exponentially in the initial stage (1–2 months), indicating that the new user published statuses with high interest. Thus, a new user account is more likely to attract other users' interests. Over time, the average number of followers and statuses changes to increase linearly and the growth slows down, indicating that the users' status interests and the ability of attract fans declines. The average number of user friends does not change over time.

User Allometric Growth with Activity Spans

In this section, we focus the allometric growth characteristics in the microblog. The number of statuses represents the production capacity for the information; the number of friends represents the user's ability to establish a relationship; and the number of followers represents the user's ability to attract other users' attention. We sorted users with different activity spans (in days) and gradually added the user

Fig. 3 Distribution of average user profile features with activity spans



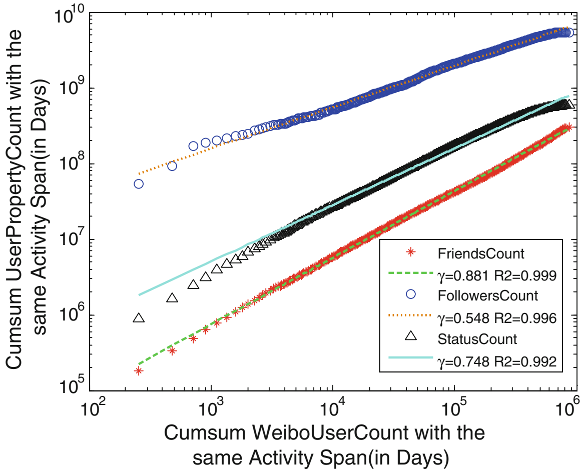
count in a larger active span. Then, we calculated the sum of the production capacity of the added users. Figure 4 shows the distribution of cumulative sum of user count and three user features count, which represent users' information production capacity. All of these user feature counts increase gradually with the user count and obtain a strict power-law form. The goodness-of-fit R^2 is greater than 0.99, and only the power-law exponent is different.

These relationships can be condensed into a unified expression as follows:

$$P(t) \propto N(t)^{\gamma} \tag{4}$$

where $P(t)$ means a certain cumulative production capacity of the network at different times of the users', $N(t)$ is the cumulative number of users of the network in different periods, and γ is the regression coefficients of the power-law. The

Fig. 4 Accumulation distribution of user count and user profile features count with same activity spans



users' production capacity is very consistent in different active spans; this includes creating the relationship, acquiring others' attention, and producing information. The whole production growth is a non-stop and self-replicating process, and users in different stages show self-similar growth characteristics. The size-dependent distribution reflects the scale-free feature and allometric growth characteristics in different user activity spans.

When $\gamma < 1$, this means negative allometric growth. In Fig. 4, we find the exponent value of the power-law as $\gamma_{followers} < \gamma_{status} < \gamma_{friends} < 1$, which means that the user's average production capabilities, including information production, relationship creation, and interest attraction, are increasing more than the rate of the user count growth. The user's ability to attract interest grows the fastest, followed by the user's ability to create information. Finally, the user's ability to create friend relations grows relatively slowest. We argue that this phenomenon is consistent with the scale-effect of user count development. Namely, when an increasing number of users join a microblog, the user productivity improves quickly.

Conclusions

In this work, we studied two kinds of user life span in Sina Weibo and its connection to the distribution of the number of followers, friends, and statuses. User activity span and user age could be well approximated by two sections of exponential distribution. These finding sheds light on the double power-law distribution of followers, friends, and statuses. We also found an allometric growth phenomenon in the cumulative sum of the user count with the user's status, friends, and followers, all of which obtain a strict power-law form. These new findings show that the users' production capacity is consistent and that users show self-similar growth ability in different activity spans. We argue that the user's average production capabilities grow faster than the rate of user accounts. This can be explained by the scale-effect of user count development. In future works, we will introduce an evolution model based on user behavior and interest to explain the allometric growth characteristics described in this paper.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ACM, San Diego, California, USA, pp 29–42
2. Ahn Y-Y, Han S, Kwak H, Moon S, Jeong H (2007) Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th international conference on world wide web, ACM, Banff, Alberta, Canada, pp. 835–844
3. Jiang J, Wilson C, Wang X, Huang P, Sha W, Dai Y, Zhao BY (2010) Understanding latent interactions in online social networks. In: Proceedings of the 10th annual conference on Internet measurement, ACM, Melbourne, Australia, pp 369–382
4. Benevenuto F, Rodrigues T, Cha M, Almeida V (2012) Characterizing user navigation and interactions in online social networks. Inf Sci 195:1–24
5. Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, Philadelphia, PA, USA, pp 611–617
6. Grabowski A (2009) Human behavior in online social systems. Eur Phys J B: Condens Matter Complex Syst 69:605–611
7. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, ACM, Raleigh, North Carolina, USA, pp 591–600
8. Yan Q, Yi L, Wu L (2012) Human dynamic model co-driven by interest and social identity in the MicroBlog community. Phys A 391:1540–1545
9. Gayon J (2000) History of the concept of allometry. Am Zool 40:748–758
10. Bettencourt LM, Lobo J, Helbing D, Kühnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. Proc Natl Acad Sci 104:7301–7306
11. Wu L, Zhang J (2011) Accelerating growth and size-dependent distribution of human online activities. Phys Rev E 84:026113
12. Ribeiro B, Gauvin W, Liu B, Towsley D (2010) On MySpace account spans and double Pareto-like distribution of friends. In: Proceedings of INFOCOM IEEE conference on computer communications workshops, pp 1–6
13. Huberman BA, Adamic LA (1999) Internet: growth dynamics of the world-wide web. Nature 401:131
14. Fan P, Li P, Jiang Z, Li W, Wang H (2011) Measurement and analysis of topology and information propagation on Sina-Microblog. In: Proceedings of IEEE international conference on intelligence and security informatics (ISI), Beijing, pp 396–401

Study and Implement of Topology Analysis Based on Hyper-Nodes in GIS

Li Zhang

Abstract In order to take full advantage of the topology analysis mechanism provided by the Geographical information system (GIS) platform software, the graphical topology method is advised to express the topology in pipe network GIS. This paper puts forward the concept of hyper-node, which has two graphic objects which each belongs to different graphics worlds. It can help implement the topology analysis across different graphics worlds. Finally, the power supply range analysis of electrical equipment (such as transformers in substation) in electric power GIS is implemented, which will validate the feasibility of the topology analysis method based on hyper-nodes.

Keywords Geographical information system (GIS) · Topology analysis · Hyper-node · Internal world

Introduction

There are many Geographical Information System (GIS) constituted by pipe network facilities such as power lines, water pipes, gas pipe and so on. The network topology analysis is one of the most important functions in pipe network GIS. GIS platform software (such as GE Smallworld, ArcInfo) provides some objects or controls with powerful topology analysis capability [1]. But it does not provide the network analysis across different graphics world, such as the topology analysis in electric power GIS from switch inside substation to feeder-lines outside which do not exist in the same graphics system [2].

L. Zhang (✉)
School of Computer Engineering, Shenzhen Polytechnic, Shenzhen,
518055 Guangdong, China
e-mail: zhangli@szpt.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_97, © Springer Science+Business Media Dordrecht 2014

Graphical Representation and Topology in GIS

In the pipe network GIS, the main window usually displays the geographical layout map whose main task is to reflect the geographical location and connection relationship of the pipeline network in the real world. There are many methods to express the topology in GIS [3]. One of them is the graphical topology method. This method combines the graphic expression with the topological relations, which requires the graphical position of pipeline facilities must be precise. The lines or equipments with the connection relationship should be really connected together in GIS. (i.e. the two geometries must overlap or cover on the coordinate position), otherwise it means that they do not connect topologically with each other [4].

In the graphical topology method, the maintenance of the topology is completely contained (hidden) in the maintenance of graphics data. So topology analysis can completely depend on the mechanisms of topology analysis provided by the GIS platform software. As a result, GIS application development workload becomes smaller.

Creating a Hyper-Node

In order to connect the feeder lines outside and the equipments inside substations topologically, we need to create such a "hyper-node" object, which has two geometries existing in two different graphics systems (called "Graphics World" in GE Smallworld). For example, *hypernode* class defined in Magik (the programming language of GE Smallworld) code is as follows:

```
def_slotted_exemplar( :hypernode, {}, { :rwo_record } )
$
_method hypernode.new_between_worlds( a_view,coord1,
world1, coord2, world2)
    _dynamic !current_dsview! << a_view
    _dynamic !current_world!
    drec << _self.target_collection.new_detached_record()
    new << _self.target_collection.insert( drec )
    !current_world! << world1
    new.make_geometry( _self.hypernode_info[:pin1], coord1
)
    !current_world! << world2
    new.make_geometry( _self.hypernode_info[:pin2], coord2
)
    >> new
_endmethod
$
```

The following Magik code is used to create a class named as *elec_hypernode* in the electric power GIS. It is a subclass of *hypernode* class. A shared constant named as *hypernode_info* should be defined. According to the code, the table name in Smallworld database is *elec_hypernode* and two point geometry field name are *pin1* and *pin2*. Another shared constant is *associated_grs* including some graphics worlds which the geometries of *elec_hypernode* may belong to. In other words, an *elec_hypernode* object has two point geometries belonging to different graphics worlds. One is *pin1* which belongs to the geographic layout map, the other is *pin2* which belongs to the substation internal world.

```
def_slotted_exemplar (:elec_hypernode,{},:hypernode)
$
elec_hypernode.define_shared_constant(:hypernode_info,
    property_list.new_with(
       :name, :elec_hypernodes,     # table name
       :pin1, :pin1,                # 2 point geometries
       :pin2, :pin2
       ), _false)
$
elec_hypernode.define_shared_constant(:associated_grs,{
       :gis,             # geographic layout map
       :subinternal,     # substation internal diagram
       }, _false)
$
```

Topology Analysis Based on Byper-Nodes

Here is an example to explore how to achieve the topology analysis based on hyper-nodes. In electric power GIS, the power range analysis of the substation equipments (e.g. a transformer inside substation) is a topological track, whose goal is to get all objects whose power is supplied by the specified station equipment.

In reality, the overhead lines or cables outside are connected to the switch inside substations to complete the power transmission. When you create a hyper-node, the geometry *pin2* should be connected to a switch inside substation and the *pin1* should be connected with the end point of an overhead line or cable.

It is more complex to express the topological relations of substation equipments, such as buses, transformers, breakers and switches. In substation internal world in GIS, the bus object is a linear object (*chain*). The other objects are *two_pin_device* objects, which have two point geometries named as *point1* and *point2*. The shape of *two_pin_device* objects are drawn based on the coordinates of the two points as a reference by the appropriate Magik code. So the inside geometry (i.e. *pin2*) of the hyper-node must be connected with the *point1* or *point2* of a switch inside the substation when a hyper-node is created.

In order to implement the power supply range analysis of substation equipments such as transformers, it is necessary to define a class named as *analyse_sub_device_tool* inherited from the *network_follower_menu* class. It is an operating dialog in which users can execute topology analysis. As a subclass of *network_follower_menu*, the *analyse_sub_device_tool* object will have a property named as follower, which is actually a *network_follower_manager* object. So you can easily to analyze the topology by calling the methods defined in *network_follower_manager*, such as *trace_out()* and *shortest_path ()* method.

```
def_slotted_exemplar( :analyse_sub_device_tool,{
    { :actual_start_point,   _unset},
    { :selected_sub_device, _unset},
    { :trace_results, _unset,    :readable},

    ……

    },:network_follower_menu )
$
```

Several properties are added when *analyse_sub_device_tool* class is defined. The *selected_sub_device* property is used to save the currently selected station equipment and the *trace_results* property will include all objects involved in tracking. The *actual_start_point* property is used to save the point geometry of the selected station equipment which is the actual start point of the track for the lower-voltage side. In addition, it includes a custom method named as *analyse_sub_device ()*, whose main task is to call the custom method *trace_down_from_subdevice()* defined in the *network_follower_manager* class.

```
_method analyse_sub_device_tool.analyse_sub_device()

   ……

   .actual_start_point<<.follower.trace_down_from_subdevi
ce(.selected_sub_device,_true,_true)

   ……

_endmethod
$
```

The idea of defining custom method *trace_down_from_subdevice()* can be descibed as follows: firstly the voltage of the selected equipment (e.g. a breaker) inside substation is saved to a variable named as *current_voltage*. Then the stop condition *stop_predicates* is set as *feeder_switch* with the status as *OFF* and breakers with the status as *OFF*. The parameter *stop_nodes* is set as the geometry *point2* of the selected station equipment. The *trace_out()* method defined in *network_follower_manager* object is called to execute topological track in one direction. The *trace_in_lower_voltage_direction?()* method will be called to determine whether it is the track for the lower-voltage side or not. If not true, the *trace_out()* method will be called again to execute topological track in the other direction while the parameter *stop_nodes* is set as the geometry *point1* of the selected station equipment. As a result, all objects whose power is supplied by the selected station equipment are saved in the property named as *trace_results* of *analyse_sub_device_tool* class. The Magik code of *trace_down_from_subdevice ()* method is as follows:

```
_method
network_follower_manager.trace_down_from_subdevice(a_subd
evice,_optional draw?)
    draw?<<draw?.default(_false)
    through_hypernode?<<through_hypernode?.default(_true)
    a_stop_predicates_group<<hash_table.new_with(:breaker,
{predicate.new(:on_off_status,:eq,"off")},
            :feeder_switch,{predicate.eq( :normal_status,
"on" )})
    .draw_trace? << _false
    current_voltage << a_subdevice.voltage.as_number()
    a_start_point<<a_subdevice.point1
    _self.trace_out(a_subdevice.point1.node,
        :via_node_rwo?,_true,
        :stop_predicates,a_stop_predicates_group,
        :stop_nodes,{a_subdevice.point2.node})
    _if
_self.trace_in_lower_voltage_direction?(current_voltage)
_is _false
    _then
        a_start_point<<a_subdevice.point2
        _self.trace_out(a_subdevice.point2.node,
            :via_node_rwo?,_true,
            :stop_predicates,a_stop_predicates_group,
            :stop_nodes,{a_subdevice.point1.node})
    _endif
    _return a_start_point
_endmethod
$
```

In *trace_in_lower_voltage_direction?()* Method, it checks the direction of the topology tracking by the following conditions: (1) If there is not any one breaker in the tracking result set, this is the track for the lower-voltage side. (2) If there exist one (or more) breaker in the tracking result set, and the highest voltage of the breakers is equal to the voltage of the check point, it is still the track for the lower-voltage side. Otherwise it is the track for the high-voltage side.

Conclusion

Because GIS platform software itself does not provide the topology analysis across different graphics worlds [5], hyper-node objects with two geometries are created to connect topologically two objects inside and outside station house in GIS. In this way, topology analysis can reflect the topology relationship in reality, such as the connection between the feeder line outside and the switch inside the substation. The hyper-nodes make it easy to topology tracking according the connection rules of pipe network. At the same time the topology analysis mechanism in GIS will play a greater role in the applications of topology relationship.

References

1. Lin F, Guo B, Qian W (2011) Common electric power grid GIS platform oriented architecture for electric power grid GIS applications. Autom Electr Power Syst 35(24):63–67
2. Liu W, Zhang J, Liu N, Zhang L (2008) Design of a distribution network intelligent planning platform based on COTS. Automation of Electr Power Syst 32(12):48–51
3. Zhang K (2008) Construction of distribution network GIS based on SmallWorld. Comput Era (1):38–41
4. Qun Y, Wei L, Guonian L (2003) Topology analysis of distribution management system based on geographical information system. Autom Electr Power Syst 27(18):80–82
5. Wu Q, Chen T, Su Y (2005) Building of spatial topological relationship based on mapobjects. Comput Simul 22(1):73–75

A New Lightweight RFID Grouping Proof Protocol

Ping Huang, Haibing Mu and Changlun Zhang

Abstract Radio Frequency Identification (RFID) is a key technology for the Internet of things. With the spread of RFID system, the security and privacy related to RFID plays a vital role in applications. At present, EPC C1G2 tags have been widely spread in the commodity retail, pharmaceutical sector, medicine distribution and many other fields. This kind of tags is passive, which means that they have limited power and memory space. Under these conditions, we present a new lightweight protocol for the RFID system with two tags and one reader. In the protocol, we develop a new function to realize the exchange between bits of a vector, which uses fewer resources in computing. In addition pieces of analysis show that the protocol reaches the capabilities of the grouping proofs and is resistant to a variety of attacks.

Keywords RFID · Lightweight · Grouping proof protocol · Privacy · Security

Introduction

Radio Frequency Identification (RFID) is a key technology of the Internet of things. An RFID system is consist of three parts, which are RFID tag, RFID reader and a backend server. RFID tags are used for preserving information of things and

P. Huang · H. Mu (✉)
Beijing Key Laboratory of Communication and Information Systems, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: hbmu@bjtu.edu.cn

P. Huang
e-mail: 12120086@bjtu.edu.cn

C. Zhang
Science School, Beijing University of Civil Engineering and Architecture,
Beijing 100044, China
e-mail: zclun@bucea.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_98, © Springer Science+Business Media Dordrecht 2014

can be sorted into three types: active tags, semi-passive tags and passive tags. Every active tag is powered by a battery, which support the tags maintain their memories and transmitting signals. Its computing power is the strongest of all. Semi-passive tags also have a battery within them but it is just used for maintain the information in the chip. Passive tags, the most widespread kind of all, without any devices to offer power, need non-volatile memory. Moreover, because of its limited ability and transmission distance, we should design special protocols for this kind of tags. This paper is based on the passive tags, considering the power limited problem.

At present, RFID is widely applied in scientific research, business, medical and many other industries. Presenting the medicine distribution of the inpatients for instance, accuracy and correctness are very important for the hospitalized patients. Only reaching the required standards can the hospitals ensure the security of patients. In other words, if privacy and security couldn't be promised during the process of distribution, a malicious adversary will prevent or interfere the corresponding medicine's being distributed to the right patients by forging information or any other means.

Aiming at the above questions, our paper presents a new lightweight RFID grouping proof for two tags. Using our protocol in medicine distribution, privacy and security can be guaranteed. There is a shared secret key between each pair of a tag and a reader. Our protocol is efficient and useful for three accounts. Firstly, considering that the EPC C1G2 tags have limit power and low computation capacities, we define a new function Con to replace the use of Hash function, which can also reach confusion and diffusion. Secondly, taking privacy and security into account, there are shared secret keys between the tag and reader. The shared information will not be transmitted in the channel during the protocol run. The last one, as for two tags in each system, information computed by the two tags depends on each other. Synthesizing the above three aspects, our lightweight protocol is efficient and secure.

The rest of this paper is organized as follows. In section Related Works, the related works are reviewed. In section Model, a proper adversary model is presented. In section A Novel Protocol, the protocol is described in details. In section Analysis, we analyze the cost and security of our protocol. In section Conclusion, we give a conclusion.

Related Works

The first approach to protect RFID users' privacy is the implementation of "Kill" [1]. That means the Reader send a command of kill the tag and the tag will be inactive forever. In this way, privacy can be well protected from adversary's scanning the tag illegally. But the tag will never work anymore. Afterwards, according to the thought of cryptography presented by Shannon, amount of researchers come up with many RFID protocols for privacy and security of RFID system. Such as the Hash-Lock protocol [2, 3]. Weis S.A. et al. [2] also described a new research direction in RFID for researchers with cryptography. An randomized Hash-Lock protocol [4] was put forward, and the authors said that low-cost Radio Frequency Identification (RFID) systems would become pervasive in our daily lives when affixed to everyday consumer items as “smart labels”. It predicts low-cost is important for RFID system’s being applied in our daily life. The hash-chain protocol [5] can ensure forward-secure privacy, because hash-function has one-way property. Hash-based enhancement of location privacy for RFID devices using varying identifiers [6] and so on [2–6] are based on Hash that cost more than some protocols using bitwise operations and didn’t consider about the limited resources in EPC C1G2 tags. Yun Tian et al. [7] created a new lightweight RFID authentication protocol with a new definition of Permutation. Kai Fan et al. [8] presented a lightweight RFID authentication Scheme to achieve efficiency. Erik-Oliver Blass et al. [9] showed a novel framework for lightweight RFID authentication protocol without using of Hash or any other complex functions. Erik-Oliver Blass et al. [10] provided a provably privacy-preserving protocol that applied for passive tags. [7–10] can decrease cost of the protocol. Roberto Di Pietro and Refik Molva [11] described two schemes that can be used independently to enhance the security of the RFID protocols. However, researches on RFID grouping authentication protocols have not been mature. The concept of “Yoking proof” was put forward by Jules [12] firstly. Soon afterwards, Saito and Sakurai [13] improved the former one with the use of timestamp and produced their protocol named “grouping proof”. Their approach can solve some problems and make up some of the vulnerabilities. Above most of the existing protocols for RFID group authentication, Pedro Peris-Lopez et al. [14] pointed the flaws in some of the existing grouping proofs and summarize a guideline for the design of this kind of protocol. Piramuthu [15] raised a new improved group proof model, but it can be suffered from attack on the communication between the tags and reader and the attack by disassembling the tags. Most of the approaches have limits, either high cost or cannot be used for more than one tags’ system.

Model

This paper is based on the EPC C1G2 tags, and can be used for medicine distribution for inpatients. So in this section, some proper assumptions are presented about the system and adversary as follows.

System Model

1. A pallet presents a hospitalized patient. In addition, Tag_A means the medicine to be distributed and Tag_B is the corresponding pallet in our protocol.

2. Every tag stores an ID and a secret key within its memory, while the reader has all the IDs and keys for the protocol.
3. A tag can operate three kinds computation, rotation Rot(*,*), exclusive or XOR(⊕) and function Con.
4. Through our protocol, a proof P_{AB} is finished and submitted to the backend verifier to that the two tags participate in the protocol simultaneously.

Adversary Model

1. A adversary can eavesdrop and intercept what reaches him. Furthermore, he can also send counterfeit information to any reader or tags.
2. As ISO14443 required, the distance between reader and tag is no more than 4 inches. Thus we assume that the adversary cannot stand between the reader and tags and he is out of the tags'communication range. Under the condition, an adversary can eavesdrop the information sent by the reader but can't eavesdrop or intercept the messages sent by the tags.

A Novel Protocol

For passive RFID tags, limited power is a vital problem for protocol designers. In order to reduce the amount of calculation and power needed, we define a novel function Con to replace the Hash used in most of RFID authentication protocols. At the same time, we just use other two kinds of bitwise operation, bitwise XOR(⊕) and Rot(*,*). There are two variables in Con, for example X, Y. We rearrange the vector X according to Y.

Definition Suppose there are two l-bit vectors X, Y, where
$X = x_1x_2...x_l$, where $x_i \in \{0, 1\}$ and $i = 1, 2, ..., l$;
$Y = y_1y_2...y_l$, where $y_j \in \{0, 1\}$ and $j = 1, 2, ..., l$;

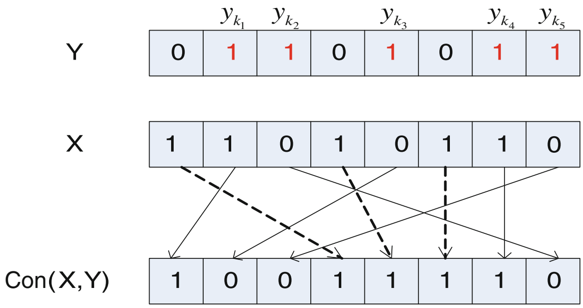
Moreover, the Hamming weight of Y is m, that means $\sum_{j=1}^{l} y_j = m$ and $y_{k_1} = y_{k_2} = ...=y_{k_m}$ and $y_{k_{m+1}} = y_{k_{m+2}} = ...=y_{k_l}$ (k_i is discontinuous, i.e. y_{k_i} just presents one of all the y_j, that equals to 0 or 1.).

Thus, $Con = x_{k_1}x_{k_3}...x_{k_{m+1}}x_{k_{m+2}}...x_{k_l}...x_{k_4}x_{k_2}$.

Example (Figure 1) X = 01101011, Y = 11010110, where m = 5, and Con(X,Y) = 10011110.

There are three kinds of entities. They are RFID tags (Tag_A, Tag_B), RFID reader and backend server. We suppose the channel between the reader and the server is secure. Under the ISO14443, we assume that the adversary cannot carry out any attack on the system between the reader and the tag on the physical. That

Fig. 1 An example of function con



also means the distance between malicious adversary and tag is longer than that between the tag and its reader. So we can adjust the tag's transmission power to keep the adversary out of the signal range.

In our protocol, every tag keeps a shared secret key k_i and its identity ID_i. Every server stores all the keys and identities.

The protocol runs as Fig. 2 shows:

1. Reader transmits request and random number (r_1, r_2) as challenges to the two tags, Tag_A and Tag_B.
2. After receiving the request and r_1, Tag_A generates a random number r_A and computes $A = Con(rot(r_A, r_1) \oplus ID_A, r_A)$ with r_A, r_1, ID_A, then he sends A and r_A as a reply to the Reader.
3. Upon receiving the A and random number r_A, the reader computes $\mathrm{V}_{\mathrm{R_A}} = Con(r_A, ID_A)$and transmits$V_{\mathrm{R_A}}$ to Tag_B.
4. When receiving the $V_{\mathrm{R_A}}, Tag_B$ generates a random number r_B and computes a $B = Con(rot(V_{R_A}, r_B) \oplus ID_B, r_B \oplus r_2)$ with the challenge number $r_2, V_{\mathrm{R_A}}$, the new random number r_B and its identity ID_B, after computing the B, it computes $V_B = Con(rot(B, r_2), k_2)$ with its secret key k_2, B and r_2. Then it transmit B, r_B and V_B to the Reader.
5. After receiving the three numbers, the reader computes $V_{R_B} = Con(r_B, ID_B)$ with r_B and ID_B and transmits V_{R_B} to the Tag_A.

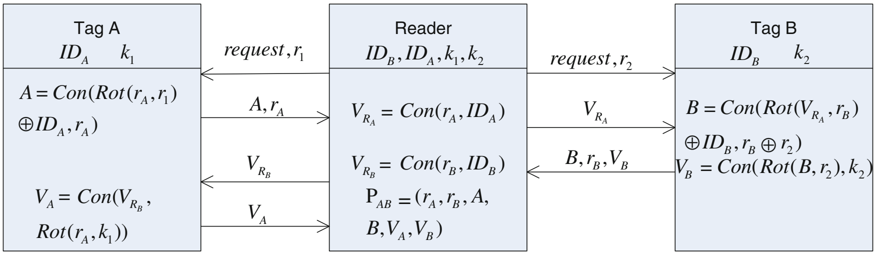


Fig. 2 A new lightweight RFID protocol for two tags

6. Then Tag_A computes $V_A = Con(V_{R_B}, rot(r_A, k_1))$ with V_{R_B}, the random number generated by Tag_A and its secret keyk_1. After this, it transmits V_A to the Reader.
7. The Reader collects r_A, r_B, A, B, V_A, V_B as a proof P_{AB}, then submits the P_{AB} to the backend server.
8. When the server receives P_{AB}, he computes a local P'_{AB} using the shared information in his database and computes the local value with what he received before. If they are equal, the protocol is successful, otherwise, the communication cannot continue anymore.

Analysis

In this section, we will evaluate our protocol on the following eight aspects. They are cost of the protocol, attack on the Tags, attack on the Reader, attack on the communication between the tags and the Reader, attack on user's privacy, tracking the tags, attack against the key and the last is disassembling the tags.

1. Overhead of the protocol.

In our protocol, we use three kinds of bitwise operations, bitwise XOR ($\oplus$), Rotation Rot (A, B) and a new designed function Con(X, Y). Compared with the existing protocols, most of them contain the hash to protect the privacy and security. But for passive tags, it is not appropriate.

2. Attack on the tags.

An adversary pretends to be a legal Reader and attack the tags with its range. From the protocol, we can learn that the legal reader computes V_{R_A} and V_{R_B} using the tags' IDs, but the adversary can't get the ID_i or r_i transmitted by the tags because the signal sent by the tags cannot reach the adversary. So the adversary cannot compute the V_{R_i} (i = A or B). Thus, our protocol can fight against this attack.

3. Attack on the Reader.

When an adversary tries to pretend a Tag and cheat the Reader, he can't success. Because the entities shared secret keys and they are not transmitted during the protocol.

4. Attack on the communication between tags and reader.

According to the ISO14443, we can assume that the adversary stays farther from the Tags than the Reader that he can't capture or intercept the information transmitted from the tags to the Reader.

5. Attack against the user's privacy.

What we can get from the description of the protocol is that there is no private information transmitted in the channel.

6. Tracking the tags.

Because the random numbers r_1, r_2, r_A, r_B are refreshed when a new protocol begins, the adversary can't use the previous information to carry out a new attack.

7. Attack against the keys.

The secret keys are stored in the backend server and the memory of each tag. When the protocol runs, they do not appear separately. No matter what kind of ability does the adversary own, he still can't get the keys.

8. Disassembling the tags.

Although the tags are not tamper-resistant, we can implant a kill-circuit. When the adversary try to get the tags physical structure by taking apart the tag, the circuit will output a signal to set up the kill mechanism and make the tag invalid forever.

9. Separating the tags.

In grouping proof protocols, resistance against separating the tags is one of the most important aims. Our protocol provides dependence on each other that can protect the tags from separating. If one tag is separated from the other, the proof P_{AB} submitted to the server in the end won't be verified successfully and the protocol will fail.

Conclusion

In this paper, we presented a lightweight RFID grouping proofs protocol that can be used for inpatient. The reader sends two different challenge numbers to the tags and the protocol begins. During the protocol runs, each tag computes verification messages using the information sent by the other one last time. At last, the reader collects enough information and submits it to the backend server to finish the protocol. Compared with other existing protocols for such RFID system that owns one reader and multiple tags, our approach achieves higher security and privacy requirement and improves efficiency. There are only three kinds of bitwise operation within the novel protocol. In the future, we will explore new schemes to solve some other practical problems and try to improve the RFID system.

Acknowledgments This work is supported by National Natural Science Foundation of China under Grant 61201159 and Fundamental Research Funds for the Central Universities under Grant 2012JBM016. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Juels A (2006) RFID security and privacy: a research survey. IEEE J Sel Areas Commun 24(2):381–394
2. Sarma SE, Weis SA, Engels DW (2003) RFID systems and privacy implications. In: Kaliski BS, Koc CK, Paar C (eds) Proceedings of the 4th international workshop on cryptographic hardware and embedded systems (CHES 2002). Lectures notes in computer science 2523. Springer, Berlin, pp 454–469
3. Sarma SE, Weis SA, Engels DW (2003) Radio-frequency identification: Secure risks and challenges. RSA Lab Cryptobytes 6(1):2–9
4. Weis SA, Sarma SE, Rivest RL, Engels DW (2004) Security and privacy aspects of low-cost radio frequency identification systems. In: Hutter D, Muller G, Stephan W, Ullmann M (eds) Proceedings of the 1st international conference on security in pervasive computing. Lectures notes in computer science 2802, Springer, Berlin, pp 201–212
5. Ohkubo M, Suzuki K, Kinoshita S (2004) Hash-chain based forward-secure privacy protection scheme for low-cost RFID. In: Proceedings of the 2004 symposium on crytography and information security (SCIS 2004), Sendai, pp 719–724
6. Henrici D, Muller P (2004) Hash-based enhancement of location privacy for radio-frequency identification devices using varying identifiers. In: Proceedings of the 2nd IEEE annual conference on pervasive computing and communications workshops (PERCOMW'04), Washington, DC, USA, pp 149–153
7. Tian Y, Chen G, Li J (2012) A new ultralightweight RFID authentication protocol with permutation. IEEE Commun Lett 16(5):702–705
8. Fan K, Li J, Li H, Liang X, Shen XS, Yang Y (2012) ESLRAS: a lightweight RFID authentication scheme with high efficiency and strong security for internet of things. In: Proceedings of 4th international conference on intelligent networking and collaborative systems, pp 323–328
9. Blass E-O, Kurmus A, Molva R, Noubir G, Shikfa A (2011) The Ff-family of protocols for RFID-privacy and authentication. IEEE Trans Dependable Secure Comput 8(3):466–480
10. Blass E-O, Elkhiyaoui K, Molva R (2012) PPS: Privacy-preserving statistics using RFID tags, world of wireless. In: Proceedings of 2012 IEEE international symposium mobile and multimedia networks (WoWMoM), pp 1–6
11. Di Pietro R, Molva R (2007) Information confinement, privacy, and security in RFID systems. In: Proceedings of the 12th European symposium on research in computer security, pp 187–202
12. Juels A (2004) Yoking proofs" for RFID tags. In: Proceedings of the first international workshop on pervasive computing and communication security, IEEE Press
13. Saito J, Sakurai K (2005) Grouping proof for RFID tags. In: Proceedings of the 19th international conference on advanced information networking and applications (AINA'05), pp 621–624
14. Peris-Lopez P, Orfila A, Hernandez-Castro JC, Van der Lubbe JC (2011) Flaws on RFID grouping-proofs. Guidelines for future sound protocols. J Netw Comput Appl 34(3):833–845
15. Piramuthu S (2007) Protocols for RFID tag/reader authentication. Decis Support Syst 43:897–914

Iris Recognition

Hwei Jen Lin, Yue Sheng Li, Yuan Sheng Wang and Shih Min Wei

Abstract An iris recognition system uses the iris to distinguish the identity of a person using the rich iris texture feature. To effectively remove noise and precisely segment the stable iris region is a crucial stage prior to recognition. Most noises on iris images are caused by occlusion of eyelids or eyelashes in certain areas. In this paper, we propose an iris recognition system which precisely locates and segments iris regions. We extract the iris feature from a relatively reliable portion of the iris region using a DoG filter. Experimental results show that the proposed iris recognition system has satisfactory results in terms of time efficiency and recognition rate.

Keywords Biometric recognition · Iris recognition · Iris segmentation · Iris normalization · Feature extraction · Difference of Gaussian (DoG)

H. J. Lin (✉) · Y. S. Li · S. M. Wei
Department of Computer Science and Information Engineering, Tamkang University,
Taipei, Taiwan, Republic of China
e-mail: 086204@mail.tku.edu.tw

Y. S. Li
e-mail: 699410519@s99.tku.edu.tw

S. M. Wei
e-mail: wesley78527@gmail.com

Y. S. Wang
Office of Physical Education, Tamkang University, Taipei, Taiwan, Republic of China
e-mail: 119391@mail.tku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_99, © Springer Science+Business Media Dordrecht 2014

A LDoS Detection Method Based on Packet Arrival Time

Kun Ding, Lin Liu and Yun Liu

Abstract Low-rate Denial of Service (LDoS) attack is a new type of Denial of Service attack, which is difficult for the router and victim site to detect because the attack packets are as many as the valid packets. In consideration of the fact that the features in time domain between attack traffic and valid traffic are different, we present a method to identify such kind of attack and try to trace the potential location of the attacker. We also carry out a simulation to illustrate the usability of this method.

Keywords Denial of service · Low-rate · TCP · Retransmission timeout · Arrival time

Introduction

As a serious challenge of the Internet security, Denial of Service (DoS) attacks abuse resources in the networks. Recently, the evolution of such type of attack, especially the Low-rate Denial of Service (LDoS) attack [1, 2], has eluded lots of

K. Ding · Y. Liu (✉)
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: liuyun@bjtu.edu.cn

K. Ding
e-mail: 12120057@bjtu.edu.cn

K. Ding · Y. Liu
Key Laboratory of Communication and Information Systems, Beijing Municipal
Commission of Education, Beijing Jiaotong University, Beijing 100044, China

L. Liu
China Information Technology Security Evaluation Center, Beijing, China
e-mail: liul@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_100, © Springer Science+Business Media Dordrecht 2014

detection methods. The common flood-based DoS attacks crash the victim site by sending high-rate invalid packets continuously which produce a large number of packets and the number is quite larger than the valid packets, so that the routers can easily detect the attack by recording the traffic counters [3, 4]. While the LDoS attacks try to repeatedly provoke the network enter the retransmission timeout phase of TCP congestion control mechanism by sending high-rate packets periodically. The burst transmission instead of continuous transmission ensures that the total number of the attack packets during the detection period of the routers is as large as the number of valid packets, so that it is difficult for the routers to identify the LDoS attacks.

LDoS Attacks Model

Generously, LDoS attacks perform at a particular rate periodically and the burst length is fixed. According to these features, we use a "square wave" model [5]. As shown in Fig. 1, we denote the duration of attack bursts as $\boldsymbol{l}$, the rate as $\boldsymbol{R}$, and the period as $\boldsymbol{T}$. Consider the number of invalid packets which are transmitted to the victim site during each attack is $\mathbf{X}$. Thus, we can easily get the corresponding equation as follows:

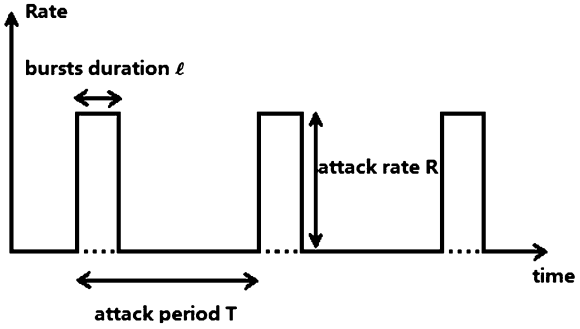
$$\mathbf{X} = \boldsymbol{l} \cdot \boldsymbol{R} \tag{1}$$

In most cases, the attackers send packets in a pulse pattern. While in some way, the rate pattern can be sinusoidal wave or cosine wave [6].

Characteristics Analysis in Time Domain

Since the attack packets during the router detection period are almost as many as the valid packets, even less than the valid packets, it is difficult for the router to detect the attack behavior. Although the packets from the normal subscriber are

Fig. 1 LDoS attack model



also burst data, there are some different characteristics between the valid packets and the attack packets. In time domain, the most difference is obviously the arrival time. The duration of the attack burst is about several hundred milliseconds. The arrival time of the attack packets is intensive during each attack, while the valid packets' arrival time is relatively dispersed.

Figure 2 depicts the arrival time of valid packets to the victim site in a simulation model. The distribution of the arrival time is random. And the simulation time is from 0 to 5.5 s. Each red line in the figure represents a packet. When there is no attack, the throughput of the network is quite high. Figure 3 depicts the arrival time of the attack packets. The most prominent feature in time domain is the periodicity and the aggregation.

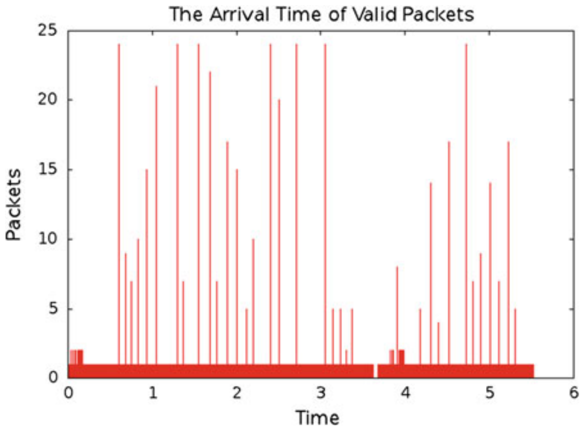
Influenced by the attack burst, the throughput of the network will be degraded. The arrival time of all packets is shown in Fig. 4. Due to the TCP congestion control mechanism, the receive window is reduced.

Detection Method

According to the above LDoS attack model, we present a detection method which is based on packet arrival time. This method is deployed in the uplink routers of the victim site or the victim site itself. Because the attack burst is high-rate traffic, the arrival time is relatively intensive. During the attack, the number of packets which enter the router per unit of time is larger than the usual.

When the traffic enters the queue of the router or passes through the router, log information will be recorded. The log information contains the source link, the destination, the arrival time, the packet type, etc. In our method, only the packets which reached the participating routers with the destination of the victim site are considered.

Fig. 2 The arrival time of valid packets



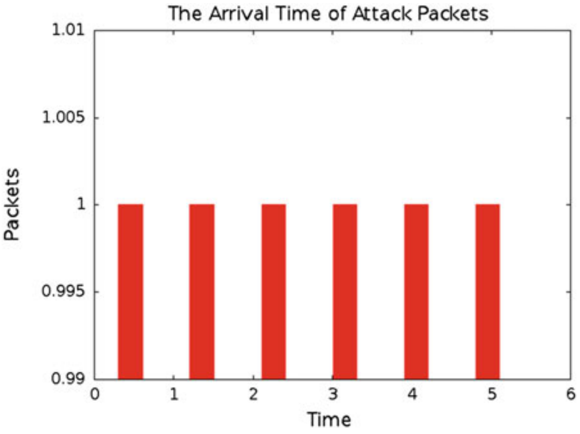


Fig. 3 The arrival time of attack packets

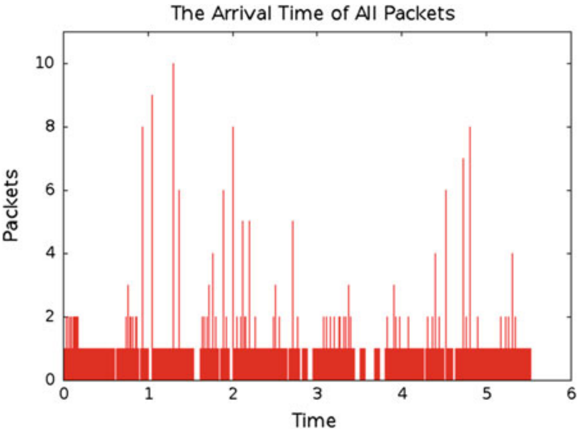


Fig. 4 The arrival time of all packets

When the network enters the retransmission timeout phase of TCP congestion control mechanic, the victim site sends control packets to all of the participating routers. The routers implement the algorithm as followed:

Definite $\mathbf{P}_i$ as the *i*th packet that reached the router, and let t_i be the time instant of the *i*th packet that reached the router. Thus, the time interval of the two adjacent packets can be depicted by Δ_i as followed:

$$\Delta_i = t_i - t_{i-1} \tag{2}$$

Definite parameter $\mathbf{r_{ci}}$ as the consequent ratio of Δ_i, $\mathbf{r_{ai}}$ as the antecedent ratio of Δ_i:

$$\mathbf{r_{ci}} = \frac{\Delta_{i+1}}{\Delta_i} \tag{3}$$

$$\mathbf{r_{ai}} = \frac{\Delta_j}{\Delta_{j+1}} \tag{4}$$

According to the characteristic of the attack packets mentioned in the previous part, $\mathbf{r_{ci}}$ will be greater at the instant when the current attack burst is over. Correspondingly, $\mathbf{r_{ai}}$ will be another numerical peak which represents the beginning of an attack burst.

Let δ_c be the threshold of $\mathbf{r_{ci}}$ and δ_a be the threshold of $\mathbf{r_{ai}}$. Analyze the log information and figure out Δ_1, Δ_2, Δ_3, …, Δ_i, … and then figure out $\mathbf{r_{c1}}$, $\mathbf{r_{c2}}$, $\mathbf{r_{c3}}$, …, $\mathbf{r_{ci}}$, … and $\mathbf{r_{a1}}$, $\mathbf{r_{a2}}$, $\mathbf{r_{a3}}$, …, $\mathbf{r_{ai}}$, … Compare $\mathbf{r_{c1}}$, $\mathbf{r_{c2}}$, $\mathbf{r_{c3}}$, …, $\mathbf{r_{ci}}$, … with δ_c, the one which is greater than δ_c would be the potential attack feature. If $\mathbf{r_{ck}}$ (k = 1, 2, 3…) is greater than δ_c, put the corresponding instant t_k of $\mathbf{P_k}$ in the "potential beginning attack" set **C**[m]:

$$\mathbf{C}[m] = t_k \, m = 1, 2, 3\ldots \tag{5}$$

m represents the mth element of set **C**. Theoretically, $\mathbf{P_k}$ represents the last packet of the current attack burst and t_k is the instant that the current attack burst is over.

Compare $\mathbf{r_{a1}}$, $\mathbf{r_{a2}}$, $\mathbf{r_{a3}}$, …, $\mathbf{r_{ai}}$,… with δ_a and if $\mathbf{r_{ak}}$ (k = 1, 2, 3…) is greater than δ_a, put the corresponding instant t_k of $\mathbf{P_k}$ in the "potential ending attack" set **A**[n]:

$$\mathbf{A}[n] = t_k \, n = 1, 2, 3\ldots \tag{6}$$

n represents the nth element of set **A**. $\mathbf{P_k}$ may be the first packet of the current attack burst.

Analyze all the log information and get set **C** and set **A**, and count the num of elements in these two sets. Definite threshold $\boldsymbol{\eta}$, if

$$\min\{\text{ element num of } \mathbf{C}, \text{ element num of } \mathbf{A}\} > \boldsymbol{\eta}$$

the alarm would be started which means the potential LDoS attack is detected.

Then, the router will analyze the log information and find out the source domain of attack packet. Finally, the router will send reply to the victim site to inform the victim site that a potential LDoS attack is on. In the reply, the attack source domain information will be included. After all of the participating routers sending replies, the victim site will get the attack route.

Simulation

Let us define the network model. As shown in Fig. 5, Router 1 is the uplink router of the victim site; Client 1 is in the local domain of Router 1; Client 2 and Attacker are in the different two domains of Router 2. Router 1 can be regarded as a gateway of the victim site. The duration of the simulation is set to 5.5 s. The arrival time of the packets from the valid clients is shown in Fig. 2. The arrival time of the packets from the attacker is shown in Fig. 3.

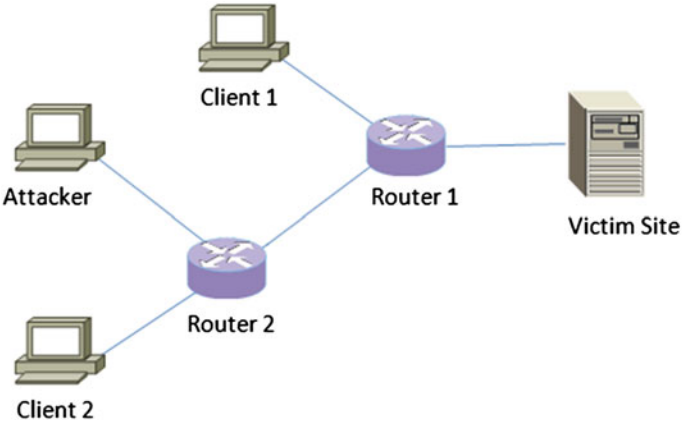
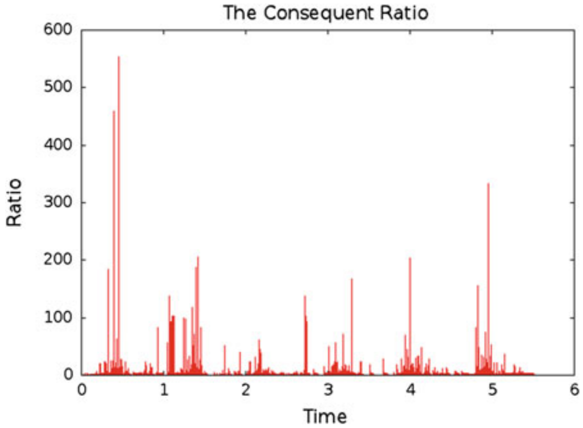


Fig. 5 The network topology of the simulation

Fig. 6 The consequent ratio



Client 1 and Client 2 send burst packets to the victim site randomly at the rate of 2 Mbps.

When there is no LDoS attack, no packet loses in the network. While the attack bursts periodically result in packets lost and invoke the network enters the retransmission timeout phase of TCP congestion control mechanism.

The period of the attack is set to 0.3 s, and according to Eq. (1), parameters are shown as followed:

$$\boldsymbol{T} = 0.9\,\mathrm{s}$$
$$\boldsymbol{R} = 3\,\mathrm{Mbps}$$
$$\boldsymbol{l} = 0.3\,\mathrm{s}$$
$$\mathbf{X} = 900\,\mathrm{Kb}$$

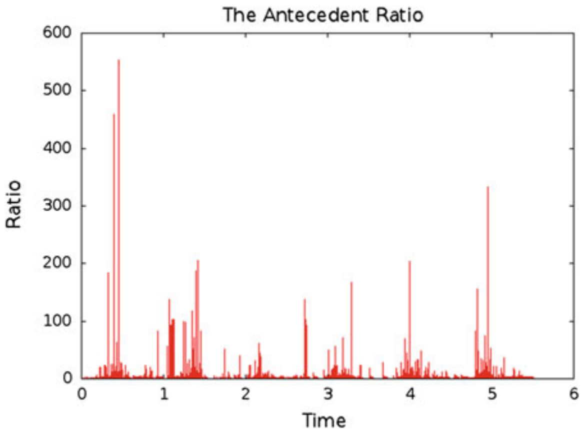


Fig. 7 The antecedent ratio

Table 1 Set A and set C

A[1]	0.328992	**C**[1]	0.399715
A[2]	0.399712	**C**[2]	0.433000
A[3]	0.453792	**C**[3]	0.453795
A[4]	1.394333	**C**[4]	1.343667
A[5]	1.423656	**C**[5]	1.394344
A[6]	3.295667	**C**[6]	1.423667
A[7]	3.998320	**C**[7]	3.095683
A[8]	4.824912	**C**[8]	3.295683
A[9]	4.949000	**C**[9]	4.115824

By analyzing the log information of Router 1, the consequent ratio and the antecedent ratio can be shown in Figs. 6 and 7. δ_c and δ_a are adjustable parameters. Different routers can set different value. In this model, they are both set to 150 in Router 1. And the adjustable parameter $\boldsymbol{\eta}$ is set to 6. According to Eqs. (5) and (6), set **C** and set **A** can be shown in Table 1.

$$\min\{\text{ element num of C},\ \text{element num of A}\} = 9 > \boldsymbol{\eta}$$

In this model, Router 2 does not detect the attack behavior. After Router 1 sends reply to the victim site. The victim site will get the message that the attacker is in the local domain of Router 1.

Conclusion

The LDoS attacks, whose number of attack packets is as small as possible, are quite different from the well-known flood-based DoS attacks. Methods which are based on statistics of traffic do not work in such attacks because the number of attack packets is too small to be detected by the routers.

We have analyzed one of the characteristics in time domain between the attack traffic and the valid traffic. By take advantage of the difference of the arrival time, we present a method to detect the potential LDoS attack behavior and try to trace the location of the attacker. Simulations show the usability of this method.

This method also provides a reference to rebuild the attack model by evaluating parameters of the attack model. And we will try to improve this method in the efficiency and reliability in the future.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024. The authors are also grateful for the comments and suggestions of the reviewers.

References

1. Yang G, Gerla M, Sanadidi MY (2004) Defense against low-rate TCP-targeted denial-of-service attacks. In: Proceedings of ninth international symposium on computers and communications ISCC 2004, vol 1. IEEE, pp 345–350
2. Kuzmanovic A, Knightly EW (2006) Low-rate TCP-targeted denial of service attacks and counter strategies. IEEE/ACM Trans Networking 14(4):683–696
3. Wong TY, Law KT, Lui JCS et al (2006) An efficient distributed algorithm to identify and traceback ddos traffic. Comput J 49(4):418–442
4. Chang RKC (2002) Defending against flooding-based distributed denial-of-service attacks: a tutorial. Commun Mag IEEE 40(10):42–51
5. Kuzmanovic A, Knightly EW (2003) Low-rate TCP-targeted denial of service attacks: the shrew vs. the mice and elephants. In: Proceedings of the 2003 conference on applications, technologies, architectures, and protocols for computer communications. ACM, pp 75–86
6. Sun H, Lui J, Yau DKY (2006) Distributed mechanism in detecting and defending against the low-rate TCP attack. Comput Netw 50(13):2312–2330

Simplifying Data Migration from Relational Database Management System to Google App Engine Datastore

Yao-Chung Chang, Ruay-Shiung Chang and Yudy Chen

Abstract Cloud computing has been widely introduced because of its ability to increase resource utilization. Furthermore, cloud computing offers resources as services that taken part as one of the next generation computing technologies. Before delivery application to cloud computing environment, the first step is the migration of the data. Migrating data from relational database management system (RDBMS) to Google App Engine is time-consuming problem. Hence, Google App Engine (GAE) Datastore provides NoSQL data storage with configuration file that contains table schema and CSV or XML file. This study presents the method for simplifying data migration from RDBMS to GAE including blob data migration. The proposed method leverages AppCfg to provide convenience way for data migration. As a result, user has eliminated at least 75 % task effort for data migration.

Keywords Cloud computing · Data migration · Google App Engine · Task queue

Y.-C. Chang (✉)
Department of Computer Science and Information Engineering,
National Taitung Unviersity, Taitung, Taiwan, Republic of China
e-mail: ycc@nttu.edu.tw

R.-S. Chang · Y. Chen
Department of Computer Science and Information Engineering,
National Dong Hwa University, Hualien, Taiwan, Republic of China
e-mail: rschang@mail.ndhu.edu.tw

Y. Chen
e-mail: m9921081@ems.ndhu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_101, © Springer Science+Business Media Dordrecht 2014

Introduction

Since cloud computing was introduced, it has been becoming more and more popular every year. Easy to use (setup, maintenance, reliability and scalability) and low cost (pay per use) are the reasons for companies to move their systems into cloud [1–3]. Along with cloud computing growth, data also become larger and more complex every day. The need of database that can handle growth of data is unavoidable. The challenges are not only data growth, data processing and analyzing but also very hard when data become bigger and bigger.

Google App Engine (GAE) is a cloud computing provider that offers platform as a service for user to host their web applications in Google's infrastructures [4]. Moreover, GAE allows developer to write and run an application above Google's infrastructure which is used by Google [5]. GAE supports distributed data storage that provides NoSql schemaless object which is built on the top of Bigtable called Datastore [6]. Datastore is used for storing common data types like integers, floats, strings, dates and binary data; and for storing binary large objects such as image, audio or video, GAE provides Blobstore.

Since data structures between RDBMS and Datastore are not the same, data migration from RDBMS to Datastore seems to be troublesome. Therefore, GAE provides one tool named "AppCfg" that can be used for uploading and downloading either data or application. Before use AppCfg for uploading data, two files should be prepared for every table. One is configuration file that contains information about table, data type of each column, etc. The other is CSV or XML file that contains the data to migrate. The objectives of this work are: (1) Minimizing error-prone while migrating data from RDBMS to GAE. (2) Reducing the time for data migration. (3) Currently, AppCfg doesn't support files bulk uploading, it is also important to provide files bulk uploading from other server into GAE blobstore.

The rest of this study is organized as follows: An overview of migrating data to GAE is described in section Related Studies. Section System Framework provides the system framework. The implementation of proposed method and subsequent result are sent out and analyzed in section Implementation. Finally, section Conclusions concludes the paper.

Related Studies

Data migration is known as one time process for migrating formatted data whereas both structure between databases are different on a conceptual and/or technical level [7]. Generally, it has some reasons for data migration [8]. Study in [9] presented that more than 75 % of their respondents encounter the problems during data migration. Some companies tackle those issues from affecting business operations by doing data migration during off-hours or weekends. However, this method increase the migration costs as a result of staff overtime, and this can negatively impact IT staff morale. Therefore, the good methodology for data migration is needed.

The relational database is a database that stores data into tables [10]. In relational database, one database can have multiple tables and if needed, the data between tables are related each other. The example of this database is MySQL. RDBMS fits for structured data, while now the data is transforming into structured (data come from RDBMS), semi structured (XML data) and unstructured data (files). Hand in hand with RDBMS, NoSQL came when the data complexity was becoming overhead.

Google App Engine is a platform as a service cloud provider. As a cloud provider, Google offers convenience services for users. Without software and hardware to buy or maintain, user can use GAE without limitation of application scaling. Rich API and simple to use is one of the advantages from GAE [11]. GAE performs load balancing and cache management automatically by allocating resources depends on traffic and resource use patterns. Every applications runs in a sandbox to prevent malicious operations. It also can improve CPU and memory utilization for multiple applications on the same physical machine [12].

Google Datastore is derived from Google Bigtable and it provides SQL-like query language called Google Query Language (GQL). The data are stored on High Replication Datastore (HRD) and are replicated across multiple data centers using Paxos algorithm and most queries are eventually consistent [13]. GAE has improvement from ordinary NoSQL database by providing transaction and roll back. Every transaction is committed if all of query on the transaction are success, otherwise it will be rolled back. The differences between traditional databases and GAE Datastore are shown in Table 1.

Since the data stored in database has different format, reliable tools for data migration is needed [14]. Some previous studies about data migration have been proposed. A middleware is created in Ref. [14] for data migration between RDBMS. It helps to reduce the time of migration, which is important when dealing with huge volumes of information. The methodology for migration database tables and their data between various types of RDBMS also has been proposed in Ref. [15]. It provides decision support during the process of data migration. Another study about data migration between document oriented and RDBMS has been proposed in Ref. [16]. It provides a model for the flexible specification of transformation rules and transforming the data schema from RDBMS into object oriented model.

Table 1 Differences between traditional databases and GAE datastore

	GAE datastore	Proposed system
Automatic scaling	Yes	No
Flexible queries	Yes	Yes
Strict data integrity	Customized on application	Yes

Moving data and applications from one cloud provider to another is possible when user has better choices [17]. When the migration between cloud providers occurs, it is not easy to move the data. The study in Ref. [17] proposed cloud broker to process the data migration between different cloud providers.

Because the Google AppCfg and AppRocket can not be used for migrating blob data, therefore the proposed approach in this study provides a convenience way to uploading data into Datastore and supporting blob data migration. This work provides simplicity and convenience way for data migration. The graphical user interface is introduced for user to input database's and Datastore's information. Also, user can upload their blob data during migration process. The configurations files and data files are automatically generated by the systems and user is welcomed to edit the configurations as needed. Not only uploading the automatic generated files, the proposed system can upload the data that has been prepared by user. Finally, the logs of data migration processes will exhibit the processes is succeed or failed.

System Framework

In this study, graphical user interface (GUI) of AppCfg is proposed for data migration from RDBMS to GAE Datastore. This method adds one layer above AppCfg that interacts with AppCfg. Some modifications of AppCfg are needed to eliminate interaction of user's authentication when migration occurs.

The proposed system architecture is separated into two parts (shown in Fig. 1), one is server side, and the other is client side. Server is the application runs on the Google App Engine. Client is the application runs on the client-side that interacts with end user.

- Server's Side

For processing blob data migration, Blob Downloading Manager is created on the server. Three additional columns will be added. The first column is "blobUrl". Due to the name of column for blob URLs is different from every user, this column holds URLs of blob data that is retrieved from original column from database that contains blob URLs. The second column is "newBlobUrl". This column is empty and will be filled with blobstore's key when the data is uploaded into blobstore. The third column is "fetchedBlob". This column also empty and will be filled "Y" when the data is uploaded into blobstore.

- Client's Side
  - Graphical User Interface (GUI) is the interface that interacts with users. It provides the input from users about RDBMS's information, GAE's information and blob's information. It also shows the result of each process.
  - Data Uploading Agent is an agent that controls each process on the client-side. It provides the GUI, queries the information and data from RDBMS,

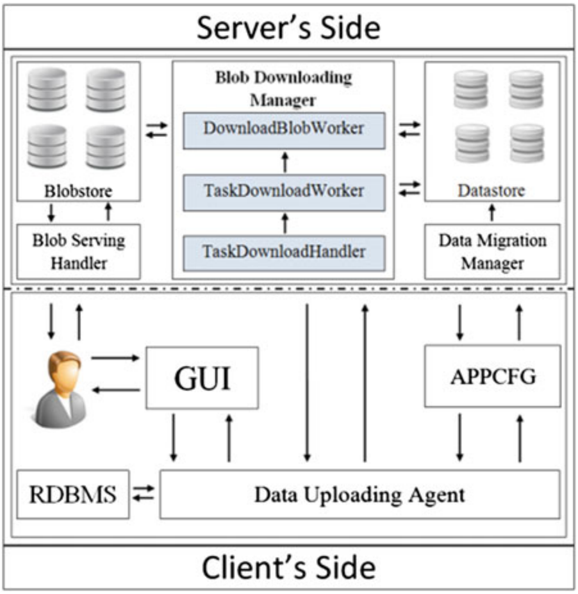


Fig. 1 Proposed system architecture

calls the TaskDownloadHandler on GAE to download the blob data and passes the message from and to AppCfg.

The correlation between TaskDownloadHandler, TaskDownloadWorker and DownloadBlobWorker is shown in Fig. 2. First, variables such as names of table and column are declared. After that, data uploading agent queries the data schema from RDBMS and gets the table's name and columns' names. For each row, the agent compares the variable tableName and the value on field table's name. If the variable tableName is empty, it means this is the new table and needs to set variable writeToFile with the variable declarations which contains the declaration needed for AppCfg's configuration file. If “tableName” is equal with the value of the field table's name, it means the current row is still belonging with same table. Otherwise, the table is new or no table anymore. Finally, it is the time for write the configuration file to local disk. Then, the system will show the configuration files to user. User can change or provide the information about blob data migration. The system needs column's name that contains blob data and/or server's path. If the server's path is blank, the proper blob path including server address should be provided as blob's path.

Due to the table's name from one user and another user is different, and GAE uses database model from the application on the TaskDownloadWorker, the database model's name from default name with the user table's name should be changed. For every row, DownloadBlobWorker will check if there is any blob has been downloaded or not. If has, the system will use the same blob's key as a “newBlobUrl”. Otherwise, the system will download the blob data.

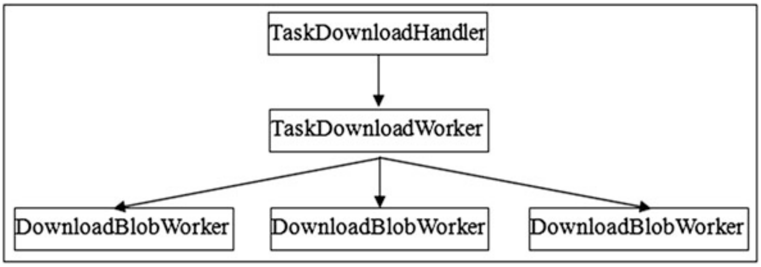


Fig. 2 Correlation between handler and workers

Implementation

- Data Types Translation

Data Types Translation is the process of creating configuration file. The system will run the query from RDBMS to get three columns. The columns needed for every table are "TABLE_NAME", "COLUMN_NAME" and "DATA_TYPE". TABLE_NAME is used for kind's value on GAE. COLUMN_NAME is used for property value and DATA_TYPE is used for "import_transform" on GAE.

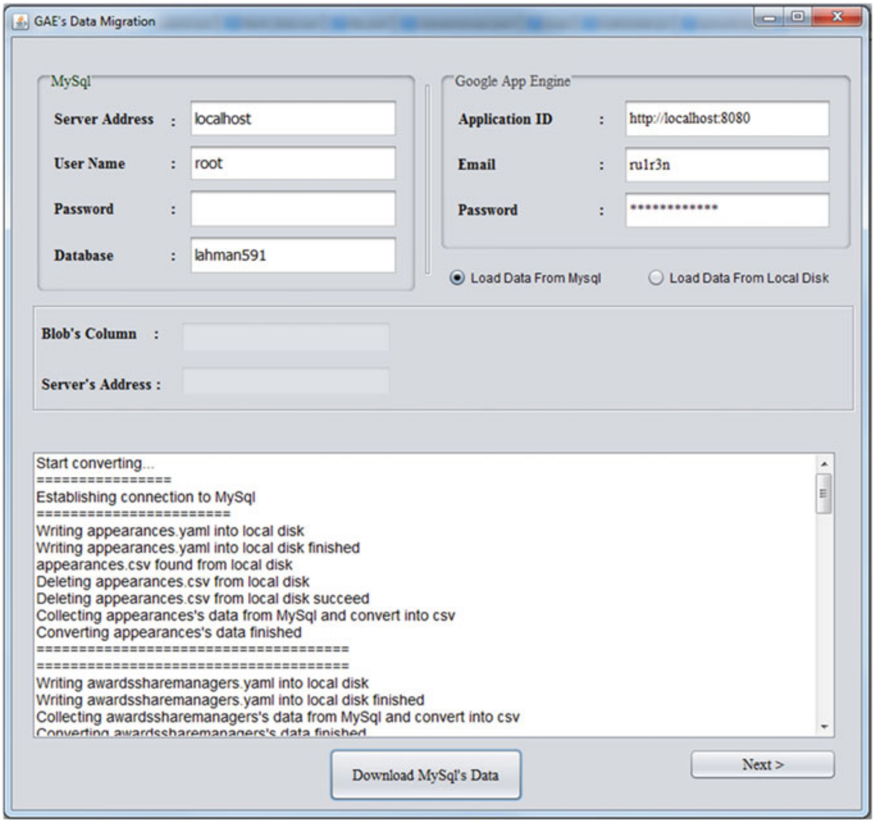
- Lahman's Baseball Data Migration

Non blob data means that data types are not binary large objects data, e.g., text and number. On the other hand, blob data means data types are binary large objects data, e.g., audio, video and image. Non blob data migration is the process for migrating non blob data from RDBMS to GAE. The data come from Lahman's Baseball Database. It contains the history of batting and pitching statistics back to 1871, plus fielding, standings, team's statistics, managerial records, post-season data, and more [18]. In this implementation, some tables from Lahman's database are used.

- ImageNet's Database Migration

For blob data uploading, user needs to provide the information about column that holds the blob's path and or server's address. The system will add three additional columns on the configuration file and data file. In this implementation, the proposed system uses the source from ImageNet. ImageNet is an image database organized according to the WordNet hierarchy, in which each node of the hierarchy is depicted by hundreds and thousands of images [19].

Figure 3 shows the user interface of data migration from RDBMS to GAE. Users need to provide RDBMS information such as RDBMS's address, username, password, and database. On the right side, user needs to provide GAE's information including application ID, email and password. On the bottom, user can see the logs of the processes. User can also choose "Load Data From Local Disk" if they have prepared data on the same directory with the application.fig.



Server Address : localhost
User Name : root
Password :
Database : lahman591

Application ID : http://localhost:8080
Email : ru1r3n
Password : ************

● Load Data From Mysql ○ Load Data From Local Disk

Blob's Column :
Server's Address :

```
Start converting...
================
Establishing connection to MySql
=======================
Writing appearances.yaml into local disk
Writing appearances.yaml into local disk finished
appearances.csv found from local disk
Deleting appearances.csv from local disk
Deleting appearances.csv from local disk succeed
Collecting appearances's data from MySql and convert into csv
Converting appearances's data finished
=====================================
=====================================
Writing awardssharemanagers.yaml into local disk
Writing awardssharemanagers.yaml into local disk finished
Collecting awardssharemanagers's data from MySql and convert into csv
Converting awardssharemanagers's data finished
```

Fig. 3 GUI for data migration from RDBMS to GAE

- Jobs Efficiency

The jobs demanded for migration are shown in Table 2 including "Creating configuration file", "Preparing and migrating the data", "Migrating blob data" and "Debugging". By using the proposed method for migration of the data, user has eliminated at least 75 % task effort. User provides the information about RDBMS and GAE and the rest of migration process will be done by proposed system. User only needs to debug if the migration process failed to run.

Table 2 Jobs needed comparison

	Manual	Our system
Creating configuration file	Yes	No
Preparing & migrating the data	Yes	No
Migrating blob data	Yes	No
Debugging	Partially yes	Partially yes

Conclusions

Simplify data migration from traditional database to Google App Engine Datastore is presented in this study. The blob data migration using task queue are also implemented in the proposed system. User only needs to interact to one system for data migration without dealing with many complicated systems. Moreover, the proposed system prepares the needed files for data migration. User is allowed to modify the prepared data, even the data types dictionary can be edited manually. Convenience way for data migration has been proposed in this work. As a result, user has eliminated at least 75 % task effort for data migration.

Acknowledgments The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially/partially supporting this research under Contract No. NSC101-2221-E-143-005-, NSC101-2221-E-259-003- and NSC101-2221-E-259-005-MY2.

References

1. Elsenpeter R, Velte AT, Velte TJ (2010) Cloud computing—a practical approach. McGraw-Hill, USA, p 23
2. Ingthorsson O (2012) 5 Reasons cloud computing is key to business success. http://www.datacenterknowledge.com/archives/2012/06/25/5-reasons-cloud-computing-is-key-to-businesss-success/. Retrieved 5 Nov 2012
3. Paul F (2012) 8 Reasons why cloud computing is even better for small businesses. http://readwrite.com/2012/04/06/8-reasons-why-cloud-computing. Retrieved 5 Nov 2012
4. Google (2012) What is Google App Engine? https://developers.google.com/appengine/docs/whatisgoogleappengine. Retrieved 29 Dec 2012
5. Severance C (2009) Using Google App Engine. O'Reilly Media, USA, p 14
6. Burrows M, Chandra T, Chang F, Dean J, Fikes A, Ghemawat S, Gruber RE, Hsieh WC, Wallach DA (2008) Bigtable: a distributed storage system for structured data. J ACM Trans Comput Syst 26(2), article no. 4, June 2008
7. Matthes F, Schulz C, Haller K (2011) Testing & quality assurance in data migration projects. In: 2011 27th IEEE international conference on software maintenance (ICSM), pp 438–447, Sept 2011
8. Kang S, Reddy ALN (2008) User-centric data migration in networked storage systems. In: IEEE international symposium on parallel and distributed processing, IPDPS 2008, pp 1–12, June 2008
9. IBM (2007) Best practices for data migration: methodologies for planning, designing, migrating and validating data migration. https://www-935.ibm.com/services/us/gts/pdf/softek-best-practices-data-migration.pdf. Retrieved 29 Jan 2013
10. Wikipedia (2013) Relational database. http://en.wikipedia.org/wiki/Relational_database. Retrieved 9 Jan 2012
11. Google (2012) Google App Engine—the platform for your next great idea
12. Prodan R, Sperk M, Osterman S (2012) Evaluating high-performance computing on Google App Engine. Software IEEE 29(2):52–58
13. Google (2012) Datastore overview. https://developers.google.com/appengine/docs/python/datastore/overview. Retrieved 15 Dec 2012

14. Elamparithi M (2010) Database migration tool (DMT)—accomplishments & future directions. In: 2010 International conference on communication and computational intelligence (INCOCCI), Dec 2010, pp 481–485
15. Walek B, Klimes C (2012) A methodology for data migration between different database management systems. Int J Comput Inf Eng 6:85–90
16. Walek B, Klimes C (2012) Data migration between document-oriented and relational databases. World Acad Sci, Eng Technol (69):894–898, Sept 2012
17. Shirazi MN, Kuan HC, Dolatabadi H () Design patterns to enable data portability between clouds' databases. In: 2012 12th International conference on computational science and its applications (ICCSA), June 2012, pp 117–120
18. Lahman S (2012) Download Lahman's baseball database. http://www.seanlahman.com/baseball-archive/statistics/. Retrieved 30 Dec 2012
19. Imagenet (n.d) ImageNet. http://www.image-net.org. Retrieved 1 Jan 2012
20. Couchbase (2012) NoSQL database technology
21. Google (2012) Blobstore python API overview. https://developers.google.com/appengine/docs/python/blobstore/overview. Retrieved 15 Dec 2012
22. Dancis K (2009) AppRocket 2.0.0. http://kaspa.rs/. Retrieved 16 Dec 2012

A Degree-Based Method to Solve Cold-Start Problem in Network-Based Recommendation

Yong Liu, Fan Jia and Wei Cao

Abstract Recommender systems have become increasingly essential in fields where mass personalization is highly valued. In this paper, we propose a model based on the analysis of the similarity between the new item and the object that the users have selected to solve cold-start problem in network-based recommendation. In order to improve the accuracy of the model, we take the degree of the items that have been collected by the user into consideration. The experiments with *MovieLens* data set indicate substantial improvements of this model in overcoming the cold-start problem in network-based recommendation.

Keywords Recommender systems · Network-based filtering · Similarity · Item degree · Cold-start

Introduction

With the rapid development of the computer network and other mass media, the amount of information that we can get is growing exponentially [1]. We are facing so much information that we have to spend a lot of time and effort to obtain the

Y. Liu · F. Jia (✉)
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: fjia@bjtu.edu.cn

Y. Liu
e-mail: 12120117@bjtu.edu.cn

Y. Liu · F. Jia
Key Laboratory of Communication and Information Systems, Beijing Municipal
Commission of Education, Beijing Jiaotong University, Beijing 100044, China

W. Cao
China Information Technology Security Evaluation Center, Beijing, China
e-mail: caow@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_102, © Springer Science+Business Media Dordrecht 2014

information that are most appropriate for us. Recommendation system is regarded as the most promising way to solve information overload problem. Recommendation systems recommend items of interest to users based on users' own explicit and implicit preferences, the preferences of other users, the attributes of users, and the attributes of items [2]. For example, a book recommender might integrate explicit ratings data (e.g., Tom rates *Introduction to Algorithms* a 3 out of 5), implicit data (e.g., Tom purchased *Algorithms in a Nutshell*), user demographic information (e.g., Tom is male), and book content information (e.g., Introduction to Algorithms is marketed as a computer-related book) to make recommendations to specific users. As recommendation system has enormous significance to the development of economy and society, a wide variety of recommender algorithms have been proposed, such as collaborative filtering algorithm [3, 4], content-based filtering algorithm [5], spectral analysis, principle component analysis, network-based algorithm [6–9], and so on.

In recent years, the network-based recommender algorithm, which was proposed by Tao Zhou in Ref. [4], has been extensively investigated over the past several years and it tends to be one of the most successful technologies for recommendation system. Some physical dynamics, including heat conduction process and mass diffusion, have been used in network-based personal recommendation. The recommender algorithm that adopted these physical approaches have been demonstrated to be both highly efficient and of low computational complexity. However, just as the collaborative filtering algorithm, such pure network-based recommendation also has the problem of cold-start. The newly added items that haven't been collected by any user are not suggested to the users.

Fortunately, content information of the newly added items can help to bridge the gap between existing and new items by inferring similarities among them. Therefore the network-based recommendation can introduce into the concept of similarities to solve the cold-start problem. In order to improve the accuracy of recommending the new items, we introduce the degree of the items that the user have collected to adjust the similarities of the user and the new items. Moreover, we introduce a free parameter β to regulate the contributions of objects degree to user-new items similarities. We carry out a benchmark experiment on *MovieLens* data set and compare it against the conventional methods to demonstrate the effectiveness. The numerical results indicate that decreasing the influence of popular objects can further improve the algorithmic accuracy.

The rest of this paper is organized as follows. In section Background, we introduce the process of network-based recommender algorithm and explain why it has cold-start problem. In section The Model, we present our model to resolve the cold-start problem. The fourth section is numerical results where we display the experiment results and explain how well our model performs. Finally, we make a summary of our work.

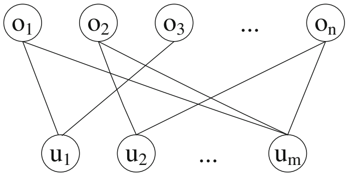


Fig. 1 Network-based recommender model

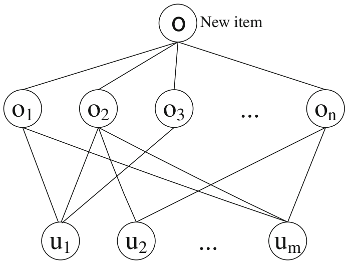
Background

In this section we describe the process of network-based recommender system and introduce the cold-start problem in it. The network-based recommendation was proposed by Tao Zhou in 2007 [4] and its model is shown in Fig. 1. Denote the object-set as $O = \{o_1, o_2, \ldots, o_n\}$ and user-set as $U = \{u_1, u_2, \ldots, u_m\}$. We give up with a binary variable a_{ij} to representation the relation of u_j with o_i. If u_j has already collected $o_i, a_{ij} = 1$ and $a_{ij} = 0$ otherwise. For example, u_1 is connected to o_1 and o_3 in Fig. 1. It represents that u_1 have collected o_1 and o_3, so $a_{11} = 1$ and $a_{31} = 1$. In this way, the relation of n items and m users can be described by an $n \times m$ adjacent matrix $\{a_{ij}\}$. Then, for a given user u_j the initial resource located on each item is a_{ij}. That's to say, if the object o_i has been collected by u_j, its initial resource is unit, otherwise it is zero. The initial resource can be understood as giving a unit recommending capacity to each collected object. According to the weighted resource allocation process discussed in Ref [4], the final resource of every object can be obtained easily. For any user u_j, all the uncollected items are sorted in descending order of the final resource, and those items with highest value of final resource are recommended. The new items haven't been collected by any user, so their final resource is zero and they wouldn't be suggested forever. This problem is the so-called cold-start.

The Model

In this section we describe a model to address the cold-start problem. The model utilizes content information of items and the record of what the user have collected In order to keep consistency with the second section, we denote the object set as $O = \{o_1, o_2, \ldots, o_n\}$ and user set as $U = \{u_1, u_2, \ldots, u_m\}$. A recommendation system can be fully described by an adjacent matrix $A = \{a_{ij}\} \in R^{n,m}$. If o_i is collected by $u_j, a_{ij} = 1$ and $a_{ij} = 0$ otherwise. As we all know, every object has its own features and they can be divided into different categories. For example, a movie may be tagged with "Action" and "Comedy". We can acquire the main features of the movie from these keywords. So if we can obtain the interest of the users, we can suggest specific items that are consistent with user hobbies. The user hobbies can be obtained from the objects that the user has collected. For instance, if most of the movies that a user has seen belong to romance, recommender system

Fig. 2 The proposed model



should suggest romantic movies to the user. Based on this method, we can calculate the similarities of new items and items that have been collected first and then get the similarities of new items and users. For a new item, all the users are sorted in the descending order according to their similarities with the new item, and the new item is recommended to those users, which are ranked in the front of all the users. This model can be described as Fig. 2.

Data related to the style of items are gathered and a "style-item matrix" is built, as shown in Fig. 3. The matrix consists of binary data. For example, object o_1 belongs to style 1, style 2 and object o_2 belongs to style 1, style Q. The "style-item matrix" can be used to calculate the similarities of two different items. Here we use the cosine theorem to determine the similarity of the two items. We use F to represent the "style-item matrix" and its column vector can be represented as f_i, which indicates the features of o_i. The elements of f_i are expressed as f_{ji}, which indicates whether o_i belongs to style j or not. Defined in the same way, the feature vector of the new item is written as fn and its elements are expressed as fn_j.

By using the cosine expression, the similarities between the new object and o_i can be written as:

$$s_{in} = \frac{1}{\sqrt{|f_i||fn|}} \sum_{j=1}^{Q} f_{ji} fn_j. \tag{1}$$

In this equation, $|f_i|$ and $|fn|$ is the modulus of f_i and fn respectively and they can be expressed as:

Fig. 3 Style-item matrix

$$\begin{array}{c} \\ \text{Style 1} \\ \text{Style 2} \\ \vdots \\ \text{StyLe Q} \end{array} \begin{array}{c} \begin{array}{cccc} O_1 & O_2 & \cdots & \end{array} \\ \left[\begin{array}{cccc} 1 & 1 & & \\ 1 & & & 1 \\ & & & \\ & 1 & & \end{array}\right] \end{array}$$

$$|f_i| = \sum_{\mathrm{j}=1}^{Q} f_{ji}, \tag{2}$$

$$|fn| = \sum_{\mathrm{j}=1}^{Q} fn_j \tag{3}$$

The predicted score, to what extent the new object will be suggested to u_j, is given as:

$$v_{jn} = \frac{\sum_{i=1}^{n} s_{in} a_{ij}}{k(u_j)}. \tag{4}$$

In Eq. (4), $k(u_j)$ represents the degree of u_j and it means how many items u_j have collected. Equation (4) can be regarded as an average of the similarities between the new item and the previous items that the users have collected.

A common problem of Eq. (1) is that they have not taken into account the influence of an item's degree, so that items, which a user has collected previously, with different degrees have the same contribution to the similarity of the new item and the user. If user u_l have selected item o_i and o_j, that is to say, they are all in line with the user's interest. Provided that item o_i is very popular (the degree of o_i is very large) and item o_j is relatively less popular (the degree of o_j is relatively small), s_{in} (the similarity between o_i and the new item) is a very ordinary similarity and it does not mean the new item and u_l are very similar. Therefore, its contributions should be small. On the other hand, s_{jn} (the similarity between o_j and the new item) is a peculiar similarity and it means the new item and u_l are very similar. Therefore its contributions should be large. In other words, it is not meaningful if the new item is similar to a very popular item that the user has collected, while if the new item is similar to an unpopular item that the user has collected, the new item is much more likely to correspond with user's interest and it should be suggested to the user with the higher probability. Accordingly, the contribution of s_{in} to the similarity v_{ln} (if u_l has collected o_i) should be negatively correlated with its degree $k(o_i)$. We suppose the s_{in}'s contribution to v_{ln} being inversely proportional to $k^{\alpha}(o_i)$ where α is a freely tunable parameter. Therefore, the proposed similarity equation should be written as:

$$v_{jn} = \frac{\sum_{i=1}^{n} s_{in} a_{ij} k^{\alpha}(o_i)}{k(u_j)}. \tag{5}$$

From now on, for a given new item, we use Eq. (5) to calculate its similarities with every user and we can get m numbers (they are $v_{1n}, v_{2n}, v_{3n}, \ldots, v_{mn}$) in the end. Then the m users should be sorted in the descending order according to the value of $v_{1n}, v_{2n}, v_{3n}, \ldots, v_{mn}$ and the new item should be suggested to the top users.

Numerical Results

In this paper, we use the standard MovieLens [10] data set to evaluate the accuracy of the proposed model. There are 1,682 movies and 943 users in this data set. In fact, MovieLens is a rating system, where each user votes movies in five discrete ratings 1–5. Hence we suppose that, the movie is collected by a user if the rating given by the user is at least 3 [11]. The data set contains 100,000 ratings and 82.52 % of the ratings are at least 3, so it contains 82,520 user-item pairs. In order to test the proposed model, we randomly select 20 numbers from 1 to 1682 as the movie ids. Then we select the corresponding records according to the 20 movie ids as test set and the remaining data serve as training set. The records in test set are seen as the new items that haven't been collected by any user. The training set is treated as known information, while no information in the probe set is allowed to be used for prediction. A recommendation algorithm generally can be evaluated with ranking score, recall and precision. As our main goal in this paper is to solve the cold-start problem and ranking score can reflect the effectiveness of the algorithm in solving this problem best, we use ranking score to evaluate the given algorithm.

Here new movies can be represented as $NewO_i$ where $i = 1, 2, \ldots, 20$. For a given $NewO_i$, we can obtain an ordered list, which contains all the users. If the relation $NewO_i - u_j$ is in the test set, we measure the position of u_j in the ordered list. For instance, if u_j is the 20th from the top, the position of u_j can be denoted by $r_{ij} = 20/943 = 0.02$. The mean value of the r_{ij}, which is calculated over all the records in test set, can be used to evaluate the effectiveness of the algorithm. The smaller the mean value of r_{ij}, the more accurate the algorithm is.

The result is shown in Fig. 4 where horizontal axis is the freely tunable parameter and longitudinal axis is the mean value of r_{ij}.The red solid line represents the situation where we don't take the item degree into account and the black

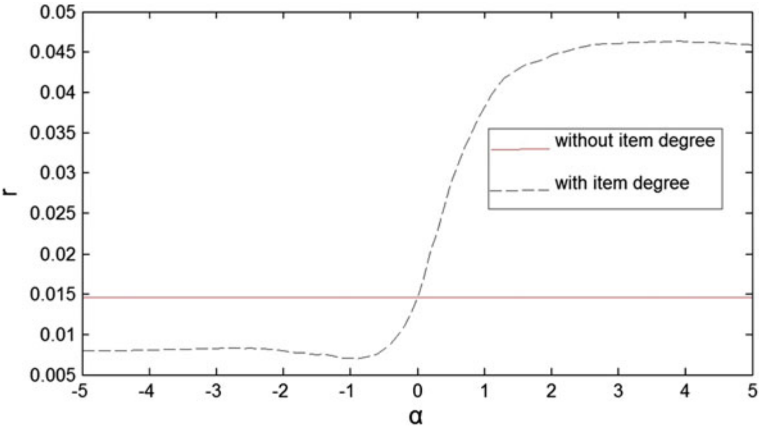


Fig. 4 The simulation results

dotted line reflects the change of r when α varies from -5 to 5. When $\alpha = 0$, the value of the solid curve is the same as the value of the dotted one. Compared with $\alpha = 0$, a positive α strengthens the influence of large-degree items, while a negative α weakens the influence of large-degree objects. We can see that, the algorithm has a better performance when $\alpha < 0$ and the curve has a clear minimum around $\alpha = -1$. When $\alpha = 0$, it means that we don't take item degree into account and $r = 0.0146$ in this case. When $\alpha = -1$, $r = 0.007$ and it is reduced by 52.05 %. It is indeed a greater improvement in solving cold-start problem.

Conclusion

In this paper, we propose a model to solve cold-start problem in network-based recommendation. The item degree is taken into account when we calculate the similarities between new items and the users. Furthermore, we introduce a free parameterα to regulate the influence of item degree. Numerical results indicate that decreasing the influence of large-degree items further improves the recommendation accuracy: When $\alpha = -1$, the r can be reduced by 52.05 % compared with the case where item degree isn't taken into account.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Fundamental Research Funds for the Central Universities under Grant 2013JBM006, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024. The authors are also grateful for the comments and suggestions of the reviewers.

References

1. Broder A, Kumar R, Moghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Comput Netw 33:309
2. Schein AI, Popescul A, Ungar LH, Pennock DM (2002) In: SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, pp 253–260
3. Adomavicius G, Tuzhilin A (2005) IEEE Trans Knowl Data Eng 17:734
4. Zhou T, Ren J, Medo M, Zhang Y-C (2007) Bipartite network projection and personal recommendation. Phys Rev E 76:046115
5. Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) ACM Trans Inf Syst 22:5
6. Ren J, Zhou T, Zhang Y-C (2008) Europhys Lett 82:58007
7. Zhou T, Jiang L-L, Su R-Q, Zhang Y-C (2008) Europhys Lett 81:58004
8. Zhang Y-C, Medo M, Ren J, Zhou T, Li T, Yang F (2007) Europhys Lett 80:68003
9. Liu C, Zhou W-X (2012) Heterogeneity in initial resource configurations improves a network-based hybrid recommendation algorithm. Physica A, pp 5704–5711
10. The MovieLens data can be downloaded from the website of GroupLens Research. http://www.grouplens.org

11. Blattner M, Zhang Y-C, Maslov S (2007) Phys A 373:753
12. Said A, Jain BJ, Albayrak S (2012) Analyzing weighting schemes in collaborative filtering: cold start, post cold start and power users. In: SAC '12 Proceedings of the 27th annual ACM symposium on applied computing, pp 2035–2040

Security Flaws of Off-Line Micro Payment Scheme with Dual Signatures

Shin-Jia Hwang

Abstract Wuu et al. proposed their off-line micro-payment scheme with dual signatures to provide customers' anonymity. However, some security flaw is pointed out. To remove this flaw, the channel between bank and trusted party and the channel between the bank and customers should be authenticated and secure.

Keywords Blind signatures · Payword chains · Micro payment

Introduction

Micro payment provides electronic payment mechanisms for small value transactions over networks. Due to the small value of each transaction, the computation and communication costs of a micro payment should be low. Among the proposed micro payment schemes [1–15], the payword chain [11] is the famous technique to reduce the computation cost of each transaction.

Wuu et al. [14] proposed the off-line micro payment scheme with dual signatures to provide anonymity for customers. To provide customers' anonymity, a trusted authority and a trusted issuer, are involved. The trusted authority authenticates and authorizes the pseudo public keys for each customer. Since the authorized pseudo public keys are validated only by the issuer, the bank needs the help of the issuer to validate the customer's pseudo public key. The issuer also generates some anonymous payword chain and obtains the bank's authorization by the blind signature scheme [2] to break the link between the payword chain and the customer. Since the bank cannot authenticate the customer and knows the blind signature on payword chains, a security flaw occurs.

S.-J. Hwang (✉)
Department of Computer Science and Information Engineering, TamKang University,
Tamsui, New Taipei City 251, Taiwan, Republic of China
e-mail: sjhwang@mail.tku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_103, © Springer Science+Business Media Dordrecht 2014

Table 1 Notation description

Notation	Description
N	An amount of coins withdrawn by a consumer
Life	Coin expiration date
T	Timestamp
H()	Public hash function
$H^n(\bullet)$	$H^n(\bullet) = H(H^{n-1}(\bullet))$ and $H^1(\bullet) = H(\bullet)$
PK_X, SK_X	Long-term public and secret key pair of the participate X
PPK_X, PSK_X	Pseudo public and secret key pair chosen by some customer
<Data>K	The ciphertext of data encrypted/decrypted with key K in public cryptosystems
Sig_X(Data)	A digital signature generated by participate X and $Sig_X(Data) = Data \| <H(Data)>SK_X$

Section Review of Off-Line Micro Payment Scheme with Dual Signatures. Section Security Flaw describes the security flaws of the off-line micro payment scheme. The final section is our Conclusion.

Review of Off-Line Micro Payment Scheme with Dual Signatures

The off-line micro payment scheme with dual signatures includes five kinds of participants: Consumer C, merchant M, bank B, issuer I, and trusted authority TA. The scheme contains four phases: Register, withdrawal, payment, and deposit phases. Table 1 defines the notations.

Register Phase

A customer C has to register at TA to obtain an encrypted payment certificated. C generates his/her pseudo public secret key pair (PPK, PSK), and submits TA his/her real identity C and PPK. Then TA gives C <PCert> PK_C, where PCert = $<Sig_{SKTA}$(C, PPK)> PK_I.

Withdrawal Phase

The withdrawal protocol is described below.

Step W1: C sends B the message (C, PCert, Sig_{SKPSK}(N, T)).
Step W2: B sends I the message (B, Life, PCert, Sig_{SKPSK}(N, T)).

Step W3: I recovers C and PPK by decrypting PCert and validates Sig_{SKTA}(C, PPK). If Sig_{SKTA}(C, PPK) is valid, then I randomly chooses two primes K_P, K_{SI} and computes $PK = K_P \times K_{SI}$. Finally, send B (H(CCert) × <r> PK_B, PK), where CCert = (B, <K_P> SK_I, PPK, Life) and r is the blind factor.

Step W4: B chooses two primes p and q such that gcd(PK, $\phi(n)$) = 1, and computes the product $n = pq$ and $K_{SC} = PK^{-1} \bmod \phi(n)$, where $\phi(n) = (p - 1)(q - 1)$. Then send I the message (<H(CCert)> $SK_B \times r$, n).

Step W5: I prepares a payword chain Coin = ((c_0, c_0'), (c_1, c_1'), …, (c_N, c_N')) by randomly choosing a integer c_N and computing $c_{i-1} = H(c_i)$ for i = N, N − 1, …, 1 and c_i' = <c_i> K_{SI} for i = 0, 1, …, N. Remove the blind factor r from <H(CCert)> $SK_B \times r$ to obtain <H(CCert)> SK_B. Then send B <Coin, Sig_{SKB}(CCert)> PPK, where Sig_{SKB}(CCert) = CCert||<H(CCert)> SK_B.

Step W6: B sends <Coin, Sig_{SKB}(CCert)> PPK and $Sig_{SKB}(K_{SC})$ to C.

Finally, customers obtains the key K_{SC}, and Coin and Sig_{SKB}(CCert).

Payment Phase

Assume that C spends j coins ((c_0, c_0'), (c_1, c_1'), …, (c_j, c_j')), where c_i' = <c_i'> $K_{SI} = (c_i)^{KSI} \bmod n$ for i = 0, 1, 2, …, j. The customer wants to spend another k coins on the merchant M with the transaction timestamp TT.

Step P1: C sends M the transaction <Order, Sig_{PSK}(PI)> PK_M, where Order is the order information and the payment information PI = (M, TT, Sig_{SKB} (CCert), $< c_j' > K_{SC} = (c_j')^{K_{SC}} \bmod n,\ c_{j+k}$).

Step P2: M obtain Order and Sig_{PSK}(PI) by decrypting <Order, Sig_{PSK}(PI)> PK_M. Then M verifies Sig_{SKB}(CCert) to check whether or not PPK is authorized by B. M checks the whether or not PI is authorized by the owner of PPK. M also obtains the authorized public key K_P from CCert and checks whether or not $\ll c_j' > K_{SC} > K_P = H^k\left(c_{j+k}\right)$. If all the verifications hold, M accepts transaction; otherwise rejects the transaction. Finally M acknowledges C the final result.

Deposit Phase

M deposits coins by sending Sig_{SKM}(M, TT, Sig_{PSK}(PI)) to B. B not only checks whether or not $\ll c_j' > K_{SC} > K_P = H^k\left(c_{j+k}\right)$ but also detects the occurrence of double spending. If double spending occurs, B asks TA to find out the owner of PPK.

Security Flaw

In the withdrawal phase, B and I do not authenticate one another, so a malicious customer can cheat B. Suppose that some malicious customer C performs the withdrawal protocol once, so C has CCert = (B, <K_P> SK_I, PPK, Life) and the public key K_P. The cheating attack is described below.

Step W1: C sends B the message (C′, PCert, Sig_{PSK}(N, T′)), where C′ is another customer real identity and T′ is another timestamp.

Step W2: B sends I the message (B, Life, PCert, Sig_{PSK}(N, T′)) but C intercepts (B, Life, PCert, Sig_{PSK}(N, T′)).

Step W3: C constructs another pseudo public and secret key pair (PPK′, PSK′). C also chooses another prime K_{SI}' and computes $PK' = K_P \times K_{SI}'$. C sends B (H(CCert′) × <r> PK_B, PK′), where CCert′ = (B, <K_P> SK_I, PPK′, Life′) and r′ is the blind factor.

Step W4: B chooses two primes p′ and q′ such that gcd(PK, $\phi(n')$) = 1, and computes the product n′ and $K'_{Sc} = PK'^{-1} \bmod \phi\,(n')$, where $\phi(n') = (p' - 1)(q' - 1)$. Then send I the message (<H(CCert′) > SK_B × r′, n′) that is intercepted by C.

Step W5: C prepares $\text{Coin}' = \left(\left(c_0,\ c_0'\right), \left(c_1,\ c_1'\right), \ldots, \left(c_N,\ c_N'\right)\right)$ by randomly choosing a integer c_N and computing $c_{i-1} = H(c_i)$ for $i = N, N - 1, \ldots, 1$ and $c_i' = \, < c_i > K_{SI}'$ for $i = 0, 1, 2, \ldots, N$. Remove the blind factor r′ from <H(CCert′)> SK_B × r′ to obtain <H(CCert′)> SK_B. Then C sends <Coin′, Sig_{SKB}(CCert′)> PPK′ to B, where
Sig_{SKB}(CCert′) =
CCert′|| <H(CCert′)> SK_B.

Step W6: B sends <Coin, Sig_{SKB}(CCert′)> PPK′ and Sig_{SKB}(K'_{SC}) to C.

Since B cannot authenticated the received messages, the malicious customer C entraps the bank and an innocent customer C′. To overcome the cheating flaw, the communication link between B and I should be authenticated. B relies on I to bind the pseudo public key and the real customer identity with the encrypted payment certificate PCert, the communication link between B and C should be also secure; otherwise, Sig_{SKB}(K_{SC}) leases the important secret key K_{SC}.

Conclusion

In Wuu et al.'s off-line micro payment scheme with dual signatures, the customer easily impersonates the issuer to entrap the bank and some innocent customer. To remove this security flaw, the channel between the bank and issuer must be secure and authenticated in the withdrawal protocol.

References

1. Bellare M, Garay J, Hauser R, Herzberg A, Krawczyk H, Steiner M, Tsudik G, Waidner M (1995) iKP—A family of secure electronic payment protocols. In: Proceeding of 1st USENIX workshop on electronic commerce, pp 89–106
2. Chaum D (1983) Blind signatures for untraceable payments. In: Advances in cryptography—Proceeding of Crypto'82. Springer, New York, pp 199–203
3. Chaum D, Fiat A, Naor M (1988) Untraceable electronic cash. In: Advances in crytology—CRYPTO'88, LNCS, vol 403. Springer, New York, pp 21–25
4. Chen L, Kudla C, Paterson KG (2004) Concurrent signatures. In: Advances in cryptology—EUROCRYPT 2004, LNCS, vol 3027. Springer, Berlin, pp 287–305
5. Furche, A, Wrightson G (1996) SubScrip—an efficient protocol for pay-per-view payments on the internet. In: Proceedings of 5th international conference on computer communications and networks (ICCCN'96), Rockville, MD, 16–19 Oct 1996
6. Glassman S, Manasse MS, Abadi M, Gauthier P, Sobalvarro P (1996) The millicent protocol for inexpensive electronic commerce. In: World wide web journal, proceeding of 4th international world wide web conference. O'Reilly, Boston, MA, pp 603–618
7. Herberg A (1998) Micropayment. In: Kou W (ed) Payment technologies for E-commerce. Springer, New York, pp 245–282
8. Huang C-W (2006) A postpaid micropayment scheme with revocable customers' anonymity. Master Thesis, Tamkang University, Taiwan, R.O.C
9. Lin S-Y (2004) Design and cryptanalysis of micropayment schemes. Master Thesis, National Central University, Taiwan, R.O.C
10. Manasee MS (1995) The millicent protocols for electronic commerce. In: Proceeding of 1st USENIX workshop on electronic commerce, pp 117–123
11. Rivest RL, Shamir A (1997) PayWord and MicroMint: two simple micropayment schemes. In: Proceeding of security protocols workshop, LNCS 1189. Springer, New York, pp 69–87
12. Stern J, Vaudenay S (1997) SVP: a flexible micropayment scheme. In: Proceeding of financial cryptography, LNCS, vol 1318. Springer, New York, pp 161–172
13. Tsou J-H (2005) The study of electronic payment scheme. Master Thesis, Tamkang University, Taiwan, R.O.C
14. Wuu L-C, Chen K-Y, Lin C-M (2008) Off-line micro payment scheme with dual signature. J Comput 19(1):23–28
15. Yen S-M (2001) PayFair: a prepaid internet micropayment scheme ensuring customer fairness. Comput Digital Tech, IEE Proc 148(6):207–213

Mobile Reference Based Localization Mechanism in Grid-Based Wireless Sensor Networks

Ying-Hong Wang, Yi-Hsun Lin and Han-Ming Chang

Abstract Wireless sensor networks (WSNs) are based on monitoring or managing the sensing area by using the location information with sensor nodes. These sensor nodes are sometimes random deployed, so they have to be aware to their location before starting their tasks. Most sensor nodes need hardware support or receive packets with location information to estimate their location, and this needs lots of time or costs, and may have a huge error. In this paper we present a localization mechanism in wireless sensor networks (MRN). This mechanism can cooperate with node localization algorithm and mobile reference node moving direction scheme. We use a mobile reference node with GPS to move to the whole environment, and we use RSSI and trilateration to estimate unknown nodes' location. We can obtain more unknown nodes location by mobile reference node moving scheme, and will decreases the energy consumption and average location error.

Keywords Wireless sensor networks (WSN) · Localization · Mobile sensor node · Received signal strength indicator (RSSI)

Introduction

Recent years wireless sensor networks [1] are getting more convenient, and the applications or researches are become skillful. So in this paper we proposed a mechanism in wireless sensor networks localization.

Y.-H. Wang (✉) · Y.-H. Lin · H.-M. Chang
Department of Computer Science and Information Engineering, Tamkang University,
Tamsui, Taiwan, Republic of China
e-mail: inhon@mail.tku.edu.tw

Y.-H. Lin
e-mail: kidsle0830@gmail.com

H.-M. Chang
e-mail: chmcicel74723@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_104, © Springer Science+Business Media Dordrecht 2014

In recent researches, we find out localization can be classified as range-based and range-free approaches. Range-free approaches do not assume the availability or validity of distance information and only rely on the connectivity measurements undetermined sensors to a number of seeds [1]. Having lower requirements on hardware, the accuracy and precision of range-free approaches are easily affected by the node densities and network conditions, which are often unacceptable for many WSN applications that demand precise localizations. Range-based approaches calculate node distances based on some measured quantity [2], whereas they usually require extra hardware support; thus, they are expensive in terms of manufacturing cost and energy consumption. And how to reduce the extra cost becomes an important task for us to find out. When the sensor node position is estimate, it would have a significant increase for the data transfer speed or other works need to do.

In this paper we proposed a mechanism using a mobile reference node (MRN) with RSSI [6] and trilateration in wireless sensor networks localization to reduce energy consumption costs and reduce the location error. In section Related Work will introduce some range-based and range-free approaches. In section Localization Mechanism with Mobile Reference Node will introduce the mechanism we proposed. And final are the simulation and conclusion.

Related Work

Many approaches have been proposed to determine sensor node locations, falling into two categories: (a) range-free approaches and (b) range-based approaches.

Range-Free Approaches

Knowing the hardware limitations and energy constraints required by range-based approaches, researchers propose range-free solutions as cost-effective alternatives. Having no distances among nodes, range-free approaches depend on the connectivity measurements from sensor nodes to a number of reference nodes, called seeds. For example, in Approximate Point-in-Triangulation Test (APIT) [2], some sensor nodes have high frequency to transmit signal with GPS or other ways to obtain sensor nodes location called Beacon. As an alternate solution, DV-Hop [3] only makes use of a constant number of seeds. Instead of single-hop broadcasts, seeds flood their locations throughout the network, maintaining a running hop count at each node along the path.

Range-Based Approaches

Range-based approaches assume that sensor nodes can measure the distance and/or the relative directions of neighbor nodes. Various techniques are employed to measure the physical distance. For example, TOA obtains range information through signal propagation times [4], and TDOA estimates the node locations by utilizing the time differences among signals that are received from multiple senders [4]. As an extension of TOA and TDOA, AOA allows nodes to estimate the relative directions between neighbors by setting an antenna array for each node [5]. All those approaches require expensive hardware. For example, TDOA needs at least two different signal generators. AOA needs antenna arrays and multiple ultrasonic receivers. RSSI is utilized to estimate the distance between two nodes with ordinary hardware [6].

Therefore, no interaction is required between nodes, avoiding cumulative errors of coordinate calculations and unnecessary communication overhead. The localization accuracy can also be improved by multiple measurements that are obtained when the mobile nodes are at different positions. With the RSSI values from mobile node to an unknown node in an ideal sense, the distance between other unknown nodes should be calculated according to the log-normal shadowing model in (1), which is widely used in range-based localization approaches [6], i.e.

$$\mathrm{RSSI}(d) = \mathrm{P}_T \text{ - } \mathrm{P}_L(d_0) - 10\alpha\log_{10}\frac{d}{d_0} + X_\sigma \tag{1}$$

where ***PT*** is the transmission power, ***PL(d0)*** is the path loss for a reference distance of ***d0***, and ***η*** is the path-loss exponent. The random variation in RSSI is expressed as a Gaussian random variable ***Xσ = N(0, σ2)***. All powers are in given in decibels relative to 1mW, and all distances are given in meters. ***η*** is set between 2 and 5. ***σ*** is set between 4 and 10, depending on the specific environment [6].

And last step we use the trilateration [7]. The trilateration is a common mathematical formula, using the distances we estimate, and then calculate the unknown node location.

Localization Mechanism with Mobile Reference Node

In this section presents our proposed mechanism. The mechanism can divide into two phases. First phase is Node localization Algorithm, and in the second phase is Mobile Reference node moving direction scheme, according to the environment size to choose which algorithm we should use. In Fig. 1 is whole mechanism structure.

Network Environment

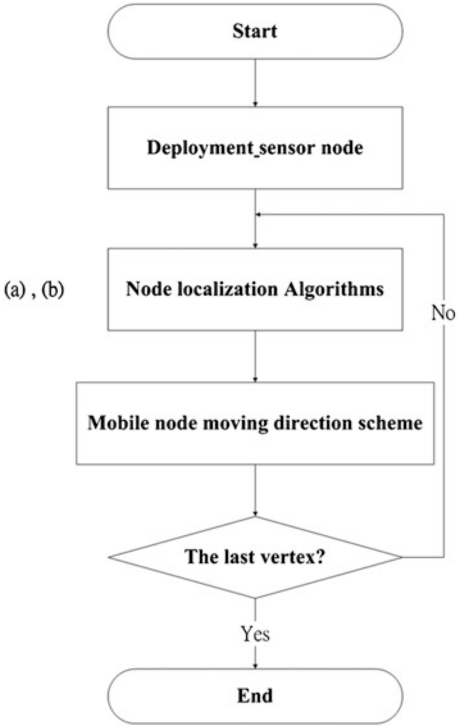
In this section, introduce our proposed localization mechanism in the environmental setting conditions. As shown in Fig. 2, we random deployment 20 unknown sensor nodes, and give each unknown sensor node its unique node ID, and mark the upper left corner is the initial position of the mobile reference node.

As shown in Fig. 2, after the sensor nodes deployed, the sink will know the length and width of the entire environment, assuming that the width of the environment define as M, and the length of the environment as L, each grid length is define as the mobile reference node transmission radius R, and then we use the M/R = k, L/R = p these two formulas to get the values as k, p, and use the results of the values of k to do the following decisions: (1) divided each virtual grid and tag an grid ID on each virtual grid (2) determine the mobile reference node start position (3) help mobile reference node to set a first direction need to turn.

Node Localization Algorithm

In this section we introduce node localization algorithm, it has two phases (a) Mobile Reference node Broadcast Algorithm, (b) Sensor node Localization Algorithm. They are parallel execution in the localization mechanism in the first phase.

Fig. 1 Grid-based mobile sensor node localization mechanism


Start
Deployment_sensor node
(a) , (b)
Node localization Algorithms
No
Mobile node moving direction scheme
The last vertex?
Yes
End

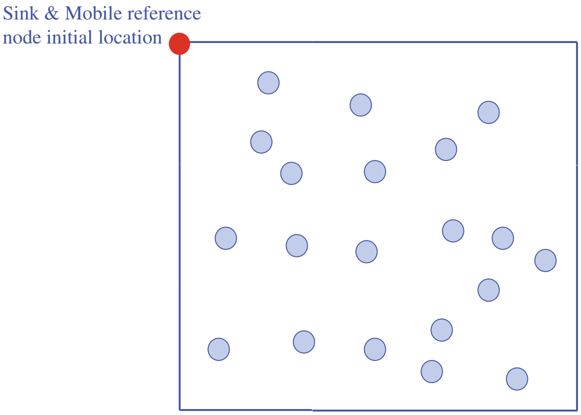


Fig. 2 Mobile reference node and Sink initial position

Mobile Reference Node Broadcast Algorithm

First, mobile reference node start broadcast a Wake_up of beacon to wake up the unknown sensor nodes where are in the virtual grid, then broadcast a Initial_start signal, then mobile reference node will move to reach the R/2 position, as the half transmission radius. To connect with unknown sensor nodes, and show the location mobile reference node is. After that mobile reference node will start moving to the end point of a virtual grid's length. Last when mobile reference node move to the end point will broadcast a Middle_stop signal means mobile reference node has finished moving a grid side length, shown in Fig. 3 is a single virtual square broadcast.

In Table 1 is the introduction of start signal, start signal packet contains two fields Start_signal_flag, and Mobile node coordinate, first Start_signal_flag is the one used to decide what kind of signal type. As shown in Fig. 5, first we use the k value to determine if it's an odd number, if so, to determine whether the special region of the grid in the grid area, as shown in Fig. 4, assume that we want to locate unknown sensor nodes in the position of x, first mobile reference node came to (1) location, unknown sensor node can receive the first start signal, then the unknown sensor node will use sensor node localization algorithm to statistics the number of receiving signal.

Sensor Node Localization Algorithm

In this section we introduce sensor node localization algorithm, the main function of this algorithm is for unknown sensor node performance in a virtual grid, through the broadcast receive from mobile reference to calculate the signal strength and signal attenuation formula to calculate the distance. We will make unknown sensor nodes that are waiting for the start signal into the sleep state until the Wake_up beacon accept in the next action. Finally, when an unknown sensor node use three signal values and calculate its node coordinate of the location through trilateration, will wait to receive any end signal to do the action, to transmit a packet, and the

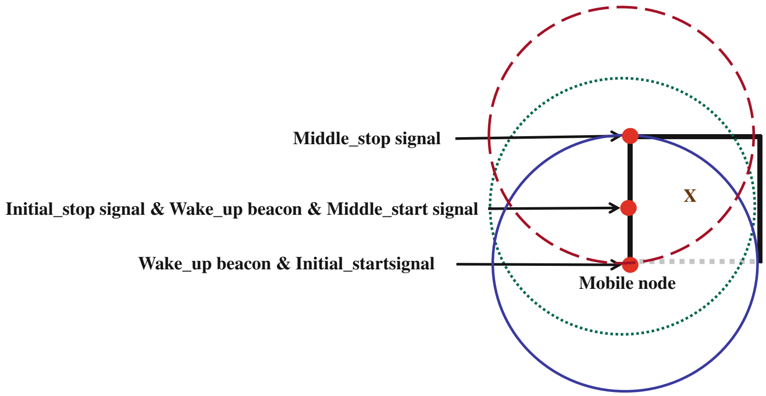
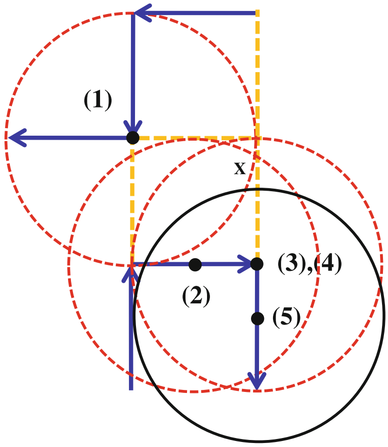

Middle_stop signal
Initial_stop signal & Wake_up beacon & Middle_start signal
Wake_up beacon & Initial_startsignal
X
Mobile node

Fig. 3 Single virtual square broadcast

Table 1 Start signal packet

Start signal packet	
Start_signal_flag	Mobile node coordinate

Fig. 4 Exception grid



format as shown in Table 2, while waiting for end signal, the unknown sensor node will continue to wait for and won't enter the sleeping mode.

Mobile Reference Node Moving Direction Scheme

After mobile reference node moves an edge, we need to determine the next move and the point of view, therefore, we propose Mobile reference node moving direction scheme, and is divided into two choices: (a) Mobile reference node

Table 2 Unknown sensor node return packet

Sensor_node_position packet	
Sensor node ID	Sensor node coordinate

moving direction for even algorithm (b) Mobile reference node moving direction for odd algorithm. And the flow charts are shown in Fig. 5.

Mobile Reference Node Moving Direction for Even Algorithm

As shown in Fig. 6, when the environment size is 4 × 4. We find out a regularity for mobile, mobile reference node change direction, when we will continue to make twice the same direction of rotation.

Mobile Reference Node Moving Direction for Odd Algorithm

We use solid lines in Fig. 7 to determine place, when the mobile reference node moves to the sides of the environment, the moving situation is different with the front, just move down linear, and we define a special switch as SPMD, shown in Fig. 8, the real line the circle marked SPMD will set to 1, to control the rotation of the position.

The mobile reference node has move in the bottom of the environment shown in Fig. 9. Left part of the execution flow, is the same we talk in the front when even × even environment size situation.

Simulation and Analysis

We use NS2 vision 2.29 as the simulator to analyze our proposed localization mechanism. The simulation parameters are similar to a method we compare in [8], and shown in Table 3.

Average Location Error Analysis

In this paper we to hope that can reduce the estimate location error rate of the unknown sensor nodes, and has reached a more complete experimental data through the different size of the environment.

Shown in Fig. 10, in different size of environments and compare the average location error in each environment size, can be found between the MRN and the PI have almost the same location error, because the PI is uninterrupted broadcast to find the strongest signal strength and use the triangle moving path in order to achieve the lowest error rate.

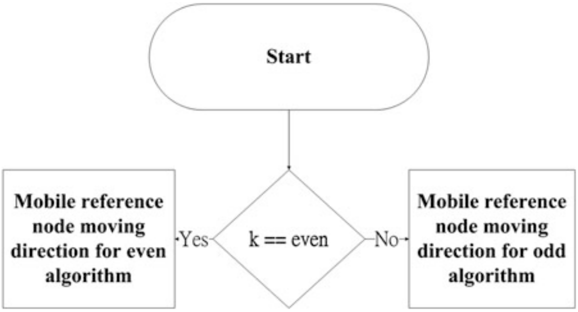

Start
Mobile reference node moving direction for even algorithm
Yes
k == even
No
Mobile reference node moving direction for odd algorithm

Fig. 5 Mobile reference node moving direction scheme

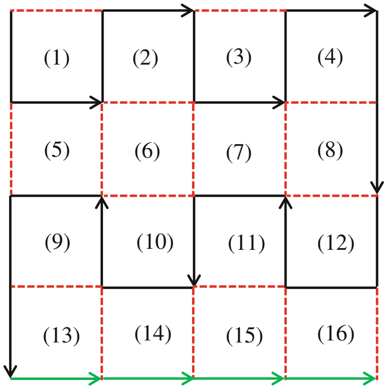


Fig. 6 Environment size (p × k) is 4 × 4

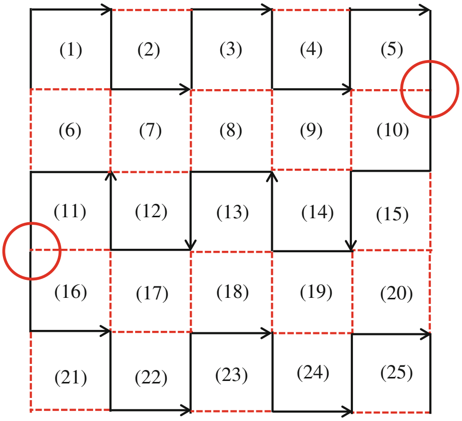


Fig. 7 Environment size (p × k) is 5 × 5

Energy Consumption Analysis

Shown in Fig. 11, in this simulation we only compared with PI and MBBGC, because the location error rate of PI and MBBGC are close with our proposed

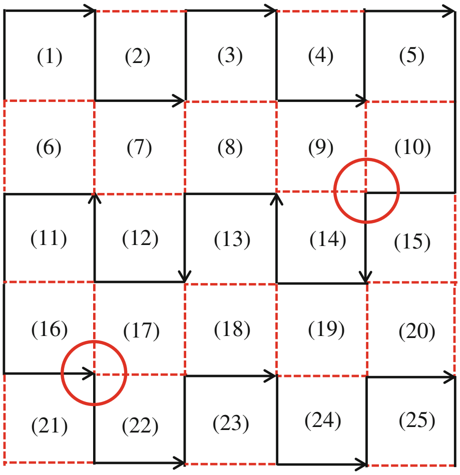


Fig. 8 Environment size (p × k) is 5 × 5

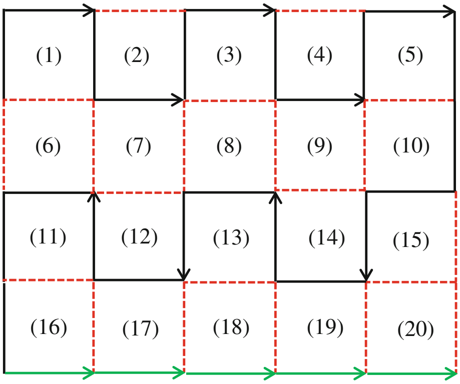


Fig. 9 Environment size (p × k) is 4 × 5

mechanism, the average location error between each other are 0.5–1 m. So in Fig. 11 can see the MRN we proposed has lower power consumption than other methods.

Conclusion and Future Work

In wireless sensor networks, it will increase lots of throughput when we can handle all sensor nodes' location in the whole environment. Nowadays there are a lot of localization technology are restriction by the cost or the natural environment, therefore in some cases made the location error can't be avoided, to reduce the location error need to find other direction to make the breakthrough.

Table 3 Parameter setting

Parameter	Value
Communication range	10 m
Mobile node moving rate	0.1 m/s
Transmission frequency	1 Hz
Sensor node power	3.6 V
Transmit energy consumption	40 mA
Receive energy consumption	35 mA
Sleep energy consumption	0.1 uA
Number of sensor nodes	80 ~ 130

Fig. 10 Average location error

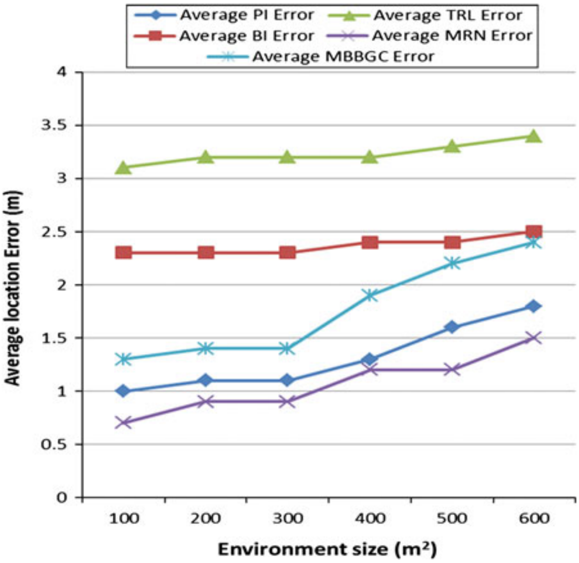
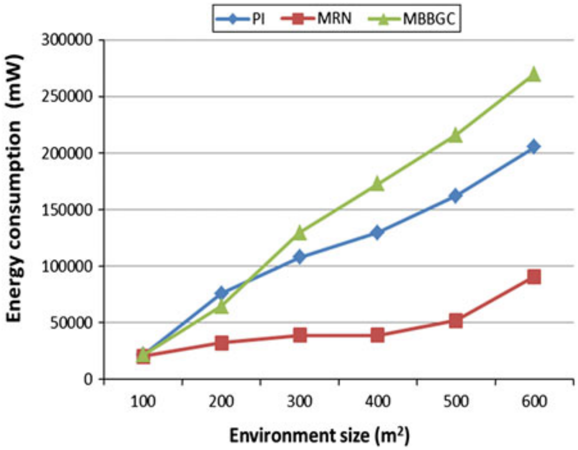


Fig. 11 Mobile node energy consumption



In this paper, we proposed a localization mechanism (MRN) using a mobile reference nod with RSSI method to estimate distances, finally we use Trilateration to ensure the location more correctly. According to the simulation results can find out that the mechanism we proposed can have the same location error between methods we compare. And in energy consumption comparison, have a very significant reduce, whether in the mobile node or unknown nodes.

In the future, we can improve our mechanism in the three-dimensional size of the environment, or Obstacles in the moving path to make sure the moving algorithm can cover whole environment.

References

1. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) A survey on sensor networks. IEEE Commun Mag 40(8):102–114
2. He T, Huang C, Blum BM, Stankovic JA, Abdelzaher TF (2005) Range-free localization schemes for large scale sensor networks. ACM Trans Embed Comput Syst 4(4):877–906
3. Niculescu D, Nath B (2003) DV-based positioning in ad hoc networks. Kluwer J Telecommun Syst 22(1):267–280
4. Frampton KD (2006) Acoustic self-localization in a distributed sensor network. IEEE Sens J 6(1):166–172
5. Dai F, Wu J (2006) Efficient broadcasting in ad hoc wireless networks using directional antennas. IEEE Trans Parallel Distrib Syst 17(4):335–347
6. Hightower J, Borriello G, Want R (2000) SpotON: an indoor 3D location sensing technology based on RF signal strength, University of Washington, Seattle, Univ. Washington, Tech. Rep. UW CSE 00-02-02, Feb 2000
7. Sun J, Yu J, Zhu L, Wu D, Cao Y (2012) Construction of generalized Ricci flow based virtual coordinates for wireless sensors network. IEEE Sens J 12(6):2109–2112
8. Guo Z, Guo Y, Hong F, Jin Z, He Y, Feng Y, Liu Y (2010) Perpendicular intersection: locating wireless sensors with mobile beacon. IEEE Trans Veh Technol 59(7):3501–3509
9. Lee S, Kim E, Kim C, Kim K (2009) Localization with a mobile beacon based on geometric constraints in wireless sensor networks. IEEE Trans Wireless Commun 8(12):5801–5805

A Delegation-Based Unlinkable Authentication Protocol for Portable Communication Systems with Non-repudiation

Shin-Jia Hwang and Cheng-Han You

Abstract For portable communication systems, the delegation-based authentication protocol provides efficient subsequent login authentication, data confidentiality, User privacy protection, and non-repudiation. However, in all proposed protocols, the non-repudiation of mobile users is based on an impractical assumption that home location registers are trusted. To reduce the HLR's trust assumption and enhance the non-repudiation of the mobile users, our new delegation-based authentication protocol is proposed. Our protocol also removes the exhaustive search problem in the subsequent login authentication to improve the subsequent login authentication performance. Moreover, the user unlinkability in the subsequent login authentication is also provided in our protocol to enhance the user identity privacy.

Keywords Delegation-based authentication · Concurrent signatures · Concurrent signcryption scheme · Portable communication systems

Introduction

Portable communication systems (PCSs) provide roaming services in wireless communication networks. In PCSs, User has to first register in some home location register (HLR) to get its legality. Before roaming, a user must login some visit location register (VLR) and VLR validates the user's legality with the help of the

S.-J. Hwang · C.-H. You (✉)
Department of Computer Science and Information Engineering, Tamkang University, Tamsui, New Taipei City 251, Taiwan, Republic of China
e-mail: 699420435@s99.tku.edu.tw

S.-J. Hwang
e-mail: sjhwang@mail.tku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_105, © Springer Science+Business Media Dordrecht 2014

HLR. If the user is legal in some HLR, VLR allows this user to use its roaming services.

Global system for mobile communications has some drawbacks [1]: No non-repudiation property, no users' identity privacy, and no mutual authentication between users and VLR. Some protocols try to improve the GSM protocol [2, 3] based on symmetric cryptosystems. But, it is hard to achieve non-repudiation property. Some public key cryptosystems-based protocols are proposed [4, 5] to provide both non-repudiation and mutual authentication.

The delegation-based authentication protocol for PCSs [1] is proposed by exhibiting off-line authentication processes to reduce the communication load between VLR and HLR; and keeping the lower computation load for users. The protocol satisfies the non-repudiation, mutual authentication, data secrecy, and identity privacy, but does not satisfy the non-repudiation property in off-line authentication processes [6]. So the enhanced protocol [6] is proposed. Both [7] and [8] indicated that Lee protocol suffers the linkable problem. To remove the link problem, [7] and [8] state two protocols, respectively. However, Wang et al. protocol still suffers forgery attack [9]. Both two protocols suffer the exhausted search problem in off-line authentication process [10]. To overcome exhaustive search problem, Chen et al. [10] proposed their protocol.

In the proposed protocols [1, 6–8, 10], a registered user obtained the proxy public key and the proxy private key generated by HLR. The proxy public key is used as the user's alias for roaming and the proxy private key is used to generate the proxy signatures to provide non-repudiation. However, HLR also knows the proxy key pair. To avoid misuse of the proxy key pair and provide the user's non-repudiation, those protocols must assume that HLR is trustworthy by all users.

To guarantee the HLR's trust means that the staffs of the HLR also are trusted. However, if some malicious staff in HLR steals the proxy private keys, then the staff can successfully forge the proxy signatures for some legal user and impersonate the legal user for roaming. Then the legal user owning the proxy private key cannot deny the roaming records from the malicious staff. The trust HLR assumption weakens users' non-repudiation. So, to reduce the trusted HLR assumption to semi-trust HLR assumption also enhances users' non-repudiation.

Our Contribution

Our new delegation-based authentication protocol for PCSs is proposed by first designing the embedded concurrent signcryption scheme with anonymity, adopting the concept of the confidential deniable authentication protocol [10]. In the embedded concurrent signcryption scheme, an initial signer and a matching signer can exchange their signatures in a fair and confidential way. So the concurrent signcryption scheme satisfies the following security properties [11, 12]: unlinkability, correctness, fairness, unforgeability, and confidentiality.

The new concurrent signcryption scheme is used to fairly exchange the user's signature on the proxy public key application and the HLR's signature on the proxy public key delegation warrants. Our delegation-based authentication protocol also satisfies the following security properties [1, 6–8, 10]: non-repudiation, mutual authentication, session key security, user identity privacy, user unlinkability, and no exhausted search.

Our new delegation-based authentication protocol is proposed in section Our Delegation-Based Unlinkable Authentication Protocol for PCS with User's Non-repudiation. The security analysis of our protocol is given in section Security Analysis and Proofs. The final section is our conclusions.

Our Delegation-Based Unlinkable Authentication Protocol for PCS with User's Non-repudiation

The parameter l is a security level parameter. The public parameters p and q are two large primes such that $q\ |(p-1)$. The public parameter g is an element in $\mathrm{Z}_{\mathrm{p}}^{*}$ with order q. Two hash functions $h()$ and $H()$ are published, where $h(.)$ maps from $\{0,1\}^{*}$ to Z_q^{*} and $H(.)$ maps from $\{0, 1\}^{*}$ to $\{0, 1\}^{l}$. Our protocol publishes a symmetric cipher satisfying indistinguishable security against adaptive chosen ciphertext attacks (IND-CCA2). The symmetric encryption function is denoted by $[M]_{\mathrm{K}}$, where M is a message and K is the symmetric secret key. Some public key based signature scheme is also published for all legal users. Notation $A = Sig[M]_x$ denotes the signature generation function and $Verify[A, M]_y$ is the signature verification function, where M is the message, A is the signature on M, x is the signer's private key, and y is the signer's public key. Notation K_{VH} denotes the shared secret key between VLR and HLR. Notations ID_V and ID_H denote the identities of VLR and HLR, respectively. Notation $m_1||m_2$ denotes the message m_1 is concatenated with the message m_2. Notation $A \rightarrow B$: M denotes that A sends the message M to B. Notations $(x_M, y_M = g^{x_M} \bmod p)$ and $(x_H, y_H = g^{x_H} \bmod p)$ denote the certificated private-public key pair of MS and HLR, respectively. Our protocol contains three phases: Initialization, login authentication, and subsequent login authentication phases.

Initialization Phase

Some mobile user submits its anonymous private-public key pair to the HLR for the first registration by the following registration protocol.

Step 1: User constructs an anonymous private-public key pair (x_K, y_K) by randomly selecting the private key $x_K \in Z_q^{*}$ and computing the public key $y_K = g^{x_K} \bmod p$.

Step 2: User generates the promise of Schnorr signature σ_M on the registration for (x_K, y_K).

Step 2.1: Select a random number $r_M \in Z_q^*$.

Step 2.2: Compute $V_M = h(g^{r_M} \bmod p, m_M || ID_M || y_K)$, the keystone $s_M = r_M + V_M x_M \bmod q$, and $S_M = g^{s_M} \bmod p (= g^{r_M + V_M x_M} \bmod p)$, where m_M is the registration letter for the anonymous private-public key pair (x_K, y_K).

Step 2.3: Store $\sigma_M = (S_M, V_M)$ and s_M in User's local database.

Step 3: User transmits $(\sigma_M, m_M || ID_M || y_K)$ to HLR through secure channels.

Step 4: HLR verifies σ_M by checking whether or not $V_M = h(S_M y_M{}^{-V_M} \bmod p, m_M || ID_M || y_K)$. If the equation does not hold, then stop.

Step 5: HLR generates the promise of Schnorr-like signature.

Step 5.1: Select a random number $r_H \in Z_q^*$.

Step 5.2: Compute $S_H = \mathrm{S_M}^{\mathrm{X_H}} \bmod p, V_H = h(g^{r_H} S_H \bmod p, m_H || ID_H || y_K), k = (r_H - V_H) x_H{}^{-1} \bmod q$, where m_H is the authorization warrant for (x_K, y_K).

Step 5.3: Store $(\sigma_H = (S_H, k, V_H), \sigma_M, m_H, m_M, ID_M, y_K)$ in HLR's temporary database by y_K's order

Step 6: HLR transmits (σ_H, m_H) to User through secure channels.

Step 7: User verifies σ_H by checking whether or not $\mathrm{V}_H = \mathrm{h}(\mathrm{g}^{\mathrm{VH}} \mathrm{y}_H \mathrm{k} \mathrm{S}_H \bmod \mathrm{p}, \mathrm{m}_H || \mathrm{ID}_H || y_K)$. If the equation does not hold, then stop.

Step 8: User sets his/her proxy private key as x_K and his/her proxy public key as y_K.

Step 9: User computes HLR's Schnorr-like signature $\rho_H = (s_H = s_M + k \bmod q, V_H)$. Then HLR easily recovers User's Schnorr signature $\rho_M = (s_M = s_H - k \bmod q, V_M)$ after obtaining ρ_H.

Login Authentication Phase

User contracts VLR to obtain services, and VLR checks the User's legality by the login authentication protocol. Suppose that the current unused proxy public key is y_K.

Step 1: User randomly selects an integer n_1 and computes a one-way hash chain $h^1(n_1), h^2(n_1), h^3(n_1), \ldots, h^{(\mathrm{n}+1)}(n_1)(=N_1)$, where $h^{i+1}(n_1) = h(h^i(n_1))$ for $n \geq i \geq 1$ and $h^1(n_1) = h(n_1)$.

Step 2: User constructs a new anonymous private-public key for the next round by selecting a random number $x_{K,new} \in Z_q^*$ and computing $y_{K,new} = g^{x_{K,new}} \bmod p$.

Step 3: User generates promise of Schnorr signature for new key pair $(x_{k,new}, y_{k,new})$.

Step 3.1: Select a random number $r_{M'} \in Z_q^*$.

Step 3.2: Compute $V_M' = h(g^{r_M'} \bmod p,\ m_M\|ID_M\|y_{K,new})$ the keystone $S_M' = r_M' + V_M' x_M \bmod q$, and $S_M' = g^{S_M'} \bmod p = g^{r_M' + V_M' X_M}$ mod p, where m_M is the registration letter for $(x_{K,new}, y_{K,new})$. Store $\sigma_M' = (S_M', V_M')$ and s_M' in its database

Step 4: User generates the promise of signcrytext.

Step 4.1: Select a random number $x \in Z_q^*$.

Step 4.2: Compute $Y = g^x$ mod p, $S_K = H(S_M'\|(y_H)^{s_M'} \bmod p)$, and $MK = H(SK)$.

Step 4.3: Generate a MAC = $H(MK,\ m_M\|Y\|ID_V\|ID_M\|V_M'\|y_{K,new})$ and encrypt them both by $C_M = [m_M\|Y\|ID_V\|ID_M\|V_M'\|y_{K,new}\|MAC]_{SK}$.

Step 5: User transmits ($\rho_H = (s_H,\ V_H) = (s_M + k$ mod $q,\ V_H)$, m_H, ID_H, y_K) to VLR.

Step 6: VLR verifies ρ_H by checking whether or not $V_H = h(g^{V_H} {y_H}^{s_H} \bmod p, m_H\|ID_H\|y_K)$. If the equation does not hold, then stop.

Step 7: VLR selects a random number n_2 and transmits User n_2, the period of validity *Per*, and ID_V.

Step 8: User generates the signature $A = Sig[N_1\|n_2\|Per\|ID_V]_{xK}$ then sends (A, (C_M, S_M'), ID_V, N_1) to VLR.

Step 9: VLR validates A on the message $N_1\|n_2\|Per\|ID_V$.

Step 9.1: Generate $[A\|Per\|N_1\|n_2\|y_K|(C_M, S_M')\|\rho_H\|Dig]K_{HV}$, if Verify$[A, N_1\|n_2\|\mathrm{Per}\|ID_V]_{y_K}$ is valid for $Dig = H(A\|Per\|N_1\|n_2\|y_K\|(C_M,\ s_M')\|\rho_H)$.

Step 9.2: Send ($[A\|Per\|N_1\|n_2\|y_K\|(C_M, S_M')\|\rho_H\|Dig]K_{HV}$, ID_H, ID_V) to HLR.

Step 10: HLR decrypts $[A\|Per\|N_1\|n_2\|y_K|(C_M, s_M')\|\rho_H\|Dig]K_{HV}$.

Step 10.1: Obtain $(A\|Per\|N_1\|n_2\|y_K\|(C_M, S_M')\|\rho_H\|Dig)$ by decrypting $[A\|\mathrm{Per}\|N_1\|n_2\|y_K\|(C_M, S_M')\|\rho_H\|\mathrm{Dig}]_{k_{HV}}$.

Step 10.2: Validate the recovered message by checking whether or not $Dig = H(A\|Per\|N_1\|n_2\|y_K|(C_M, S_M')\|\rho_H\|Dig)$.

Step 11: HLR validates the certificate ρ_H of User's anonymous public key y_k.

Step 11.1: Find $\sigma_H = (S_H,\ k,\ V_H)$ and $\sigma_M = (S_M,\ V_M)$ in HLR's temporary database using y_K as the searching key.

Step 11.2: Recover the User's signature $\rho_M = (s_M,\ V_M)$ by computing $s_M = s_H - k$ mod q.

Step 11.3: Verify ρ_M by checking whether or not If $V_M = h(g^{s_M y_M{}^{-V_M}} \bmod p, m_M \| ID_M \| y_K)$. If $\rho\ _M$ is valid, HLR believes that User is some legal User.

Step 11.4: Validate the signature A by checking *Verify* $[A, N_1 \| n_2 \| Per \| ID_V]_{y_K}$ to confirm whether the specified User is the one knowing the secret key x_K.

Step 12: HLR validates User's promise of the Schnorr signature on $y_{k,\ new}$.

Step 12.1: Compute $S_H{}' = s_M{}'^{x_H}$ mod p and $SK = H(S_M | S_H{}'$ mod $p)$.

Step 12.2: Obtain $\mathrm{m_M \| Y \| ID_V \| ID_M \|} V_M{}' \| \mathrm{y_{K,new} \| h(MK, m_M \| Y \| ID_V \| ID_M \|} V_\mathrm{M}{}' \| \ \mathrm{y_{K,new}})$ by decrypting $\mathrm{C_M}$ with SK.

Step 12.3: Compute $MK = H(\mathrm{SK})$ and check $h(MK,\ m_M \| Y \| ID\text{-}_V \| ID_M | V_M{}' \| y_{K,new})$ to authenticate message.

Step 12.4: Check ID_V.

Step 12.5: Validate the promise $(S_M{}',\ V_M{}')$ by the equation $V_M{}' = h(S_M{}' y_K{}^{-V_M{}'} \bmod p, m_M \| ID_M \| y_{K,new})$.

Step 13: HLR generates the response

Step 13.1: Select two random numbers $r_H{}'$ and R'.

Step 13.2: Compute $V_H{}' = h(g^{r_H{}'} S_M{}' \bmod p, m_H \| ID_H \| y_{K,new})$ and $K' = (r_H{}' - V_H{}') x_H{}^{-1}$ mod q. So the new promise of the certificate is $\sigma_H{}' = (S_H{}',\ K',\ V_H{}')$

Step 13.3: Compute $SK' = H(Y^{X_H}$ mod p).

Step 13.4: Choose a nonce n_3, compute $RK_1 = H(N_1 \| n_2 \| n_3 \| y_K)$ and store $L = N_1$.

Step 13.5: Generate and transmit $([[N_1 \| n_3 \| ID_V \| R\ ' \| \sigma_{H'}]_{SK'} \| Y_K \| n_2 \| N_1 \| RK_1]_{k_{HV}}, ID_H, ID_V)$ to VLR.

Step 14: VLR validates HLR's response.

Step 14.1: Obtain $[N_1 \| n_3 \| ID_V \| R' \| \sigma_{H'}]_{SK'} \| Y_K \| n_2 \| N_1 \| RK_1$ by decrypting the ciphertext $([[N_1 \| n_3 \| ID_V \| R' \| \sigma_{H'}]_{SK'} \| Y_K \| n_2 \| N_1 \| RK_1]_{k_{HV}}$.

Step 14.2: Check the freshness of n_2. If n_2 is not fresh, stop.

Step 14.3: Compute $ID_1 = H([N_1 \| n_3 \| ID_V \| R' \| \sigma_{H'}]_{SK'} \| RK_1)$.

Step 14.4: Store $((\rho_H,\ m_H,\ ID_H,\ ID_1),\ RK_1,\ A,\ L = N_1,\ 1)$ into its local database according to the order of ID_1.

Step 14.5: Transmit $([N_1 \| n_3 \| ID_V \| R' \| \sigma_{H'}]_{SK'}, ID_V$ to User.

Step 15: User decrypts HLR's ciphertext.

Step 15.1: Get $N_1 \| n_3 \| ID_V \| R' \| \sigma_H{}'$ by decrypting $[N_1 \| n_3 \| ID_V \| R' \| \sigma_H{}']_{SK'}$, where $SK' = H(y_H{}^x \bmod p)$.

Step 15.2: Check the freshness of N_1. If N_1 is not fresh, then stop.

Step 16: User validates the promise $\sigma_H{}'$ on new key $y_{K,new}$ by checking $V_H{}' = h(g^{V_H{}'} y_H{}^{K'} S_{H'}\ mod\ p, m_H \| ID_H \| y_{K,new})$.If $V_H{}' = h(g^{V_H{}'} y_H{}^{K'} S_{H'}\ mod\ p, m_H \| ID_H \| y_{K,new})$. then accept; otherwise, reject.

Step 17: User sets the proxy private key as $x_{K,new}$ and the proxy public key as $y_{K,new}$ for the next round.

Step 18: User computes HLR‘s Schnorr-like signature $\rho_H{}' = (S_H{}', V_H{}')$, where $S_H{}' = s_M{}' + K'$ mod q.

Step 19: User stores $(RK_1 = H(N_1||n_2||n_3||y_K)$, $(h^1(n_1)$, $h^2(n_1)$, $h^3(n_1)$, …, $h^{(n+1)}(n_1))$, y_K, round number $= 1$, $ID_1 = H([N_1 \| n_3 \| ID_V \| R' \| \sigma_{H'}]_{SK'} \| RK_1)$.) into User’s database.

When $\rho_H{}' = (S_H{}', V_H{}')$ is used, HLR recovers User’s Schnorr signature $\rho_M{}' = (s_M{}' = S_H{}' - K'$ mod q, $V_M{}')$ after obtaining $\rho_H{}'$.

Subsequent Login Authentication Phase

VLR authenticates User repeatedly without contracting HLR in the subsequent login authentication described below. Suppose that this is ith round subsequent login authentication with the session key $RK_{\rm i}$, where $i \leq n$, $RK_i = H(L, RK_{i-1})$, $RK_1 = H(N_1||n_2||n_3||y_K)$, and $L = H^{(n-i\ +\ 2)}(n_1)$. Therefore, User retrieves the record $(RK_{\rm i}$, $(h^1(n_1)$, $h^2(n_1)$, $h^3(n_1)$, …, $h^{(n+1)}(n_1))$, y_K, round number $=$ (1) first.

Step 1: User transmits $(ID_{\rm i}, [h^{(n-i+1)}(n_1)]_{RK_{\rm i}})$ to VLR.

Step 2: VLR finds $((\rho_H, m_H, ID_H, ID_i), RK_1, A, L, i)$ by searching its local database according to $ID_{\rm i}$ and obtain $h^{(n-i\ +\ 1)}(n_1)$ by decrypting $[h^{(n-i+1)}(n_1)]_{RK_{\rm i}}$

Step 3: VLR updates $L = h^{(n-i\ +\ 1)}(n_1)$, $i = i$+1, $RK_{i\ +\ 1} = h(L, RK_i)$, and $ID_{i\ +\ 1} = H(ID_i \| RK_{i\ +\ 1})$ for next round if $h(h^{(n-i\ +\ 1)}(n_1)) = L$.

Step 4: VLR sends $[ACK]_{RK_i}$back to User, where $ACK = h(h^{(n-i\ +\ 1)}(n_1) \| ID_i)$.

Step 5: User obtains ACK by decrypting the ciphertext and validates ACK by checking whether or not $ACK = h(h^{(n-i\ +\ 1)}(n_1) \| ID_i)$.

If $i = n + 1$, then User has to repeat the login authentication by setting $y_K = y_{K,new}$.

Security Analysis and Proofs

Security assumptions of our protocol are stated first. One is the symmetric encryption/decryption scheme is IND-CCA2. One is the DDH assumption that no probabilistic polynomial-time (PPT) algorithm solves DDH problem with non-negligible probability. Our protocol satisfies unlinkability, correctness, fairness, unforgeability, confidentiality, non-repudiation, mutual authentication, session key security, and User identity privacy, and User unlinkability.

Confidentiality

Our protocol satisfies IND-CCA2, if there is no PPT adversary α wins the IND-CCA2 game with the non-negligible probability.

Theorem 1: Assume the underlying symmetric encryption/decryption scheme is IND-CCA2. The proposed protocol provides IND-CCA2 if no PPT algorithm wins the IND-CCA2 game with non-negligible probability based on the DDH assumption in random oracle model.

Unlinkability Between the Promise of Signcrytext and HLR's Ciphertext

Unlinkability among difference services means no one can deduce any two different services belonging to the same User. Two cases cause linkability problem in our protocol. One case is that adversaries may find out the link between $\rho_H{'}$ and $(C_M, s_M{'})$ from the same User on two successive roaming. Since $(C_M, s_M{'})$ is the signcryptext of the promise $\sigma_M{'} = (s_M{'}, V_M{'})$, the adversary may find out the link between $\rho_H{'}$ and $\sigma_M{'}$ with the help of the link find between $\rho_H{'}$ and $(C_M, s_M{'})$, if the decryption of C_M is feasible. Therefore the link find between $\rho_H{'}$ and $(C_M, s_M{'})$ is at least harder than the link find between $\rho_H{'}$ and $\sigma_M{'}$. The link find problem between $\rho_H{'}$ and $\sigma_M{'}$ is also the link find problem in Nguyen's scheme that is our underlying scheme. Fortunately, this link find is infeasible for the unlinkability of Nguyen's scheme [12].

The other case is to find out the link between the promise of signcrytext $(C_M, s_M{'})$ and HLR's ciphertext $[N_1\|n_3\|ID_V\|R'\|\sigma_{H'}]_{SK'}$. To prove this link find is infeasible, the unlinkability game is defined first. Our protocol is unlinkabile between the promise of signcrytext $(C_M, s_M{'})$ and HLR's ciphertext $[N_1\|n_3\|ID_V\|R'\|\sigma_{H'}]_{SK'}$ if no PPT adversary α wins the unlinkability game with a non-negligible probability.

Unlinkability Game Between the Promise of Signcrytext and HLR's Ciphertext

This game has two participators, Adversary α and Challenger C. C controls all hash oracles whom α is allowed to query. These hash oracles are the same as the hash oracles in the security proof of confidentiality. This game consists of two phases: Setup, and challenging and guessing phases. In the setup phase, all systems parameters, public functions, and the public keys/private keys of all users are constructed. The public parameters and all public keys are sent to the adversary.

In the challenging and guessing phase, the adversary first chooses two anonymous public keys that may belong to two different users to the challenger. The challenges have two types. Type 0 is that the challenge consists of two promises of signcryptext using two anonymous public keys and one HLR's ciphertext for one anonymous public key. The adversary guesses which the promise matches the HLR's ciphertext. Type 1 is that the challenge consists of two HLR's ciphertexts using two anonymous public keys and one promise of signcryptext for one anonymous public key. The adversary guesses that the HLR's ciphertext matches the promise of signcryptext.

Theorem 2: Our protocol is unlinkable between the promise of signcrytext and HLR's ciphertext if no PPT algorithm wins the above unlinkability game with non-negligible probability based on the DDH assumption in random oracle model.

Fairness

The fairness in our protocol means that User (HLR) can obtain the HLR's (User's) signature but HLR (User) cannot. There are two cases violating the fairness.

Case 1: Given the promises $\sigma_M = (S_M, V_M)$ satisfying $V_M = H(S_M {y_M}^{-V_M} \mathit{mod}\, p, m_M \| ID_M \| y_K)$, HLR generates alone the corresponding $\rho_M = (s_M, V_M)$ satisfying $V_M = H(g^{s_M} {y_M}^{-V_M} \mathit{mod}\, p, m_M \| ID_M \| y_K)$..

Case 2: Given two promises $\sigma_M = (S_M, V_M)$ and $\sigma_H = (S_H, k, V_H)$ satisfying $V_M = H(g^{s_M} {y_M}^{-V_M} \mathit{mod}\, p, m_M \| ID_M \| y_K)$. and $V_H = H(g^{V_H} {y_H}^{k} S_H \mathit{mod}\, p, m_H \| ID_H \| y_K)$, respectively, User obtains $\rho_H = (s_H, V_H)$ satisfying $V_H = H(g^{V_H} {y_H}^{s_H} \mathit{mod}\, p, m_H \| ID_H \| y_K)$but HLR cannot obtain $\rho_M = (s_M, V_M)$ satisfying $V_M = H(g^{s_M} {y_M}^{-V_M} \mathit{mod}\, p, m_M \| ID_M \| y_K)$.

Theorem 3: Our protocol is fair since neither Case 1 nor Case 2 occurs based on the discrete logarithm and DDH assumption.

Unforgeability

The unforgeability means that both User's Schnorr signatures and HLR's Schnorr-like signatures are unforgeability. Fortunately, User's Schnorr signatures are existentially unforgeable against chosen-message attacks [13, 14]. HLR's Schnorr-like signatures are unforgeability [12].

Our protocol also satisfied incorrectness, on-reputation, session key security, the mutual authentication, users' identity privacy, and users' unlinkability.

Conclusion

By utilizing our concurrent signcryption, HLR cannot fore users' signatures during roaming. Consequently, the users' non-repudiation is enhanced. Moreover, the performance of the subsequent login authentication is improved by removing the exhaustive search problem when VLR has to find the alias of an anonymous User. Since the User adopts different aliases to login VLR in subsequent login authentication each time, our protocol supposes the subsequent login User unlinkability to protect the User identity privacy. The concurrent signcryption scheme also supposes the fairness property to fairly protect User's and HLR's benefits.

References

1. Lee W-B, Yeh C-K (2005) A new delegation-based authentication protocol for use in portable communication systems. IEEE Trans Wireless Commun 4(1):57–64
2. Al-Tawill K, Akrami A, Youssef H (1999) A new authentication protocol for GSM networks. In: Proceedings of 23rd annual IEEE conference on local computer networks, pp 21–30
3. Lee C-H, Hwang M-S, Yang W-P (1999) Enhanced privacy and authentication for the global system for mobile communications. Wireless Netw 5(4):231–243
4. Beller MJ, Chang L-F, Yacobi Y (1993) Privacy and authentication on a portable communication system. IEEE J Sel Areas Commun 11(6):821–829
5. Lo C-C, Chen Y-J (1999) Secure communication mechanisms for GSM networks. IEEE Trans Consum Electron 45(4):1074–1080
6. Lee T-F, Chang S-H, Hwang T, Chong SK (2009) Enhanced delegation-based authentication protocol for PCSs. IEEE Trans Wireless Commun 8(5):2166–2171
7. Youn T-Y, Lim J (2011) Improved delegation-based authentication protocol for secure roaming service with unlinkability. IEEE Commun Lett 14(9):791–793
8. Wang R-C, Juang W-S, Lei CL (2011) A privacy and delegation-enhanced user authentication protocol for portable communication systems. Int J Ad Hoc Ubiquitous Comput 6(3):183–190
9. Hwang S.-J, You C.-H (2011) Weakness of Wang et al.'s privacy and delegation enhanced user authentication protocol for PCSs," CSCIST 2011 and iCube 2011, Taipei, 2011
10. Chen H.-B, Lai Y.-H, Chen K.-W, Lee W.-B (2011) Enhanced delegation based authentication protocol for secure roaming service with synchronization. J Electronic Sci Technol 9(4), pp 345–351
11. Hwang S-J, Sung Y-H (2011) Confidential deniable authentication using promised signcryption. J Syst Softw 84:1652–1659
12. Nguyen K (2005) Asymmetric concurrent signatures. In: Proceedings of information and communications security conference (ICICS 2005), LNCS 3783, Springer, New York, pp 181–193
13. Schnorr C (1991) Efficient signature generation by smart cards. J Cryptology 14(3), pp 161–174
14. Pointcheval D, Stern J (2000) Security arguments for digital signatures and blind signatures. J Cryptography 13(3), pp 361–396

A Lightweight Mutual Authentication Protocol for RFID

Changlun Zhang and Haibing Mu

Abstract In order to protect the tag and its communication, the authentication between the tag and the reader as well as its backend database is necessary. The paper proposed a mutual authentication protocol by introducing pointer and check number pool which make a simple random number to hide the ID information and disturb the static answers in challenge-response of the protocol. The check number pair selected from the pool in each authentication turn is also pointed out by a random number which controls the shifting of the pointer in pool. The analysis shows that the protocol can resist the common attacks in RFID communication with low computation overhead.

Keywords IoT · RFID · Check number pool · Pointer · Security

Introduction

The Internet of Things (IoT) is another milestone after the Internet in the development of information technology. It combines the virtual Internet with human and objects in real world to achieve the interrelation of human with human, human with object and objects with objects. The terminal nodes in IoT are extended to a large number of sensors, Radio Frequency Identification (RFID) labels and intelligent equipment besides computers and human. The recipient and analyzer of

C. Zhang
Science School, Beijing University of Civil Engineering and Architecture,
Beijing 100044, China
e-mail: zclun@bucea.edu.cn

H. Mu (✉)
Beijing Key Laboratory of Communication and Information Systems, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
e-mail: hbmu@bjtu.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_106, © Springer Science+Business Media Dordrecht 2014

the information in the network may be the device itself rather than the thinking man. The device has limited ability to identify the information source and need new authentication protocols the support the secure communication in IoT.

RFID is a main kind of IoT terminals and provides non-contact data transmission with a certain distance which makes an attack window for intruder. We must implement mutual authentication between tag and reader before data transmission to protect the communication. Due to the limited resources of the RFID label, an ideal authentication protocol should be secure, efficient and lightweight. The simplest form of authentication protocols in the RFID system works in this way: the reader sends inquiry request information to scan the target label, and label responds ID items to help identify their own identity. Although this form of authentication protocol is simple and fast, it is vulnerable to forgery, eavesdropping, tracking, impersonation and other attacks. An RFID authentication protocol should be of confidentiality and forward security to prevent tracking, forging and cloning with good certifying efficiency.

Related Work

In recent years, more and more research focuses on the IoT and the RFID with its security. A variety of RFID security authentication protocols have been proposed. Hash-Lock protocol [1, 2] by Sarma et al. proposed a security protocol, which basically solved the privacy protection and access control, and can be implemented on low-cost tags for the tag needs only hash operation. The data of the answer in each turn of the challenge remains unchanged which made the tag is vulnerable to position tracking attack. Furthermore, the tag with an ID sent in plain text through an insecure channel will be vulnerable to replay and forgery attack. Random Hash-Lock protocol [3] is a modified form for Hash-Lock protocol, the use of random numbers prevents location tracking according to the particular output. But the ID with plain text makes it lack of forward security and an attacker can forge the tag once it got the value of the tag's ID. The protocol also need a large quantity of data communication and computation to get H(IDj||R) for each tag. The large amount of calculation, the efficiency is not high. Hash chain protocol [4] is also based on shared secret inquiry—response protocol with indistinguishability and forward security. The protocol implements one-way authentication for tag and is vulnerable to replay and forgery attack. In addition, each authentication requires a larger amount of computation and comparison, and is not suitable for a large number of tags in the system. Distributed request-response authentication protocol [5] is a mutual authentication protocol against eavesdropping, replay, tracking and other common attacks with better security. However the main drawback of the protocol is the large number of complex function computation according to the number of the readers and the tags.

Recent studies [6–10] introduce pseudo-random number generator and cyclic redundancy code in EPC to improve efficiency and reduce the overhead at the same time. Tian et al. [11] proposed an ultra-lightweight RFID authentication protocol with permutation. The tags only operate bitwise XOR, left rotation and permutation, a new defined rule that can make diffusion and confusion. In this way, the tags avoid using unbalanced OR and AND operations to improve security. Pietro and Molva [12] made two contributions within their new identification and authentication protocol. Firstly, when the servers can be compromised, they devised a technique that makes RFID identification sever-dependent that can protect the tags from malicious servers. Secondly, a probabilistic tag identification scheme was proposed only using bitwise operations that can speed up the authentication process. Inspired by [12], Blass et al. [13]. put forward a lightweight F_f family of privacy-preserving authentication protocols for RFID system. The protocol offers strong authentication. Although it is based on a algebraic framework, it can resist algebraic attacks. Different from lightweight authentication protocols for RFID system, AX Liu and LRA Bailey [14] put forward a privacy and authentication protocol (PAP). In this paper, they distinguished privacy from authentication. The protocol was placed in a concrete environment, store. It requires only an extremely small amount of computations and can protect privacy from malicious attacks. Molnar and Wagner [15] suggested novel architectures to solve privacy issues related to RFID in libraries. They gave a general scheme to build private authentication using work logarithmic in the number of tags. In addition, they also presented a simple scheme that provided security against a passive eavesdropper using XOR, without any pseudo-random functions. Lee et al. [16]. presented a RFID authentication protocol using both XOR and hash chains to realize low-cost and high security. The scheme can resist to traceability and cloning.

But because of the new characteristics of RFID based IoT, these protocols still cannot meet special environmental requirements in the application. In this paper, a new protocol based on the index is proposed to reduce the computation overhead in tag and database as well as provide necessary security for RFID. The paper is organized as follows: section The Authentication Protocol Based on Check Number Pool describes the details of the protocol, and section Analysis of the Protocol analyzes security and performance of the protocol. At last the conclusion is provided in section Conclusion.

The Authentication Protocol Based on Check Number Pool

The protocol combines small pseudo-random verification number and the random number generator to achieve nearly random number function with low computation overhead.

Initialization and Parameters

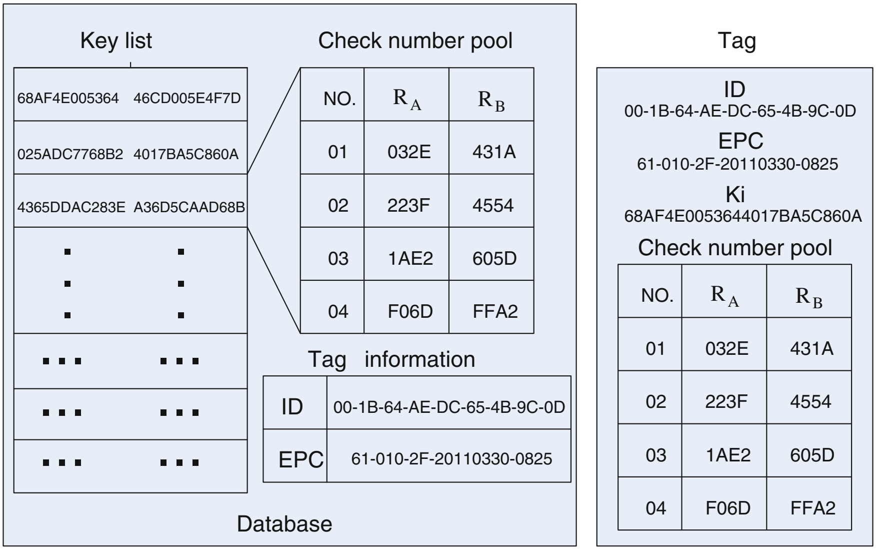
There are several parameters to be generated during the system initialization phase, such as encryption and decryption keys K_i and the corresponding check number pools. There are records in each pool with three factors in each record which are the serial number and the random number R_A and R_B. Each RFID tag is written with a key and its corresponding check number pool as well as its ID and EPC information. At the same time, the ID of the tag is also synchronized to the backend database, so that each tag ID is combined with the corresponding key and check number pool. There is a pointer in backend database to keep the synchronization of the pool.

The initialization and the parameters are shown in Fig. 1.

Detail of the Authentication

The authentication process between the RFID tag and the reader supported by backend database can be described as following communication interaction.

1. The reader generates a random number R and sends it together with the communication request to the target tag.



Key list

68AF4E005364	46CD005E4F7D
025ADC7768B2	4017BA5C860A
4365DDAC283E	A36D5CAAD68B
⋮	⋮
...	...
...	...
...	...

Check number pool

NO.	R_A	R_B
01	032E	431A
02	223F	4554
03	1AE2	605D
04	F06D	FFA2

Tag information

ID	00-1B-64-AE-DC-65-4B-9C-0D
EPC	61-010-2F-20110330-0825

Database

Tag

ID
00-1B-64-AE-DC-65-4B-9C-0D
EPC
61-010-2F-20110330-0825
Ki
68AF4E0053644017BA5C860A

Check number pool

NO.	R_A	R_B
01	032E	431A
02	223F	4554
03	1AE2	605D
04	F06D	FFA2

Fig. 1 Initialization and the parameters

2. The tag receives R and takes the check number R_A according to the current pointer. It then calculates $P = E_{ki}(R \oplus R_A \oplus ID)$and sends the P and R_A to the reader.
3. Reader receives the data and sends P, R, R_A to the back-end database.
4. The database searches the check number pool and gets the subset according to R_A. It then computes $P' = E_{ki}(R \oplus R_A \oplus ID')$ of each record using R, R_A with the corresponding key K_i', and compares results with received P respectively. If there is not any matching record, the tag will be considered as illegal and communication ends. If there is any result equal to P, the back-end database will get the ID of the tag and prove it a legal one.

After authentication, the database will send to the reader another random number R_B in this pair of check number. In order to update the exchange information during authentication turn, the generator in database will generate a random number r to shift the pointer value. The new pointer and the previous one are stored in database until next turn. Database also sends this value to the reader in order to synchronize the pointer in the tag.

5. The reader receives the information and then sends R_B and r to tag.
6. Tag compares the received R_B with the stored one in the pool according to the current pointer. If the two numbers are of the same, the reader is considered as legal one. And the pointer is updated according to r. Check number pools are synchronized completely. Otherwise, if this communication is failed, the tag will use the old pointer in next authentication. This may be also successful because the previous one is still stored in database and the authentication achieved.

Details of one turn of the authentication are shown in Fig. 2.

Analysis of the Protocol

For any newly proposed security protocol, analysis of the protocol on whether it can protect the entities in RFID system from existing security threats or not is necessary.

1. Eavesdropping

Eliminating eavesdropping in most wireless networks thoroughly is impossible. Authentication protocols just limit the amount of information transmitted in unsecure physical channel only. By encrypting authentication messages, communication protocol can lower the value of information eavesdropped by malicious adversary. In this way, effectiveness of such an attack can be reduced.

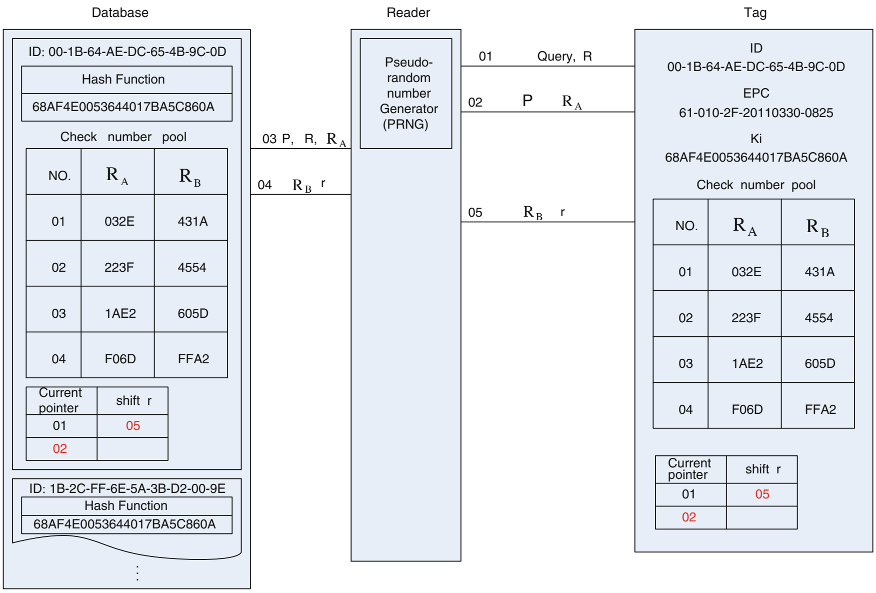


Fig. 2 Authentication detail

2. Replay attack

Our protocol can resist replay attack effectively. The main weakness of random Hash-Lock protocol is lead by the authentication message from reader to tag, which is the static ID of a tag. Our protocol did not send ID and any other stead information during authentication instead of random numbers.

If an adversary can capture a large number of verified sequence pairs sent during every authentication process, when a repeated verified sequence is captured, the adversary will be clear about the corresponding response. Aiming at the situation above, an adversary has to monitoring a mass of verified sequence pairs. However, it is impractical for most of the adversaries.

3. Dos attack

There is no need to compute P for every ID in this protocol, because there is a process of retrieving matching values before computing. There are only simple computations of retrieving instead of a large number of computations. Computing on P won't start until the matching values are obtained. Even if there are amount of challenge information, it can be confirmed whether the data is true or fault by simply filtering, because the adversary doesn't know an accurate RA of next hop. Although Dos attack is possible, the probability of its success has been lower comparing with original protocol.

4. Tracking

This protocol can be effective against identity tracking. Tag ID information cannot be obtained by attacker because it is not transmitted in plain text or other steady forms. The responses to the challenge are no longer the same ID information, but randomized check number thus avoids being tracked.

5. De-synchronization attack

The database will store the previous pointer to shift the check number pairs for one more turn in order to avoid the de-synchronization for the failure of the last information sent to tag.

Although it is impossible to design a security protocol to solve all the problems, different protocol will be suitable to different environment.

Conclusion

In this paper, an authentication protocol for RFID security in IoT focus on improving the original random Hash-Lock protocol security issues, especially against replay attacks. It introduces check number pool with shift pointer which makes a simple random number pair, effectively improve the randomized Hash-Lock protocol against replay, tracking, DoS and other attacks. At the same time the protocol improves the performance in computing greatly by matching check number firstly before other operation. The further work of the protocol is to get more simulation data to compare and analyze its performance in more detail.

Acknowledgments This work is supported by National Natural Science Foundation of China under Grant 61201159, the Beijing Municipal Organization Department of talents training-funded project (2010D005017000008), Beijing Institute of Architectural Engineering School research fund (Z10053) and Jilin University Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education research fund (93 K-17-2012-02).

References

1. Sarma SE, Wreis SA, Engels DW (2003) Radio frequency identification: Secure risks and challenges. RSA Laboratories Cryptobytes 6(1):2–9
2. Sarma SE, Wreis SA, Engels DW (2003) RFID systems and security and privacy implications. In: Proceedings of the 4th intemational workshop on cryptographic hardware and embedded systems, pp 454–469
3. Juels A, Rivest RL, Szydlo M (2003) The blocker tag: selective blocking of RFID tags for consumer privacy. In: Proceedings of the 10th ACM conference on computer and communications security, Washington DC, USA, pp 103–111
4. Ohkubo M, Suzuki K, Kinoshita S (2004) Hash-chain based forward-secure privacy protection scheme for low-cost RFID. In: Proceedings of the 2004 symposium on cryptography and information security (scis 2004), Sendai, pp 719–724

5. Rhee K, Kwak J, Kim S (2005) Challenge-response based RFID authentication protocol for distributed database environment. In: Proceedings of the 2nd international conference on security in pervasive computing (SPC 2005). Lectures Notes in Computer Science 3450. Springer, Berlin, pp 70–84
6. Duc DN, Park J, Lee H, et a1 (2006) Enhancing security of EPC global gen 2 RFID tag against traceability and cloning. In: Symposium on cryptography and information security-SCIS 2006, Hiroshima, Japan
7. Chien H, Chen C (2007) Mutual authentication protocol for RFID conforming to EPC class 1 generation 2 standards. Comput Stand Interfaces 29(2):254–259
8. Yuan S, Dai H, Lai S (2008) Hash based RFID authentication protocol. Comput Eng 12:51
9. Yang L, Chen Z (2010) A mutual authentication protocol for low-cost RFID
10. Juels A (2007) RFID security and privacy: a research survey. IEEE J Sel Areas in Commun 24:381
11. Tian Y, Chen G, Li J (2012) A new ultralightweight RFID authentication protocol with permutation. Commun Lett IEEE 16(5):702–705
12. Di Pietro R, Molva R. Information confinement, privacy, and security in RFID systems. In: Computer security–ESORICS 2007. Springer, Berlin, pp 187–202
13. Blass EO, Kurmus A, Molva R et al (2011) The F_f-family of protocols for RFID-privacy and authentication. IEEE Trans Dependable Secure Comput 8(3):466–480
14. Liu AX, Bailey LRA (2009) PAP: a privacy and authentication protocol for passive RFID tags. Comput Commun 32(7):1194–1199
15. Molnar D, Wagner D (2004) Privacy and security in library RFID: issues, practices, and architectures. In: Proceedings of the 11th ACM conference on computer and communications security, pp 210–219
16. Lee S, Asano T, Kim K (2006) RFID mutual authentication scheme based on synchronized secret information. In: Symposium on cryptography and information security

An Approach for Detecting Flooding Attack Based on Integrated Entropy Measurement in E-Mail Server

Hsing-Chung Chen, Shian-Shyong Tseng, Chuan-Hsien Mao, Chao-Ching Lee and Rendabel Churniawan

Abstract The aim of this study is to protect an electronic mail (email) server system based on an integrated Entropy calculation via detecting flooding attacks. Lots of approaches have been proposed by many researchers to detect packets accessing email whether are belonging to the normal or abnormal packets. Entropy is an approach of the mathematical theory of Communication; it can be used to measure the uncertainty or randomness in a random variable. A normal email server usually supports the four protocols consists of Simple Mail Transfer Protocol (SMTP), Post Office Protocol version 3 (POP3), Internet Message Access Protocol version 4 (IMAP4), and HTTPS being used by remote web-based email. However, in Internet, there are many flooding attacks will try to paralyze email server system. Therefore, we propose a new approach for detecting flooding attack based on Integrated Entropy Measurement in email server. Our approach can reduce the misjudge rate compared to conventional approaches.

Keywords Entropy · Flooding attack · Email server

H.-C. Chen (✉) · S.-S. Tseng · C.-H. Mao · C.-C. Lee · R. Churniawan
Department of Computer Science and Information Engineering, Asia University, Taichung 41354, Taiwan, Republic of China
e-mail: shin8409@ms6.hinet.net; cdma2000@asia.edu.tw

S.-S. Tseng
e-mail: sstseng@asia.edu.tw

C.-H. Mao
e-mail: 101267005@live.asia.edu.tw

C.-C. Lee
e-mail: johnson10723@gmail.com

R. Churniawan
e-mail: rendabel@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_107, © Springer Science+Business Media Dordrecht 2014

Introduction

In recent year, the rapid development of technologies helps people to communicate any information and sharing information via Internet. Email has become one of necessary communication services for Internet users. The using of Electronic Mail (email) is a method of exchanging digital messages from one person to one or more recipients, via connecting internet or computer network. There are many kind purposes of using email services, from private purposes to business purposes. Email Service Provider (ESP) is an organization which provides email server to send, receive and store emails for personal and or organization necessity. Some ESP who may provide the services to general public to personal email are Gmail, Yahoo! Mail, Hotmail and many others.

Each email server is able to support many kind of protocol. In 1982, the early stage of email development, the Simple Mail Transfer Protocol (SMTP, for short) which is formulated in RFC (Request for Comments) 821 [1, 2]. SMTP is a protocol for a mail sender communicates with a mail receiver. On certain types of smaller nodes in the Internet it is often impractical to maintain a message transport system [3]. For example, a workstation may not have sufficient resources (cycles, disk space) in order to permit a SMTP server [RFC821] [3]. To solve this problem, The Post Office Protocol—Version 3 (POP3, for short) is intended to permit a workstation to dynamically access a mail drop on a server host in a useful fashion. Usually, this means that the POP3 protocol is used to allow a workstation to retrieve mail that the server is holding for it. POP3 is an application layer Internet standard protocol used by local email clients to retrieve email from remote server over a TCP/IP connection. The other protocol which is also supported by email server is Internet Message Access Protocol (IMAP, for short) [4]. In 2003, M. Crispin [5] proposed the latest version IMAP version 4rev1 (IMAP4, for short) which is formulated in RFC 2060. It has also been publish to the Internet Engineering Task Force (IETF). The MAP4 allows a client to access and manipulate electronic mail messages on a server [5]. IMAP4 permits manipulation of remote message folders, called "mailboxes", in a way that is functionally equivalent to local mailboxes. IMAP4 also provides the capability for an offline client to re-synchronize with the server. There is the other approach for supporting email services called webmail (or web-based email) [6, 7]. It is any email client implemented as a web application accessed via a web browser. For example, when accessing webmail at https://webmail.asia.edu.tw, you will be redirected to a SSL [6] secured address and your connection will be encrypted. The Secure Sockets Layer (SSL) Protocol Version 3.0 was proposed by P. Karlton [6], in 2011, which is formulated in RFC 6101[6]. It is a security protocol that provides communications privacy over the Internet via HTTPS (Hypertext Transfer Protocol Secure). The protocol allows client/server applications to communicate in a way that is designed to prevent eavesdropping, tampering, or message forgery [6].

However, a simple email server has a lot of users; it is an important attacked target in Internet. The methods of attacks include SMTP Flooding Attack, spam attacks and the malicious attachment etc. in email [8–12]. The various flooding attacks will increase the loading of the server. In this paper, we propose an approach for detecting flooding attack based on integrated entropy measurement in email server. Then, we use the entropy operations [13–15] to analyse the received packets, in order to estimate normal packets and abnormal packets from SMTP together with POP3, IMAP4, HTTP messages flows, and then evaluate its corresponding risk information. Therefore, the risk information will be used to describe the status of the serving server. According to the value of this status, the server will determine whether it is suffered by flooding attacks.

The remainder of this paper is organized as follows. "Related Work" describes the SMTP, POP3, IMAP4 and entropy operation related work. In the "Email Server Prevention Against Flooding Attack Based on Entropy", we propose a new approach for detecting flooding attack based on integrated entropy measurement in email server, and describe how to calculate the evaluate value of risk information of email server. Finally, we draw conclusions in "Conclusions".

Related Work

In our proposed approach, we use the entropy measurement to detect the behaviour of the SMTPFA. Therefore, in "Related Work", we will describe the normal message flows of SMTP standard [1, 2], and the entropy operations [13–15].

SMTP

First, SMTP had been defined in the RFC 821[1, 2, 16]. It is an independent subsystem in special communication system. In this communication system, it only needs a reliable channel to transmit the related sequence message flows. SMTP has an important simple delivering email protocol which it can forward an email between two different networks. The architecture of SMTP is shown in Fig. 1. In the SMTP architecture [16], it consists of a Sender, a sender-SMTP, a receiver-SMTP and a Receiver. When a Sender (user or file server) will connect to another receiver, it will send a request message of Establishes Connection to the sender-SMTP. Then, the sender-SMTP will establish a two-way transmission channel in order to connect the Receiver. The receiver-SMTP will be as a destination point or a relay point. Thus, the sender-SMTP will send the related SMTP commands to the receiver-SMTP. Finally, the receiver-SMTP will follow these commands to send back a SMTP response message to sender-SMTP. According to the above steps, if the command-respond pair has been completed during one normal time-period, it means that a round of SMTP session has been completed. The established SMTP message flows are divided into seven stages [16] as below: Establishes Connection, HELO, MAIL FROM, RCPT TO, DATA, DATA TRANSFER, and QUIT. The SMTP message flows are shown in Fig. 2 [16].

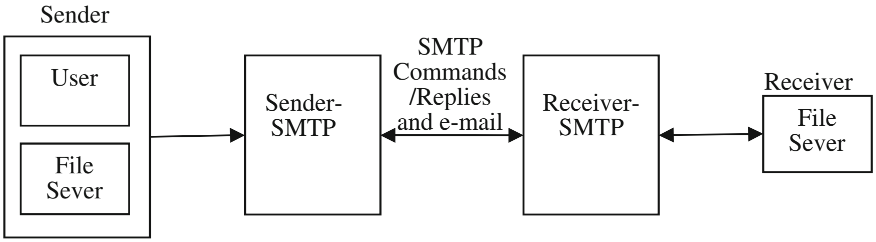


Fig. 1 The SMTP architecture [1, 2, 16]

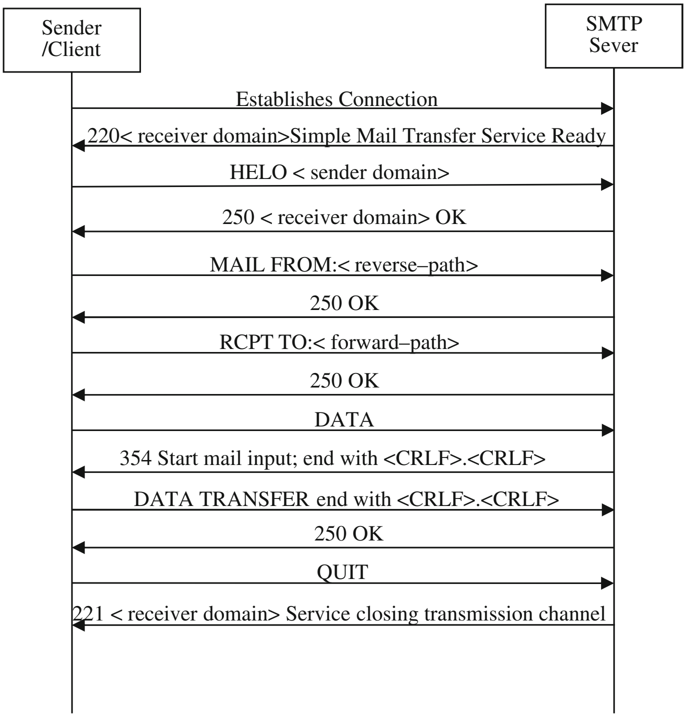

Sender /Client
SMTP Sever
Establishes Connection
220< receiver domain>Simple Mail Transfer Service Ready
HELO < sender domain>
250 < receiver domain> OK
MAIL FROM:< reverse–path>
250 OK
RCPT TO:< forward–path>
250 OK
DATA
354 Start mail input; end with <CRLF>.<CRLF>
DATA TRANSFER end with <CRLF>.<CRLF>
250 OK
QUIT
221 < receiver domain> Service closing transmission channel

Fig. 2 SMTP message flows [1, 2, 16]

POP 3

Post Office protocol (POP) [3] is an application layer internet standard protocol used by local email clients to retrieve email from a remoter server over TCP/IP connection. The POP 3 messages Flow is shown as Fig. 3 [3]. The POP3 flow of Transaction state is shown as Fig. 4.

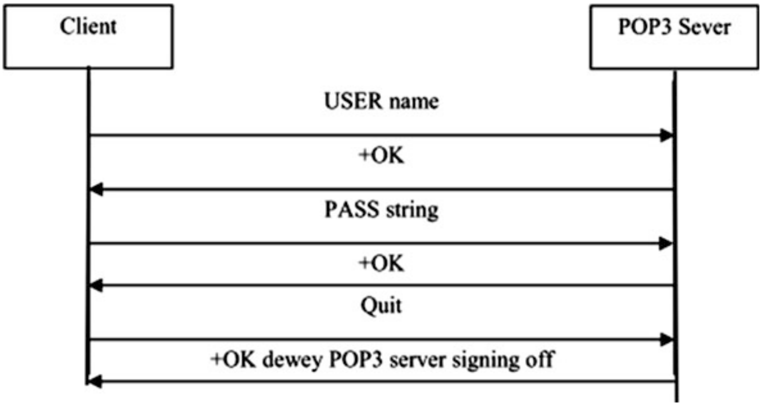


Fig. 3 POP3 flow Authorization state [3]

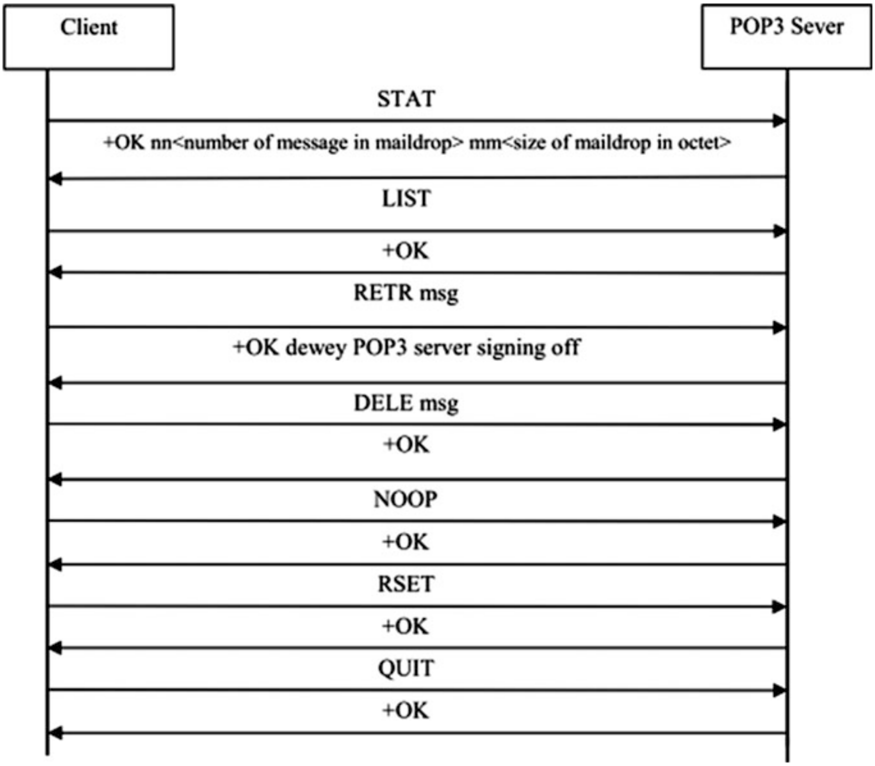

Client
POP3 Sever
STAT
+OK nn<number of message in maildrop> mm<size of maildrop in octet>
LIST
+OK
RETR msg
+OK dewey POP3 server signing off
DELE msg
+OK
NOOP
+OK
RSET
+OK
QUIT
+OK

Fig. 4 POP3 flow Transaction state [3]

Internet Message Access Protocol

Internet Message Access Protocol (IMAP) is one of two general protocols for receiving electronic mail. The Internet Message Access Protocol version 4rev1 (IMAP4rev1) [5] allow a client to access and manipulate electronic mail message in the server. IMAP4rev1 permits manipulation of mailboxes in a way that is functionally equivalent to local folders.

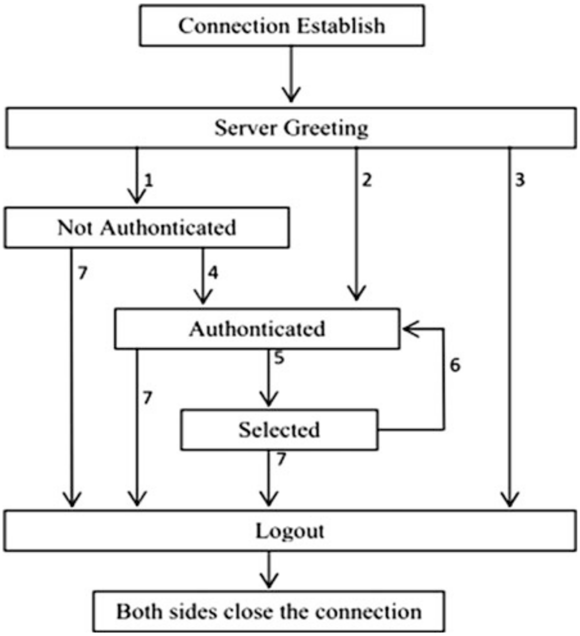
a. IMAP Message Flow

IMAP version 4 has been defined in RFC 3501 [5]. This protocol is an interaction of client/server; consist of a client command, server data and completion result response. The main architecture of IMAP is shown as Fig. 5.

where the information means as below.

1. Connection without pre-authentication (OK greeting)
2. Pre-authenticated connection (PREAUTH greeting)
3. Rejected connection (BYE greeting)
4. Successful LOGIN or AUTHENTICATE command
5. Successful SELECT or EXAMINE command
6. CLOSE command, or failed SELECT or EXAMINE command
7. LOGOUT command, server shutdown, or connection closed

Fig. 5 IMAP architecture [5]



b. IMAP Client Commands—Any State

In the any state client command, there are three commands: CAPABILITY, NOOP, and LOGOUT. These command is always be used in NOT Authenticated state, Authenticated state, and selected state. In the communication flows of client server for any state command are shown in Fig. 6.

c. IMAP Client Commands—Not Authenticated State

In the not authenticated state, the AUTHENTICATED or LOGIN command establishes authentication and enters the authenticated state. The AUTHENTICATE command provides a general mechanism for a variety of authentication techniques, privacy protection, and integrity checking [5].

The STARTTLS command is an alternate form of establishing session privacy protection and integrity checking, but doesn't establish authentication or enter the authenticated state.

Server implementations may allow access to certain mailboxes without establishing authentication. This can be done by means of the ANONYMOUS authenticator. An older convention is a LOGIN command using use rid "anonymous"; in this case, a password is required although the server may choose to accept any password. The restrictions placed on anonymous user are implementation-dependent [5].

Once authenticated, it is not possible to re-enter not authenticated state. In addition the universal command (any state command) is valid in the NOT Authenticated state. For description of not authenticated state client server communication, is shown in the Fig. 7 as below.

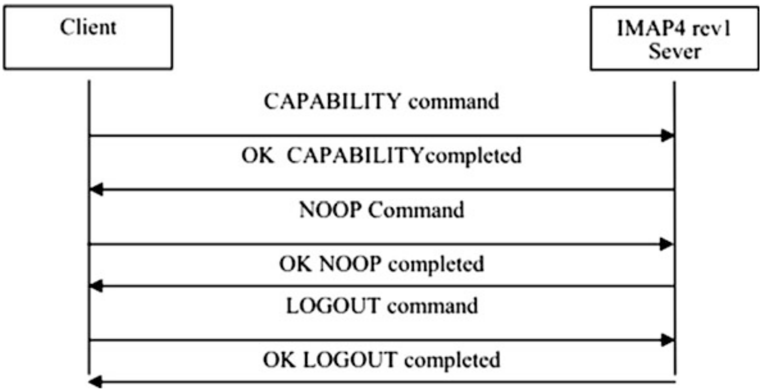


Fig. 6 IMAP flow any state [5]

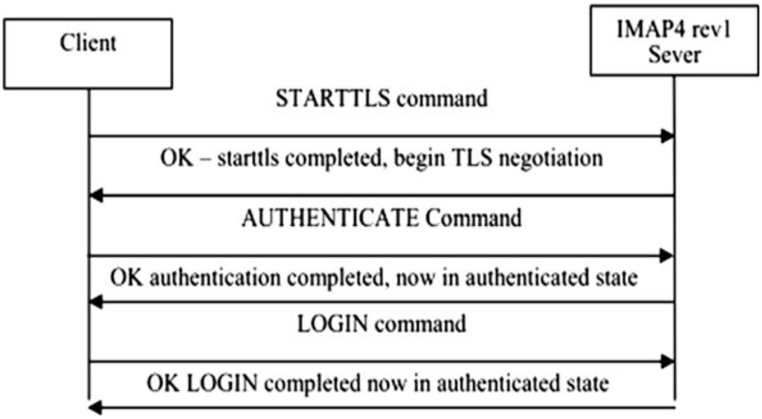

Client
IMAP4 rev1 Sever
STARTTLS command
OK – starttls completed, begin TLS negotiation
AUTHENTICATE Command
OK authentication completed, now in authenticated state
LOGIN command
OK LOGIN completed now in authenticated state

Fig. 7 IMAP flow not authenticated state [5]

d. IMAP Client Commands—Authenticated State

In the authenticated state, commands that manipulate mailboxes are permitted. Of these commands, the SELECT and EXAMINE commands will select a mailbox for access and enter the selected state.

In authenticated state, the universal commands (CAPABILITY, NOOP, and LOGOUT), are also valid to be used. And for the commands in authenticated state are: SELECT, EXAMINE, CREATE, DELETE, RENAME, SUBSCRIBE, UNSUBSCRIBE, LIST, LSUB, STATUS, and APPEND. For a description of client server communication in the authenticated state for IMAP, is shown in the Fig. 8.

HTTPS

HTTPS is a communications protocol for secure communication over Internet. Technically, it is not a protocol in and of itself; rather, it is the result of simply layering the Hypertext Transfer Protocol (HTTP) on top of the SSL/TLS (Secure Sockets Layer/Transport Layer Security) protocol [6], thus adding the security capabilities of SSL/TLS to standard HTTP communications. In its popular deployment on the internet, HTTPS provides authentication of the web site and associated web server that one is communicating with, which protects against man-in-the-middle attacks [7].

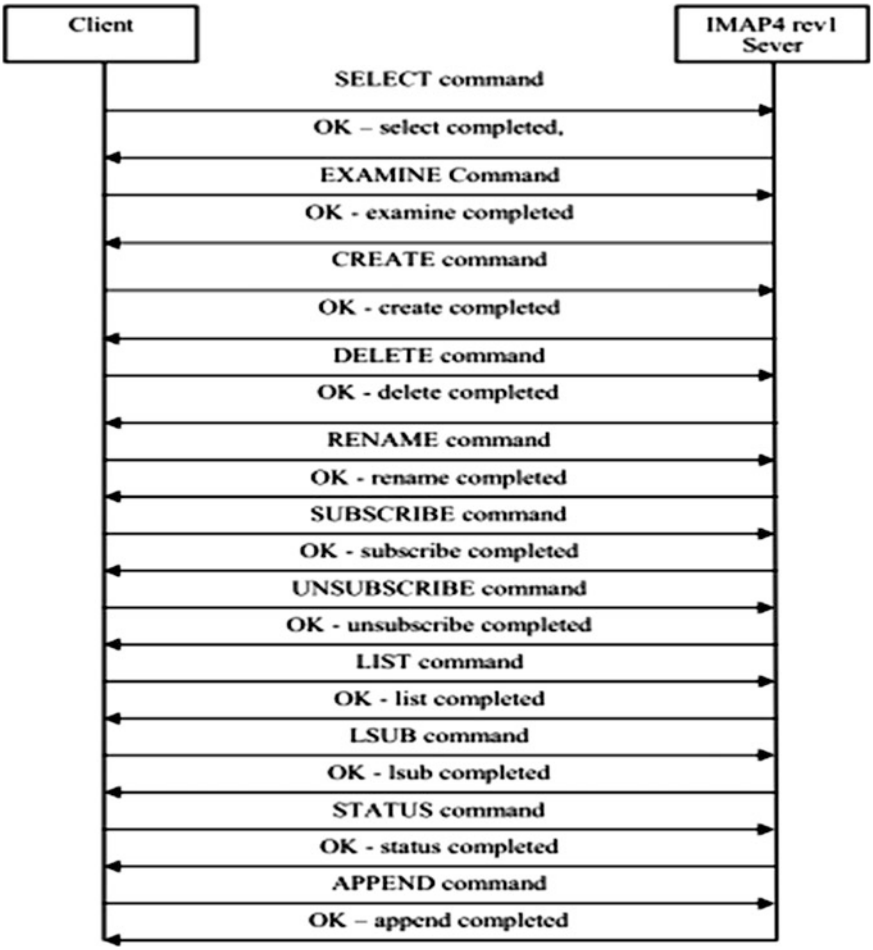


Fig. 8 IMAP flow authenticated state [5]

Entropy Operation

In the Information Theory, Entropy is an approach used to measure the uncertainty or randomness in random variable [13–16]. Entropy measurement approach is proposed by Shannon [13] and Weaver [15]. In the entropy operation, a random entropy value $X \in \{x_1, x_2, x_3, \ldots, x_n\}$, the entropy calculation formula [13–16] as below:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{1}$$

where $P(x_i) = \frac{m_i}{m}$, $m = \sum_{i=1}^{n} m_i$, m_i is the observation frequency or numbers of the x_i from X. It can represent [13–16] as:

$$H(X) = -\sum_{i=1}^{n} \left(\frac{m_i}{m}\right) \log P\left(\frac{m_i}{m}\right) \tag{2}$$

From the example above, we know that the coin is thrown according to the positive and negative probability to determine the entropy. The entropy value is inversely proportional to the probability value. With this feature, the value of results we calculate has dependability [16].

Email server Prevention Against Flooding Attack Based on Entropy

In approach of detecting flooding attack in email server, each protocol message flows are divided into two one-group pairing between server and client. Then, we use the entropy operations to calculate the entropy values of normal packets and abnormal packets for each protocol, individually. The formula for normal packet and abnormal packet are listed as below.

$$Entropy : H(X_{K_{T_w},n}) = -\left(\sum_{i=1}^{t} \frac{P_{(K_{T_w},n)}}{P_{K_{T_w}}} \log_2 \frac{P_{(K_{T_w},n)}}{P_{K_{T_w}}}\right), \quad T_w = 1, 2, \ldots, t; \tag{3}$$

$$Entropy : H(X_{K_{T_w},a}) = -\left(\sum_{i=1}^{t} \frac{P_{(K_{T_w},a)}}{P_{K_{T_w}}} \log_2 \frac{P_{(K_{T_w},a)}}{P_{K_{T_w}}}\right), \quad T_w = 1, 2, \ldots, t; \tag{4}$$

where

$X_{K_{T_w},n} \in \{x_{K_{T_w},n1}, x_{K_{T_w},n2}, x_{K_{T_w},n3}, \ldots, x_{K_{T_w},nt}\}$ is a random entropy value of normal packet during sampling time duration T_w, and $X_{K_{T_w},a} \in \{x_{K_{T_w},a1}, x_{K_{T_w},a2}, x_{K_{T_w},a3}, \ldots, x_{K_{T_w},at}\}$ is a random entropy value of abnormal packet during sampling time duration T_w;
$H(X_{K_{T_w},n})$ is an Entropy value set for normal packet; $H(X_{K_{T_w},a})$ is an Entropy value set for abnormal packet;
$K \in \{\text{SMTP, IMAP4, POP3, HTTPS}\}$;
$P_{(K_{T_w},n)} + P_{(K_{T_w},a)} = 1$, where $P_{(K_{T_w},n)}$ means the probability of normal packets of K during a sampling time duration T_w; $P_{(K_{T_w},a)}$ means the probability of abnormal packets of K during a sampling time duration T_w.

According to formula (3) and (4), we define the *Evolution Algorithm* as below.

Evolution Algorithm

Input: $\left(H(X_{K_{T_w},n}),\ H(X_{K_{T_w},a})\right)$, an email service message flow pair is including two Entropy value for normal packets and abnormal packets during in sampling time duration T_w, where $T_w = 1, 2, \ldots, t$.

Output: Result cost values S of the email server status in T_w. The cost values $\mathrm{C}_{K_{T_w}} \in \{Critical\ High,\ Very\ High,\ High,\ Normal,\ Low,\ Very\ Low\} = \{CH,\ VH,\ H,\ N,\ L,\ VL\}$ are the evolution in order to support the further judgements for the email whether under the flooding attacks or not. Also, the result will show what kind email protocol if the system is under flooding attacks.

Begin

```
$x_{K,T_w} \leftarrow \max\left(-\sum_{i=1}^{t} \frac{P_{(K_{T_w},n)}}{P_{K_{T_w}}} \log_2 \frac{P_{(K_{T_w},n)}}{P_{K_{T_w}}}\right),\ T_w = 1, 2, \ldots, t;$
$y_{K,T_w} \leftarrow \max\left(-\sum_{i=1}^{t} \frac{P_{(K_{T_w},a)}}{P_{K_{T_w}}} \log_2 \frac{P_{(K_{T_w},a)}}{P_{K_{T_w}}}\right),\ T_w = 1, 2, \ldots, t;$
if $\left(y_{K,T_w} - x_{K,T_w}\right) \gg 0$ then return $\mathrm{C}_{K_{T_w}} = CH;$
else if $\left(y_{K,T_w} - x_{K,T_w}\right) > 0$ AND $\left(y_{K,T_w} - \frac{y_{K,T_{w-1}} + y_{K,T_{w-2}} + y_{K,T_{w-3}}}{3}\right) > 0$, then return VH;
return H;
if $\left(y_{K,T_w} - x_{K,T_w}\right) < <0$ then return $\mathrm{C}_{K_{T_w}} = VL;$
else if $\left(y_{K,T_w} - x_{K,T_w}\right) < 0$ AND $\left(x_{K,T_w} - \frac{x_{K,T_{w-1}} + x_{K,T_{w-2}} + x_{K,T_{w-3}}}{3}\right) > 0$ then return L;
return N;
```

End;

□

Finally, the algorithm returns the cost values for the current sampling time duration T_w. The cost values can indicate the evolution cost in order to support the judgements for the email server whether under the flooding attacks or not. Also, the result will show what kind email protocol if the system is under flooding attacks.

Conclusions

In this paper, we propose a new approach for detecting flooding attack based on Integrated Entropy Measurement in email server. Our approach can reduce the misjudge rate compared to conventional approaches. By using this approach, we can quickly analyse the current status of the email server, and determine whether the server is attacked by flooding attacks or not. Finally, according to the evolution cost value of email server by using the integrated entropy measurement we proposed, it can detect flooding attacks easily and quickly.

Acknowledgments This work was supported in part by Asia University, Taiwan, under Grant 101-asia-28, and by the National Science Council, Taiwan, Republic of China, under Grant NSC99-2221-E-468-011.

References

1. Postel JB (1982) A simple mail transfer protocol. RFC821
2. Klensin J (2008) A simple mail transfer protocol. RFC5321
3. Myers J, Rose M (1996) Post office protocol—Version 3. RFC 1939
4. Crispin M (1996) Request for comments: 2060. Standards Track, Network Working Group, Dec 1996
5. Cripsin M (2003) Internet message access protocol—version 4rev1. RFC3501
6. Karlton P (2011) Request for comments: 6101. Standards Track, Network Working Group, Aug 2011
7. Wikipedia (2013) HTTP secure. http://en.wikipedia.org/wiki/HTTP_Secure
8. Chen H-C, Sun J-Z, Wu Z-D (2010) Dynamic forensics system with intrusion tolerance based on hierarchical colour petri-nets. In: BWCCA 2010: international conference on broadband and wireless computing, communication and applications, also NGWMN-2010: the third international workshop on next generation of wireless and mobile networks, , Fukuoka, Japan, 4–6 Nov, pp 660–665
9. O'Donnell AJ (2007) The evolutionary microcosm of stock spam. Sec Priv IEEE 5:70–75
10. Bass T, Watt G (1997) A simple framework for filtering queued SMTP email. In: MILCOM 97 proceedings, vol. 3, pp 1140–1144
11. Bass T, Freyre A, Gruber D, Watt G (1998) Email bombs and countermeasure: cyber attack on availability and brand integrity. IEEE Network 12(2):10–17
12. Wang X, Chellappan S, Boyer P, Xuan D (2006) On the effectiveness of secure overlay forwarding systems under intelligent distributed DoS attacks. IEEE Trans Parallel Distrib Syst 17:619–632
13. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656
14. Absolute Astronomy (2012) Information entropy. Available from: http://www.absoluteastronomy.com/topics/Information_entropy
15. Weaver W, Shannon CE (1963) The mathematical theory of communication, 1949, republished in paperback
16. Chen H-C, Sun J-Z, Tseng S-S, Weng C-E (2012) A new approach for detecting smtpfa based on entropy measurement. In: The 9th IFIP international conference on network and parallel computing (NPC 2012), Gwangju, Korea, 6–8 Sept 2012

A Forecast Method for Network Security Situation Based on Fuzzy Markov Chain

Yicun Wang, Weijie Li and Yun Liu

Abstract In order to solve some problems associated with network security situation forecast, this study proposed a new forecast method based on fuzzy Markov chain. In this work, we focus on forecasting the threat value of network combines historical data of safe behavior with the level of threat. We establish unified information database based on multi-source log data mining techniques. By using text categorization and the threat level division, it is capable of calculating the threat value of a period of time. Due to the discrete nature of the threat values of each time and its unfollow-up effect property, considering the fuzziness of safety state, we use fuzzy Markov chain to predict the threat value in next period of time.

Keywords Network security situation · Log audit · Text mining · Fuzzy Markov chain · Forecast

Introduction

In recent years, how to grasp the network security situation and forecast the development trend has become one of the hottest spots of researches in network emergency response at home and abroad [1, 2]. Traditional network security

Y. Wang
School of Communication and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: 12120209@bjtu.edu.cn

W. Li
Key Laboratory of Communication and Information Systems, Beijing Municipal
Commission of Education, Beijing Jiaotong University, Beijing 100044, China
e-mail: liwj@itsec.gov.cn

Y. Liu (✉)
China Information Technology Security Evaluation Center, Beijing, China
e-mail: liuyun@bjtu.edu.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_108, © Springer Science+Business Media Dordrecht 2014

situation prediction algorithms are based on Naïve Bayes and grey incidence model [3], these methods can only provide network managers past and current network security situation, but cannot predict the next stage of network security status. Markov prediction model is suitable for the data prediction which meets the condition of large fluctuations, many uncertain factors with complex correlation and enough data [4]. Considering the correlation between the security states and the fuzziness of network security status, we use fuzzy Markov chain to predict the network security situation [5].

In this paper, we firstly preprocess the large amount of logs generated by network equipment, convert heterogeneous data formats, establish a unified database structure, and then use the method of text mining to classify all kinds of attacks. We will introduce it in "Log Audit". In "Network Situational Awareness", we will rank the level of threat based on past experience and evaluate the current situation of network security, so as to arrive threaten values in each period. Predication is covered in "Prediction", we use the threaten values of each period to predict the network security situation in next period based on fuzzy Markov chain.

Log Audit

Security situational awareness refers to awareness and access to the secure element in a certain time and space, integrating and analyzing the obtained data, and predicting the future trend based on the results of analysis. In recent years, this technology is gradually applied to computer network. In this paper, we use the log audit to analyze the correlation between various logs, get security event information, and calculate the theoretical security threat value, in order to predict the network security status.

Text Preprocessing

In a computer network, there are a large number of hosts, servers and network equipments, which will generate numerous logs when the system running. There is a certain correlation between these logs, which includes the security event information. However, in the process of information construction, the use of network devices becomes increasingly complicated. This generated a large number of heterogeneous, distributed security audit data, which makes it difficult for data analysis and decision supporting. The solution is preprocessing the large amount of logs by converting heterogeneous data formats and establishing a unified database structure. Because there exist considerable research efforts in this subtask, we will not detail or discuss it in this paper.

Text Mining

Security audit can be seen as a classification problem: we hope to classify each audit records to the possible categories, such as the normal or a particular intrusion or an abnormal operation. The audit log is usually based on the rule base, mathematical statistics, or data mining. In this paper, we use text mining to classify the log entries according to the type of attacks. Typical classification algorithms include Naïve Bayes, KNN and TFIDF algorithm. SVM owns the highest accuracy, but the computation time is too much; Naïve Bayes is the fastest in classifying time as well as the most efficient in terms of system memory usage although it not owns the highest accuracy [6]. As there are huge amount of logs and network security forecast has a strong real-time characteristic, we use Naïve Bayes to classify the log entry in this paper.

According to the Bayes formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$, the task of Naïve Bayes text classification is to classify the text vector $X(x_1, x_2, \ldots, x_n)$ to its most closely correlated categories $C(c_1, c_2, \ldots, c_j)$, where $X(x_1, x_2, \ldots, x_n)$ is the feature vector of X_q, and $C(c_1, c_2, \ldots, c_j)$ is the given category system. The assumption of Naïve Bayes is that any two terms from $X(x_1, x_2, \ldots, x_n)$ representing a text T and classified under category c are statistically independent of each other. This can be expressed by:

$$P(X|c) = \prod_{i=1}^{n} P(x_i|c) \tag{1}$$

The category predicted for T is based on the highest probability given by:

$$c_{NB} = \arg\max p\left(x_1, x_2, \ldots, x_n|c_j\right)P\left(c_j\right) = \arg\max P\left(c_j\right)\prod_{i=1}^{n} P(x_i|c_j) \quad c_j \in C \tag{2}$$

Network Situational Awareness

There are many types of system attacks, and different security events have different threat level. From the angle of attack level, this paper will divide all attacks into four levels [7] (Fig. 1).

Level 1 includes the denial-of-service attacks and the mail bombs, and DoS includes distributed denial of service attacks, reflective distributed denial of service attacks, the DNS distributed denial of service attack and FTP attacks. Level 2 includes such events, such as local users illegally obtain the read and write permissions. Level 3 is the problems related to that external access to internal documents. Due to the server configuration error, harmful CGI and overflow problem, related vulnerabilities will appear in large numbers. Level 4 occurs in the

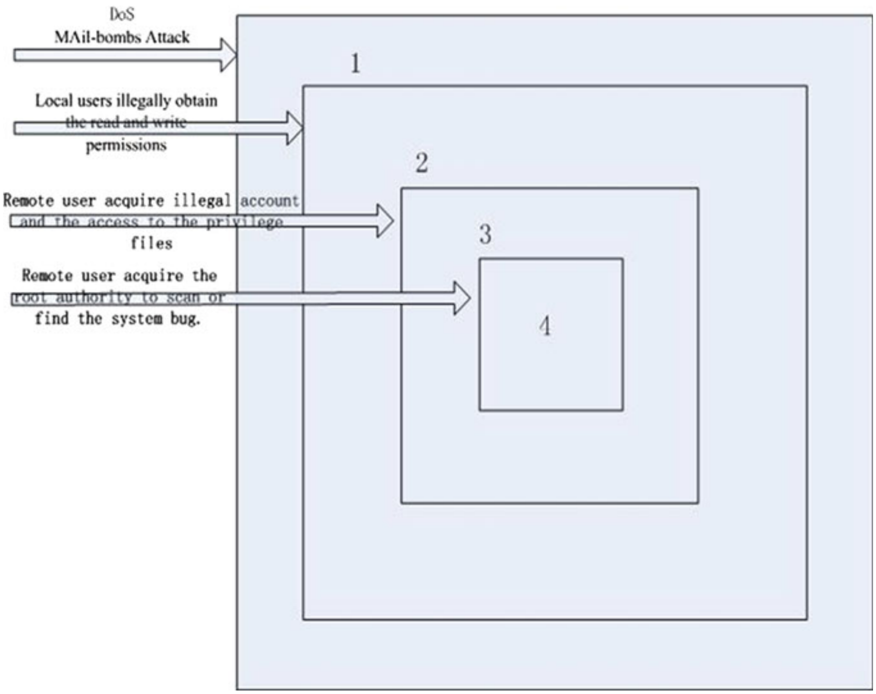

DoS
MAil-bombs Attack
Local users illegally obtain the read and write permissions
Remote user acquire illegal account and the access to the privilege files
Remote user acquire the root authority to scan or find the system bug.
1
2
3
4

Fig. 1 Attack levels

Table 1 Threaten values of the attack levels

Attack level	0	1	2	3	4
Threaten values	0	2	4	8	10

environment that what happened should never be allowed. Main attacks include TCP/IP continuous theft, the passive channel eavesdrop, packet interception, and attackers gain root privileges.

According to the four attack levels, we can divide different security events into different grades, and assign a weight to each of the attack levels (Table 1).

Where 0 stands for the security incidents, and threaten values of 0 to 10 correspond to the above attack classification. We can use the following formula to normalize the threat level:

$$p(j) = \frac{M(j) - MIN}{MAX - MIN} \tag{3}$$

Where MAX is the supremum of threat of attacks, and MIN is the infimum of the threat of attacks, M(i) is the current threat.

We denote c_{ij} as the security event which belongs to the category i and the attack level j. Such as in Naïve Bayes classification, $c_1, c_2, \ldots, c_m$ are different attacks, their attack level are all 1, then we denote them as $c_{11}, c_{12}, \ldots, c_{1m}$. Thus, the threaten value of network in a period of time can be calculated by the following formula:

$$A = \sum_{j=0}^{4} [p(j) \cdot \sum_{i=1}^{m} c_{ij}] \tag{4}$$

Thus, we get the threaten value of the network in a period of time.

Prediction

In Part 3, we combined the threaten values of different attacks with the number of attacks, and get the threat value of the network in a period of time, which reflects the network security situation. Network security state is not a specific subset, and we cannot make a clear division to the network security situation according to A-value. So we have to use the prediction method based on fuzzy Markov chain.

Several Concepts of Markov Chain

Markov chain is a mathematical system that undergoes state transitions from one state to another, between a finite or countable number of possible states. It is a random process characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it. This specific kind of "memorylessness" is called the Markov property.

The changes of state of the system are called transitions, and the probabilities associated with various state-changes are called transition probabilities: $p_{ij} = P(X_{n+1} = j | X_n = i)$. The set of all states and the transition probabilities completely characterizes a Markov chain. A Markov chain is a sequence of random variables $x_1, x_2, x_3 \ldots$ with the Markov property, namely that, given the present state, the future and past states are independent.

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n) \tag{5}$$

The possible values of X_i form a countable set S called the state space of the chain. And transition Matrix can be stated as:

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0j} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1j} & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \cdots \\ p_{i0} & p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \cdots \end{bmatrix}$$

Fuzzy Markov Chain Model

1. We set random variable X_i as the threaten value A of a period of time of the network security situation, the range is set to U. Set up fuzzy state $\tilde{N}_1, \tilde{N}_2, \ldots, \tilde{N}_r$ on U, in the condition of $\forall x \in U, \sum_{i=1}^{r} \tilde{N}_i(x) = 1$.
2. In period K to period K + 1, the fuzzy state transition frequency number of state $\tilde{N}_i$ transfer to state $\tilde{N}_j$ can be expressed as:

$$\tilde{S}_{ij} = \bigvee_{n-1}^{k=1} \left[\tilde{N}_i(x_k) \wedge \tilde{N}_j(x_{k+1})\right] \tag{6}$$

Fuzzy state transition frequency can be calculated by:

$$\tilde{P}_{ij} = \tilde{S}_{ij} \Big/ \sum\nolimits_{j=1}^{r} \tilde{S}_{ij}, \; i = 1, 2, \ldots, r \tag{7}$$

Experimental Validation and Analysis

The experimental data set is KDD99 [8], which includes numerous varied invasions simulated in a military network environment and have more than 4,000,000 event items. Each event item contains the feature value extracted from the original network data collected during the simulated invasion stage. We use the data set as the text set which has been classified during the log preprocessing process. We take the top 20 % of KDD99 as our training set and the 1 % of the followed data of KDD99 as our test set. Attacks and their categories are shown in Table 2:

1. DoS(Denial of Service), useslarge amount of legal requests to take up too much service resource.
2. U2R(User to Root), refers to the unauthorized access attacks that local non-privileged user instead of local super-user launched.

Table 2 Attacks and their categories

Category	Attack
DoS	Back, land, neptune, pod, smurf, teardrop
U2R	Buffer_overflow, loadmodule, perl
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy,
PROBE	warezclient, warzmaster, lpsweep, nmap, portsweep

Table 3 Threaten value of each period of time

Period	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}
Threat Value	1,926	6,859	7,745	8,930	10,000	10,000	7,968	9,270	9,865	7,428
Period	τ_{11}	τ_{12}	τ_{13}	τ_{14}	τ_{15}	τ_{16}	τ_{17}	τ_{18}	τ_{19}	τ_{20}
Threat Value	5,621	4,215	6,098	9,707	13,359	14,695	17,520	11,593	8,596	5,207

3. R2L(Remote to Local) refers to the unauthorized access of the remote host.
4. Probing, means that the attackers scan computers on the network to collect information or to find the known system vulnerabilities.

We assume that the records are evenly distributed in time τ, and the total time of collecting training data set. We divide the total time into 20 equal sections which are stated by $\tau_1,\ \tau_2,\ldots,\tau_{20}$ respectively, and calculate the network threats of each section, classify threats according to the security level stated in part 3 and compute the threaten value of each period of time. The computed results are shown in the Table 3.

According to the needs of different systems, fuzzy state level setting and membership degree equations are different. The most suitable membership degree equation can be derived from training. In order to illustrate our forecast method, in this paper, we set up fuzzy states on the set of real numbers R as follows, where x is the network threaten value of each period of time.

$$\tilde{N}_1 = \begin{cases} 1 & \mathrm{x} \in [0, 1000) \\ \frac{3000-\mathrm{X}}{2000} & \mathrm{x} \in [1000, 3000) \\ 0 & \text{else} \end{cases} \qquad \tilde{N}_3 = \begin{cases} \frac{\mathrm{X}-6000}{2000} & x \in [6000, 8000) \\ 1 & x \in [8000,11000) \\ \frac{13000-\mathrm{X}}{2000} & x \in [11000, 13000) \\ 0 & \text{else} \end{cases}$$

$$\tilde{N}_2 = \begin{cases} \frac{\mathrm{X}-1000}{2000} & x \in [1000, 3000) \\ 1 & x \in [3000,6000) \\ \frac{8000-\mathrm{X}}{2000} & x \in [6000, 8000) \\ 0 & \text{else} \end{cases} \qquad \tilde{N}_4 = \begin{cases} \frac{\mathrm{X}-11000}{2000} & x \in [11000, 13000) \\ 1 & x \in [13000,16000) \\ \frac{18000-\mathrm{X}}{2000} & x \in [16000, 18000) \\ 0 & \text{else} \end{cases}$$

$$\tilde{N}_5 = \begin{cases} \frac{\mathrm{X}-16000}{2000} & \mathrm{x} \in [16000, 18000) \\ 1 & \mathrm{x} \in [18000, 50000) \end{cases}$$

Easy to calculate the threat of fuzzy state membership in each period as follows (Table 4):

Based on formula (6) and (7), it is easy to get Tables 5 and 6:

Thus, we can get the fuzzy state transition Markov chain:

Table 4 Fuzzy state membership in each period

	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}
$\tilde{N}_1$	0.537	0	0	0	0	0	0	0	0	0
$\tilde{N}_2$	0.463	0.5705	0.1275	0	0	0	0.016	0	0	0.286
$\tilde{N}_3$	0	0.4295	0.8725	1	1	1	0.984	1	1	0.714
$\tilde{N}_4$	0	0	0	0	0	0	0	0	0	0
$\tilde{N}_5$	0	0	0	0	0	0	0	0	0	0
	τ_{11}	τ_{12}	τ_{13}	τ_{14}	τ_{15}	τ_{16}	τ_{17}	τ_{18}	τ_{19}	τ_{20}
$\tilde{N}_1$	0	0	0	0	0	0	0	0	0	0
$\tilde{N}_2$	1	1	0.951	0	0	0	0	0	0	1
$\tilde{N}_3$	0	0	0.049	1	0	0	0	0.7035	1	0
$\tilde{N}_4$	0	0	0	0	1	1	0.24	0.2965	0	0
$\tilde{N}_5$	0	0	0	0	0	0	0.76	0	0	0

Table 5 Fuzzy state transition frequency number

$\tilde{s}_{11} = 0$	$\tilde{s}_{12} = 0.537$	$\tilde{s}_{13} = 0.4295$	$\tilde{s}_{14} = 0$	$\tilde{s}_{15} = 0$
$\tilde{s}_{21} = 0$	$\tilde{s}_{22} = 0.714$	$\tilde{s}_{23} = 0.951$	$\tilde{s}_{24} = 0$	$\tilde{s}_{25} = 0$
$\tilde{s}_{31} = 0$	$\tilde{s}_{32} = 1$	$\tilde{s}_{33} = 1$	$\tilde{s}_{34} = 1$	$\tilde{s}_{35} = 0$
$\tilde{s}_{41} = 0$	$\tilde{s}_{42} = 0$	$\tilde{s}_{43} = 0.24$	$\tilde{s}_{44} = 1$	$\tilde{s}_{45} = 0.76$
$\tilde{s}_{51} = 0$	$\tilde{s}_{52} = 0$	$\tilde{s}_{53} = 0.7035$	$\tilde{s}_{54} = 0.2965$	$\tilde{s}_{55} = 0$

Table 6 Fuzzy state transition frequency

$\tilde{p}_{11} = 0$	$\tilde{p}_{12} = 0.5556$	$\tilde{p}_{13} = 0.4444$	$\tilde{p}_{14} = 0$	$\tilde{p}_{15} = 0$
$\tilde{p}_{21} = 0$	$\tilde{p}_{22} = 0.4183$	$\tilde{p}_{23} = 0.5817$	$\tilde{p}_{24} = 0$	$\tilde{p}_{25} = 0$
$\tilde{p}_{31} = 0.1252$	$\tilde{p}_{32} = 0.2916$	$\tilde{p}_{33} = 0.2916$	$\tilde{p}_{34} = 0.2916$	$\tilde{p}_{35} = 0$
$\tilde{p}_{41} = 0$	$\tilde{p}_{42} = 0$	$\tilde{p}_{43} = 0.12$	$\tilde{p}_{44} = 0.5$	$\tilde{p}_{45} = 0.38$
$\tilde{p}_{51} = 0$	$\tilde{p}_{52} = 0$	$\tilde{p}_{53} = 0.7035$	$\tilde{p}_{54} = 0.2965$	$\tilde{p}_{55} = 0$

$$\tilde{P} = \begin{pmatrix} 0 & 0.5556 & 0.4444 & 0 & 0 \\ 0 & 0.4183 & 0.5817 & 0 & 0 \\ 0.1252 & 0.2916 & 0.2916 & 0.2916 & 0 \\ 0 & 0 & 0.12 & 0.5 & 0.38 \\ 0 & 0 & 0.7035 & 0.2965 & 0 \end{pmatrix}$$

The threat value of the twentieth section is 5,207 and the subordination allocation vector of its fuzzy state is (0, 1, 0, 0, 0), so we can presume the probabilities of the network threat value of the next time section belong to the five states respectively according to the transition probability matrix $\tilde{P}$, as shown in the formula:

$$(0, 1, 0, 0, 0)\tilde{P} = (0, 0.4183, 0.5817, 0, 0) \tag{8}$$

According to the maximum subordination principle, it is easily to be seen that the threat value of the next time section most possibly belongs to the fuzzy state $\tilde{N}_3$.

In the test set, the actual value of the next period is 7041 and the security situation of the experiment value and the actual value of the next period both belong to the same fuzzy state $\tilde{N}_3$.

Conclusions and Future Work

We have introduced a method for forecasting the network security situation based on fuzzy Markov chain. We firstly preprocessed the large amount of logs and established a unified database structure, and used the text mining to classify all kinds of attacks. Based on the past experience, we assigned to each attack a different weight according to their danger degrees. Through the calculation, we obtained the Markov state transition matrix, and realized the prediction of the network security situation. Because the attacks are correlative, this method showed a good predictive performance.

Different systems have different requirements, and the fuzzy membership function is different from each other. We can always train the best membership function according to the training set. Another interesting aspect or further study is how to deal with the unknown threats. Our current approach is based on the classification of known threats. However, it would be desirable to explore other more sophisticated approaches.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Onwubiko C (2009) Functional requirements of situational awareness in computer network security. In: IEEE international conference on on intelligence and security informatics ISI'09, pp 209–213
2. Wei-wei X, Hai-feng W (2010) Prediction model of network security situation based on regression analysis. In: IEEE international conference on wireless communications, networking and information security (WCNIS 2010), pp 616–619
3. Jibao L, Huiqiang W, Liang Z (2006) Study of network security situation awareness model based on simple additive weight and grey theory. in: International conference on computational intelligence and security, pp 1545–1548
4. Zhang Y, Tan XB, Cui XL et al (2011) Network security situation awareness approach based on Markov game model. J Software 22(3):495–508

5. Kuang GC, Wang XF, Yin LR (2012) A fuzzy forecast method for network security situation based on Markov. In: International conference on computer science and information processing (CSIP 2012), pp 785–789
6. García Adeva JJ, Pikatza Atxa JM (2007) Intrusion detection in web applications using text mining. Eng Appl Art Intell 20(4):555–566
7. http://www.ibm.com/developerworks/cn/security/se-parseatt/
8. http://www.sigkdd.org/kdd-cup-1999-computer-network-intrusion-detection

IT Architecture of Multiple Heterogeneous Data

Yun Liu, Qi Wang and Haiqiang Chen

Abstract This paper studies the multiple heterogeneous data IT infrastructure issues in Internet, focusing on the data storage management, the integration analysis and the efficient data transmission. On this basis, we propose architectural solutions and key technology programs which can effectively solve and handle the problems of large data trend analysis and risk assessment. Based on the characteristics of big data, considering the flexibility and scalability of the framework and taking in-depth mining the potential value of big data as the future goal, this paper studies and designs data storage and processing framework which can be adapted to developing more application functionality and meet future demand for Internet services and new business for big data processing requirements.

Keywords Big data · Data mining · Multiple heterogeneous data · Data warehouse

Y. Liu (✉) · Q. Wang
School of Communication and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: liuyun@bjtu.edu.cn

Q. Wang
e-mail: 12125042@bjtu.edu.cn

Y. Liu · Q. Wang
Key Laboratory of Communication and Information Systems, Beijing Municipal
Commission of Education, Beijing Jiaotong University, Beijing 100044, China

H. Chen
China Information Technology Security Evaluation Center, Beijing, China
e-mail: chenhq@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_109, © Springer Science+Business Media Dordrecht 2014

Introduction

The rapid growth of the number of data, to a large extent, is promoted by the revolutionary network applications, such as micro-blog, e-commerce, messaging, social networking and location-related information search and aggregation. Such kind of big data not only has mass and high-speed characteristics, but also the diversity and variability. The diversity also known as heterogeneity, is considered to be generated for the widely using of Internet search and another reason is the new multi- structured data, including the data types of blogs, social media, Internet, phone call records and sensor networks. Especially sensors, each sensor increases the diversity of the data, due to the different installation locations and functions. Variability of big data is reflected in the multi-layer structure and varied forms of the data.

The larger scale data, the processing and storage is more difficult, but its mining potential value is greater. Large data contains the value, but there are still many difficulties and challenges in the use of big data technology. Due to the potential value of big data and the features of specific business and data source, this paper studied and proposed a multiple heterogeneous data IT infrastructure which is appropriate to the specific business needs. We focus on studying the key technologies such as the data storage management, the integration of analysis and the efficient data transmission, and present technical architecture level solutions and key technical solutions based on the data trend analysis and the risk assessment. Based on the characteristics of big data, considering the flexibility and scalability of the framework and taking in-depth mining the potential value of big data as the future goal, this paper studies and designs data storage and processing framework which can be adapted to developing more application functionality and meet future demand for Internet services and new business for big data processing requirements.

Related Work

Google launched its own distributed file system GFS [1] in 2003. GFS is a scalable distributed file system. It is a dedicated file system designed to store massive amounts of search data. GFS is used for large-scale, distributed, large amounts of data access applications and runs on inexpensive commodity hardware, but can provide fault tolerance, and higher overall performance of the services provided to a large number of users. Dean et.al. [2], proposed a parallel computing software framework for large data sets (more than 1 TB)—MapReduce. Chang et.al. [3] proposed a framework Bigtable. Bigtable is a distributed structured data storage system, which is designed to handle massive amounts of data: Usually the PB level data which is distributed in the thousands of ordinary servers. Current most of mainstream data processing frameworks are customized and improved according to Hadoop [4, 5] which is developed based on the early Google's GFS/MapReduce core idea. These customizations and improvements some focus on improving system reliability and manageability, and some focus on improving the processing speed of the system and reducing the delay, while others focus on considering the data processing capabilities of the system, as well as the specific hardware or service—specific optimization. The difficulties faced by a big data processing focus on three aspects of large-capacity, heterogeneity and high-speed, and thus the core technologies focus to address these issues. In addition, big data usually has the characteristic of multi-source diversity, the effective integration of multivariate data is the key link of fully mining the data value.

Contributions

In this paper, we study the technologies of large data storage, integration of heterogeneous data, real-time computing architecture and efficient data transmission mode based on the characteristics of big data and the existing big data technologies. Then we presented key technologies and key considerations of big data system and proposed referenced big data system architecture.

Multiple Heterogeneous Data IT Infrastructure

Multiple heterogeneous big data IT infrastructure mainly consists of four parts: the heterogeneous big data tiered storage and storage management, the integration of heterogeneous data, the off-line and real-time computing architecture, the efficient transmission mode of big data. Typical big data frameworks are based on MapReduce and Hadoop and combined with the specific services to be designed. In this paper, we designed a complete technical architecture according to the multiple heterogeneous data processing and storage requirements of the upper services. This technical structure consists of the underlying file system, structured data storage system, semi- structured data and unstructured data storage system, the distributed storage of data, the unified data access interface, data indexing and positioning, processing task decomposition and scheduling management, e distributed execution of processing tasks, e service interface of the system and the secondary development interface and system manageability and security and completely defined a data storage and processing platform meeting the specific needs. In the following paper, we will present the four aspects of our big data framework respectively.

Multiple Heterogeneous Data Tiered Storage and Storage Management

Multiple heterogeneous big data have the characteristics of multiple dimensions, multiple sources, structured + semi-structured + unstructured heterogeneous data and the huge amount of data, thus the traditional relational database has been very difficult to store such data. In addition, the demand for the processing of such data is not limited to the mode supported by SQL, but shows the features of the connection and interweaving between the data on the Internet. There are a lot of storage methods for heterogeneous data, such as serial method, graph method, tree methods, file system method, database field methods and object methods. Due to the difference between the services of application layer, the characteristics of multi-source heterogeneous data also are different with different services, so our data storage and management must use corresponding methods for differing services, for example, Bigtable, the Keyvalue of NoSQL and so on. We use the Multiple heterogeneous data tiered storage as our storage method.

Big data needs to mass storage, so we cannot storage and manage them like the traditional solutions treating them as equally important. However, Hierarchical storage is using different storage methods corresponding to the characteristics and values of different data and utilize the hierarchical storage management software to achieve automatic sorting, automatic storage, which greatly increases the effectiveness and speed of data storage and meets the storage needs of different kinds of data. In our system, there are four kinds of data which are static data, dynamic data, offline data and online data respectively. The framework of hierarchical storage is shown in Fig. 1.

In order to achieve the storage of big data, especially the Heterogeneous data storage of the structured data, semi-structured data and unstructured data, we superimpose a relational database and a distributed non-relational database on a distributed file system to store huge amounts of data. In the file system level, data are managed by the master + multi-slave physical structure. The master node is used to store metadata, and slave nodes are used to store the actual data. Logical data level is also constituted by the same master + multi-slave structure, where the master node stores metadata, making the access to RDBMS and NoSQL is consistent and also playing the role of slave control. Slave nodes in the logical data level are used to store logical data. Distributed file system is responsible for the management and maintenance of the persistence and positioning of the data on the storage medium and provides read and write physical media I/O for the logical data layer. The logical data layer is responsible for the management of the logical structure of the data and provides services of access to the logical data for the high level. Figure 2 shows us the storage framework of big data.

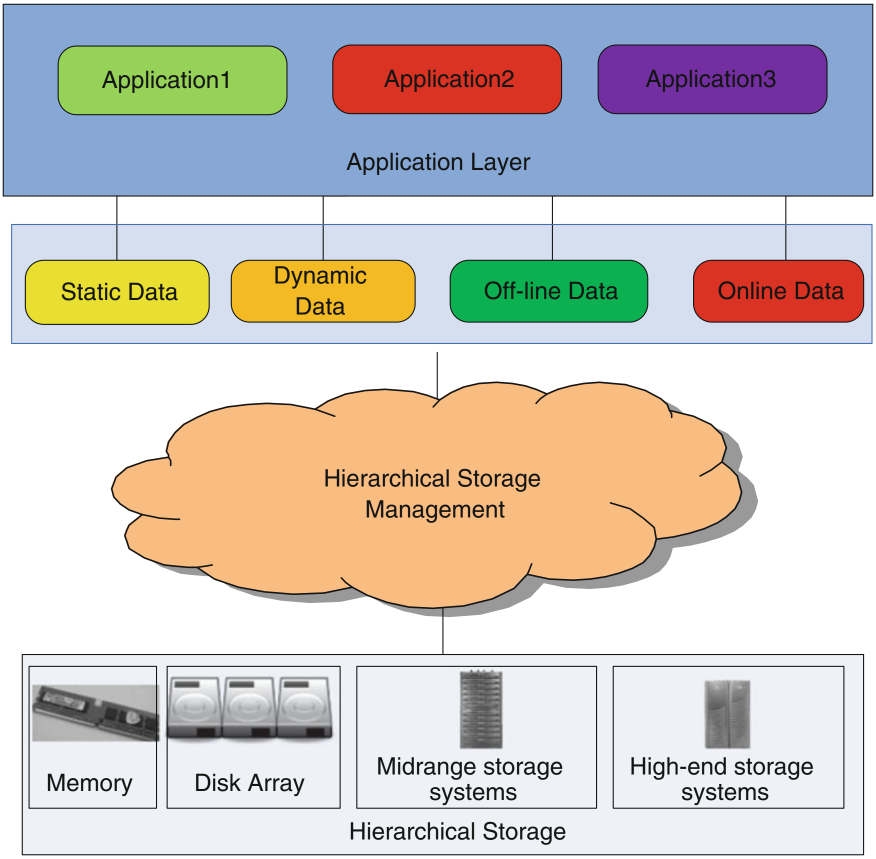


Fig. 1 Hierarchical storage diagram

Heterogeneous Data Fusion

Heterogeneous data fusion is the core part of the deal with heterogeneous data. Fusion of the data can be divided into three levels of pixel-based fusion, feature-level fusion and decision fusion. The pixel-based fusion also has a function similar to the data de-noising and the cleaning process in addition to the significance of fusion. Feature-level fusion classifies data based on the features and data attributes. Decision Fusion uses data to implement the trend assessment and the macro perspective service processing at the highest level. According to the specific services and the specific characteristics of the data, the corresponding fusion algorithms and computational structures are used to treat on different data. Utilizing data fusion technology, we can remove redundant information to reduce the amount of data increase the access efficiency. Moreover, we can extract useful

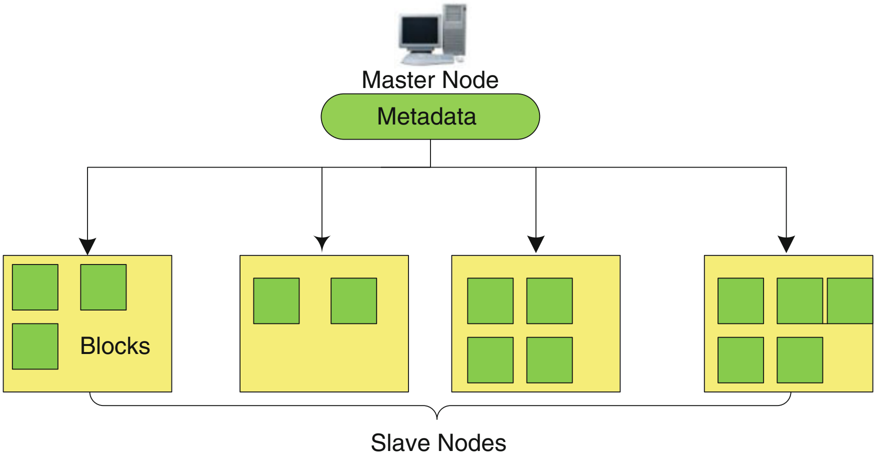
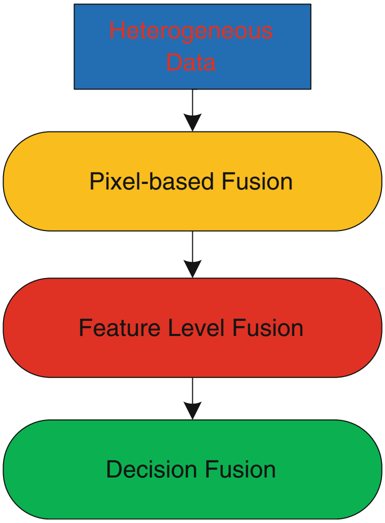


Fig. 2 Distributed file system structure

information from a large number of heterogeneous data, providing services and support for the business of the upper services. Figure 3) shows the functional model of heterogeneous data fusion.

Fig. 3 Heterogeneous data fusion diagram



Off-line and Online Computation Framework

There are off-line and online computing demands in a large data environment. Off-line computation is more relaxed in the computing time requirement. However, under the conditions of mass data, there are still computing time limit and the balance between recalculation and the incremental calculation becomes an inevitable problem with the increasing of data. Online calculation has a higher requirement on the computing time, which is a quit hard problem for the mass data and needs to a variety of methods to guarantee the calculation real-time.

In this paper, we class the data into static data and dynamic data. Static data are the historical read-only data, but dynamic data are the read and write data including intermediate result. For static data, when we implement off-line computation, we must design a reasonable storage structure and create effective index to improve the processing efficiency according to the demands of specific services. Hadoop use the idea of Mapreduce. It slices the data to deal with large amount of off-line data and then assign computing tasks to more than one computer, which will greatly improve the speed and efficiency of computation. Online computation needs to use reasonable caching mechanism to solve the massive data processing problem. For dynamic data, we will combine data structure designing, index designing and caching mechanism designing together to complete the data processing. Storm is a common stream computing framework which has been widely used for real-time log processing, real-time statistics, real-time risk control. Storm also is used to process the data preliminarily and it can store the data in a distributed database like HBase in order to facilitate the subsequent queries. Storm is a scalable, low-latency, reliable and fault-tolerant distributed computing platform. In addition, we will employ the distributed computing mode for data computation and processing and the assignment, scheming and management of the computing tasks. The off-line and online computation framework is shown in Fig. 4.

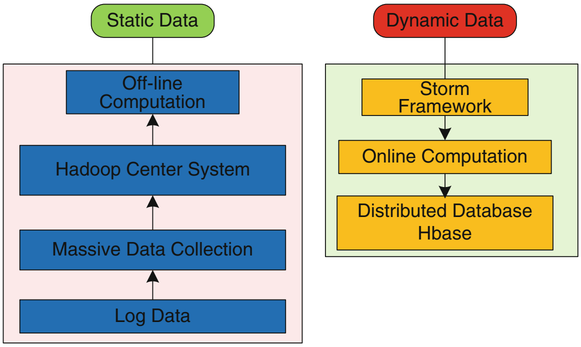


Fig. 4 Off-line and online computation

Efficient Interactive Transmission Mode of Big Data

Storage and processing of big data necessarily involves the network-based distributed storage and computation, the collection of multi-source data, the remote data management on the basis of network and data sharing and exchange. So Data exchange and interactive data service is an indispensable supporting technology.

In this paper, we designed the data exchange framework according to two modes of data service and heterogeneous data exchange. Data service is established data access function for large-scale predefined data transferring and migration and heterogeneous data exchange refers to the data exchange between the different modules within a framework and clients. Data service mode uses connection-oriented channel exchange. However, Heterogeneous Data uses connectionless datagram exchange and defines the structure of datagram by xml or json.

The framework defines three types of service interfaces for the data interactive interface: the interfaces between the modules within the system, which include two modes of service and API; the external interfaces of the system functions, which appear in the form of API and are used for the developing and data access of the service applications; interactive service interfaces, which have two modes of service and REST API. System interfaces can support each other to form the overall function. The external interfaces in the form of API can be used for the developing the service applications and REST API can be used for extending the service applications of the system and create a more convenient remote data access. Data exchange modes are shown in Fig. 5.

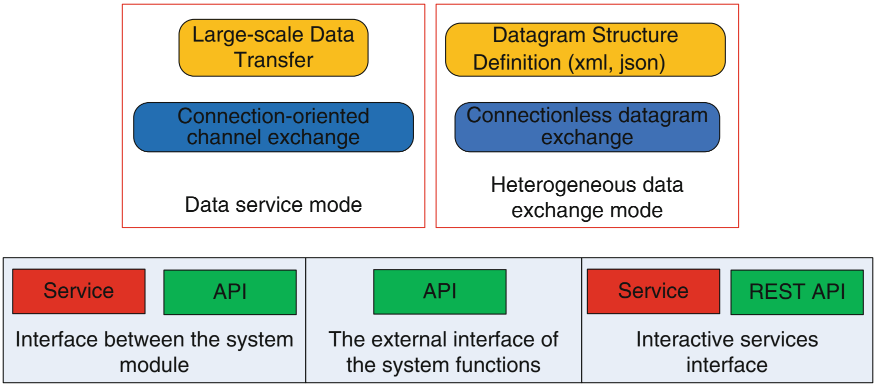

Large-scale Data Transfer
Connection-oriented channel exchange
Data service mode
Datagram Structure Definition (xml, json)
Connectionless datagram exchange
Heterogeneous data exchange mode
Service
API
Interface between the system module
API
The external interface of the system functions
Service
REST API
Interactive services interface

Fig. 5 Data exchange module structure

Conclusion

In this paper we talked about the framework of multiple heterogeneous big data from four aspects of the storage management, the data fusion, the online and offline computation and the efficient interactive transmission mode of big data. We then proposed a framework of multiple heterogeneous data which can be adapted to developing more application functionality and meet future demand for Internet services and new business for big data processing requirements.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. Ghemawat S, Gobioff H, Leung ST (2003) The google file system. ACM SIGOPS Oper Syst Rev 37(5):29–43
2. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113
3. Chang F, Dean J, Ghemawat S et al (2008) Bigtable: a distributed storage system for structured data. ACM Trans Comput Syst (TOCS) 26(2):4
4. Hadoop: the definitive guide. O'Reilly Media, California
5. White T (2012) Hadoop: the definitive guide. O'Reilly Media, California

An Implementation Scheme of BB84-Protocol-Based Quantum Key Distribution System

Shao-sheng Jiang, Rui-nan Chi, Xiao-jun Wen and Junbin Fang

Abstract The BB84 protocol is a secure communication solution that has been proved to be unconditionally secure. The quantum key distribution system that is based on this protocol has been gradually developed towards the application stage. This paper proposes a scheme to realize quantum key distribution system based on BB84 protocol using weak coherent light source, polarization controller and delay multiplexing detection technology to prepare single photon, to design photon polarization states orthogonal to each other, and to detect single photon respectively.

Keywords Quantum key distribution · Single photon signal · Photon polarization states · Delay multiplexing detection

Introduction

Quantum key distribution system is formed through the operation of quantum bits on the basis of the inherent properties of such quantum bits. Because the quantum bits comply with the no-cloning theorem and the uncertainty principle, the quantum key distribution system theoretically has the capability of unconditional security protection, and thus the technology and scheme for its realization are considered as an important direction in the research of present information security.

S. Jiang · R. Chi · X. Wen (✉)
School of Computer Engineering, ShenZhen POLYTECHNIC, Shenzhen 518055, China
e-mail: szwxjun@sina.com

J. Fang (✉)
Department of Optoelectronic Engineering, Jinan University, Guangzhou 510632, China
e-mail: junbinfang@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_110, © Springer Science+Business Media Dordrecht 2014

BB84 protocol is the first quantum key distribution protocol proposed by Bennett and Brassard. It is a four-state scheme based on two conjugate bases, which uses complementarity of linearly polarized photon state and circularly polarized photon state to satisfy the uncertainty principle in single photon quantum channel so as to ensure the unconditional security of quantum channel, and then complete the distribution of quantum key through the quantum transmission channel, quantum measurement channel and classic auxiliary channel [1]. Thus, the realization of quantum key distribution system mainly depends on the relevant technology design in the preparation, transformation and detection of quantum bits.

At present, the existing literatures revealed the research on the security of BB84 protocol of quantum key distribution system [2]. For system realization, Capraro et al. used Java to simulate the BB84 system [3], and Chen et al. used the method of polarization feedback compensation to achieve the BB84 protocol polarization encoding key distribution experiments [4]. This paper mainly studies the realization scheme of BB84-protocol-based quantum key distribution system; that is, by fully taking physical features on which the key system is based into account in the design, the key distribution system is ultimately achieved as a basis for the completion of all kinds of application based on such a system.

Generation of Quantum Signal

Realization of quantum key is based on the physical laws of quantum bits. According to quantum bits attribute, classical signals obviously do not have conditions to carry the quantum bits, so only quantum signals are used to complete [1] The quantum signal is the evolution of quantum system with a particular parameter, and can be described by Schrodinger equations as follows:

$$\mathrm{p} = \frac{\mathrm{h}}{\gimel} \tag{1}$$

General, considering the quantum system's characteristics of physics value A satisfies the following relations:

$$\mathrm{A} \sim \mathrm{h} \tag{2}$$

Quantum efficiency cannot be ignored, and then A can be used as quantum signal. Where h is the normalized Planck constant. If A $\gg$ h, A is described as the classical signal. Micro-systems all have the obvious quantum effects, so all these systems can be used as a carrier of quantum signals.

Because laser signal is relatively excellent in fiber transmission, a single photon is inseparable, so a single photon is an ideal carrier of quantum signals. The evolution of photons can be described by a state vector |Ψ>, so |Ψ> can be used to describe the quantum signals that are carried, referred to as the single photon sign.

But the preparation of single photon signal in the technical realization is difficult, so considering realizing single photon signal with weak laser pulse to be emitted by common lasers. The average photon number in a single laser pulse shall satisfy the following conditions:

$$\overline{N} \leq 0.1 \tag{3}$$

In the above formula $\overline{N}$ represents the average photon number in a single laser pulse. Calculation shows that, the laser pulse signal in this case has obvious quantum effect, with features approximated to the real single photon quantum signal.

The specific realization method is: design a quantum signal transmitter that can randomly emit four pulse lasers at the frequency of 1 MHz, and the center wave length is 850 nm, so each photon energy is

$$\mathrm{E_p} = \frac{\mathrm{hc}}{\lambda} = 2.337 \times 10^{-19}\mathrm{W} \tag{4}$$

The number of photons in each pulse is

$$\mathrm{n} = \frac{\mathrm{E}}{\mathrm{E_p}} = 100 \times 10^{12} / \left(2.337 \times 10^{-19}\right) = 4.279 \times 10^{8} \tag{5}$$

In order to ensure the photon number per pulse satisfying $\mathrm{N} \leq 0.1$, attenuation is required. When using an attenuator with multiple attenuation of 96 dB, the average photon number per pulse becomes

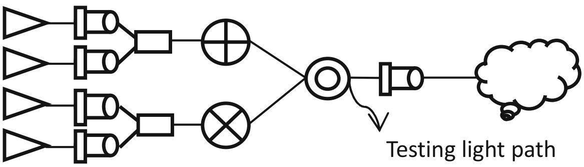
$$\overline{N} = \mathrm{n} \times 10^{-10} = 0.043 < 0.1 \tag{6}$$

Therefore, the single photons signals are generated.

Transformation of Quantum Signals

According to the protocol of BB84, four single photon signals produced by quantum signal transmitting mechanism must be transformed into two orthogonal polarization states of photons, including one pair of linear polarized photon and another pair for circularly polarized photons, which meets the uncertainty principle [5]. However, concerning that the preparation and transformation of circularly polarized photons in implementation is difficult, we design two pairs of linearly polarized photons, as long as it meets two pairs of photon state that are not mutually orthogonal and undistinguishable. The designed diagram is shown in Fig. 1:

Fig. 1 Schematic diagram of the quantum signal transform



In the above diagram:

▷ is light path emitted by the laser machine;

□ represents Polarization Beam Splitter;

⊐ represents Adjustable attenuator;

⊕ ⊗ represents Manual Polarization Controller;

◎ represents Single Mode Coupler;

☁ represents Fibernet;

First of all, we set two Polarization Beam Splitters (hereinafter referred to as PBS), which will be used to provide two photonic inputs for beam, with each polarization direction perpendicular to each other, thus forming the two orthogonal polarization states of photons.

Then we design two Manual Polarization Controllers (hereinafter referred to as MPC), which will be used to design two photon states mutually non-orthogonal. Here only let one pair of photons go through MPC rotation at the angle of 45° and then transfer, and another pair of the photons is transferred with the original mutual vertical states. Two MPC can also make two pairs of photons by adjusting the polarization feedback to well compensate the light's polarization changes in the path.

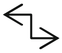
Finally, we use the Single Mode Coupler (hereinafter referred to as SMC) to form four kinds of polarized light into a beam, as a sender of quantum key distribution system.

Another important thing is to form a single photon pulse in the implementation. As mentioned before, the attenuation of polarized light into 96 dB is required. The direct attenuation with large coefficient is neither safe nor conducive to regulate the system, so generally we use 2 $\sim$ 3 concatenated attenuators for attenuation, and an optical power meter for step-by-step measurement, until it finally reaches the total attenuation of 96 dB. Usually after the emissions of four paths of light, the first attenuation is about 50 dB, and afterwards the attenuation is about 46 dB, and the error can be allowed within the range of 1 dB.

The testing light path in the diagram is used to help calibrate the polarization direction of MPC. The two properly-set MPCs can make the beam of two mutually perpendicular polarized light being prepared into |0°>, |90°>, |+45°>, |−45°> four kinds of polarized light (later we referred to these four kinds of light condition for ↔, ↕, ↗, ↖) and complement each other, so as to meet the requirements of the implementation of BB84 protocol.

Detection of Quantum Bits

The receiver of quantum key must be a photoelectric detection system, which can detect very weak quantum signals, and identify the quantum key carried by quantum signals. In the system only a single photon detector is set, by using avalanche effect to probe photons. Electric pulse signal output is processed by the main control plate, and after processing the information is transferred to a computer, with the sender completing basis vectors alignment by classical network. The realization diagram of the receiver is designed as Fig. 2:

Above  indicates a 2 m-long jumper, we use the jumper with multiple SMCs to make four polarization-state photon line to form a beam into a single photon detector, then use delay multiplexing technology to separately detect photons of various polarization states. As shown in Fig. 2, through the optical fiber transmission, a single photon will randomly select a path in the first SMC transmission. (For example, the photon of ↔ photon polarization state had a 50 % probability of going to the wrong path, and by choosing ⊗ state MPC, it is equivalent to the receiver choosing a wrong measurement basis vector, this part of detection signals will be discarded directly. When the photons of polarization states select correct measurement basis vectors, the MPC is mainly used for polarization mode dispersion for feedback compensating polarized light in the system, which makes it consistent with the polarization beam direction.)

In Fig. 2, the main function of PBS is to split polarized light of the two mutually perpendicular states into beams. For the detection of ↔, ↕, ↗, ↖ four states of polarized light, the optical receiver increases 0, 2, 4, 6 m jumpers

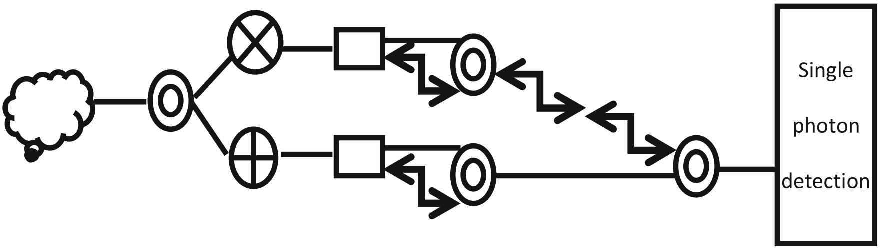


Fig. 2 Schematic diagram of the realization

respectively. If adopting the conversion of 1 m to 5 ns, i.e., delay of X + 0 ns for detecting ↔ photons, delay of X + 10 ns for detecting ↕ photons, delay of X + 20 ns for detecting ↗ photons, and delay of X + 30 ns for detecting ↖ photon. The X value is the value acquired by the laser pulse detector in roughly scanning at the delays.

Conclusion

The BB84 protocol is the first quantum cryptography communication agreement on the basis of single photon of a four-state scheme based on two conjugate bases. This paper proposes an implementation technology and realization circuit schematic diagram based on the protocol, with the key components including the preparation of single photon signals, the design of two pairs of orthogonal polarization of photons, the delay multiplexing detection technology of single photon, which has been realized in the laboratory. It has been proved to be unconditionally secure [5, 6]. At present, many applications have been realized on the basis of the key distribution system, which has the very good application and practical value.

References

1. Zeng GH (2006) Quantum cryptography. Science Press, Beijing
2. Chen ZX, Tang ZL, Liao CJ et al (2006) Analysis of BB84 protocol security eavesdropping extended practical quantum key distribution. Acta Photonica Sinica 35:126–129
3. Ivan C, Tommaso O (2007) Implementation of real time high level protocol software for quantum key distribution. In: IEEE international conference on signal processing and communications, Dubai, United Arab Emirates, pp 24–27
4. Chen J, Li Y, Wu G et al (2007) Quantum key distribution in polarization stability control. J Phys 56:5243–5247
5. Wen XJ, Chen YZ (2012) Quantum signature and its application. Aviation Industry Press, Beijing
6. Liu JF, Liang RS, Tang ZL, et al (2004) Eavesdropping of practical QKD system based on BB84 protocol. Acta Photonica Sinica 33:1356–1359

Precise Abdominal Aortic Aneurysm Tracking and Segmentation

Shwu-Huey Yen, Hung-Zhi Wang and Hao-Yu Yeh

Abstract In this paper we propose a mean-shift based technique for a precise tracking and segmentation of abdominal aortic aneurysm (AAA) from computed tomography (CT) angiography images. Output data from the proposed method can be used for measurement of aortic shape and dimensions. Knowledge of aortic shape and size is very important for selection of appropriate stent graft device for treatment of AAA. Comparing to conventional approaches, our method is very efficient and can save a lot of manual labors.

Keywords Aneurysm · Abdominal aortic aneurysm · Mean-shift · Computerized tomorgraphy

S.-H. Yen (✉) · H.-Z. Wang · H.-Y. Yeh
Department of Computer Science and Information Engineering, Tamkang University, New Taipei, Taiwan, Republic of China
e-mail: 105390@mail.tku.edu.tw

H.-Z. Wang
e-mail: 600410640@s00.tku.edu.tw

H.-Y. Yeh
e-mail: 601411191@s01.tku.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_111, © Springer Science+Business Media Dordrecht 2014

Part X
Advances in Multimedia Algorithms, Architectures, and Applications

Mining High Utility Itemsets Based on Transaction Deletion

Chun-Wei Lin, Guo-Cheng Lan, Tzung-Pei Hong and Linping Kong

Abstract In the past, an incremental algorithm for mining high utility itemsets was proposed to derive high utility itemsets in an incrementally inserted way. In real-world applications, transactions are not only inserted into but also deleted from a database. In this paper, a maintenance algorithm is thus proposed for reducing the execution time of maintaining high utility itemsets due to transaction deletion. Experimental results also show that the proposed maintenance algorithm runs much faster than the batch approach.

Keywords Utility mining · Maintenance · Transaction deletion · Two-phase approach · FUP concept

C.-W. Lin (✉) · L. Kong
IIIRC, School of Computer Science and Technology, Institute of Technology Shenzhen Graduate School, Xili, Shenzhen, People's Republic of China
e-mail: jerrylin@ieee.org

L. Kong
e-mail: konglingping@utsz.edu.cn

C.-W. Lin
Shenzhen Key Laboratory of Internet Information Collaboration Harbin, Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, People's Republic of China

G.-C. Lan
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China
e-mail: rrfoheiay@gmail.com

T.-P. Hong
Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, Republic of China
e-mail: tphong@nuk.edu.tw

T.-P. Hong
Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_112, © Springer Science+Business Media Dordrecht 2014

Introduction

Association rules mining from binary database is the fundamental approach for knowledge discovery [1–4]. In some applications, frequent itemsets may only contribute a small portion to the overall profit, while non-frequent ones may contribute a large portion to the profit [5]. In the past, utility mining [6] was thus proposed to partially solve the above limitations. It may be thought of as an extension of frequent itemset mining with the sold quantities as a quantitative database with the item profits considered [5, 7–10].

In real-world applications, transactions in the database do not usually remain in a stable condition. Some transactions may be inserted or deleted from an original database. The discovered frequent itemsets may become invalid, or some new implicit information may emerge in the whole updated database. Lin et al. thus proposed an incremental algorithm for transaction insertion in high utility mining [11] based on the Fast UPdated (FUP) concepts [12]. In addition to transaction insertion, transaction deletion is also commonly seen in real-world applications. In this paper, a maintenance algorithm is thus proposed to update the mined utility itemsets for deleted transactions. The FUP2 (Fast UPdated) algorithm [13] is then adopted in the proposed algorithm to reduce the time of re-processing the whole updated database. In the experiments, it shows that the proposed algorithm has the better performance than the two-phase algorithm [9].

Review of Related Works

FUP Concepts

Data mining can be divided into many specific areas due to its applications [14–16], but the most common approach is to extract patterns or rules from data sets in a particular representation [1, 2, 4]. In real-world applications, transaction databases usually grow over time and the mined association rules must be re-evaluated. Considering an original database and some transactions to be deleted, four cases may arise based on the FUP2 concepts [13] showing in Fig. 1. Each case is then executed by its designed procedure.

Mining High Utility Itemsets

Yao et al. proposed the utility model by considering quantities and profits of items [10]. Chan et al. proposed the topic of utility mining to discover high utility itemsets [7]. Liu et al. then presented a two-phase algorithm for fast discovering high utility itemsets by adopting the downward-closure property [9] and named it

Fig. 1 Four cases when transactions are deleted from an existing database

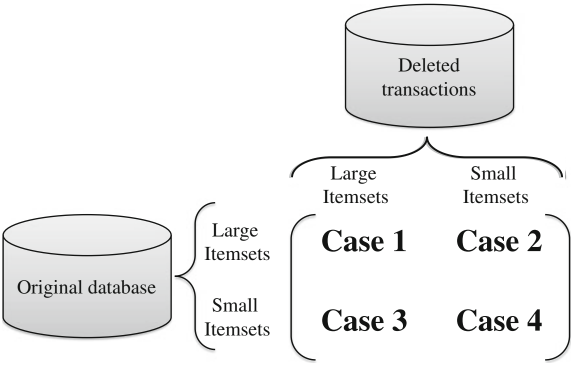


Table 1 An illustrated database

TID	A	B	C	D	E	tu
1	6	0	0	0	0	36
2	0	1	0	5	0	37
3	0	6	0	0	0	12
4	0	0	5	0	0	75
5	0	0	0	0	8	80
6	2	0	2	0	4	82
7	0	1	0	4	0	30
8	0	4	0	0	0	8
9	0	2	3	7	0	98
10	0	0	0	0	1	10
11	0	5	2	5	0	75
12	0	3	0	3	0	27
twu	118	287	330	267	172	Total utility = 570

as the transaction-weighted-utilization (TWU) model. Lin et al. then proposed an incremental approach [11] for efficiently mining the high utility itemsets based on TWU mode [9]. An example is given below to briefly demonstrate the TWU model. Assume the original database is shown in Table 1 consisting of 12 transactions with 5 items, denoted by A to E. The profits for items are $\{A{:}6, B{:}2, C{:}15, D{:}7, E{:}10\}$.

The *transaction utility* (***tu***) in Table 1 is the summation of the item utilities in the transaction. The *transaction-weighted utilization* (***twu***) value of each item is the summation of the ***tu*** value where the item existing in the transaction. Suppose the upper (minimum high utility) threshold is set at 30 %. The minimum high utility count is thus calculated as (570 × 0.3) (= 171). The final results for the high (large) transaction-weighted utilization itemsets and their *actual utilities* (***AU***) are $\{twu(B) = 287,\ AU(B) = 44,\ twu(C) = 330,\ AU(C) = 180,\ twu(D) = 267,$ $AU(D) = 168,\ twu(E) = 172,\ AU(E) = 130,\ twu(B, C) = 173,\ AU(B, C) = 104,$

$twu(B, D) = 267, AU(B, D) = 192, twu(C, D) = 173, AU(C, D) = 159, twu(B, C, D) = 173, AU(B, C, D) = 174\}$.

The Proposed Maintenance Algorithm

The transaction-weighted utilization itemsets and its actual utility values are firstly derived from the original database before transaction deletion. The details of the proposed maintenance algorithm for transaction deletion are described below.

The maintenance algorithm for transaction deletion:

INPUT: A profit table P of items, an original database D, a minimum high utility threshold λ, the total utility TU^D of D, the large (high) transaction-weighted utilization itemsets $HTWU^D$ with their transaction-weighted utilization values and actual utility values discovered from D, and a set of deleted transactions d extracted from the original database D.

OUTPUT: A set of high utility itemsets (HU^U) for the updated database U (= $D - d$).

STEP 1: Calculate the utility value u_{jk} of each item i_j in each deleted transaction t_k as $u_{jk} = q_{jk} \times p_j$, where q_{jk} is the quantity of i_j in t_k and p_j is the profit of i_j. Accumulate the utility values all items in each deleted transaction t_k as the transaction utility tu_k. That is:

$$tu_k = \sum_{j=1}^{m} u_{jk}.$$

The total utility in the deleted transactions is the summation of all transaction utilities in the deleted transactions as:

$$TU^d = \sum_{k=1}^{n} tu_k.$$

STEP 2: Calculate the updated minimum high utility count muc^U (= $TU^D - TU^d$) $\times$ λ.

STEP 3: Set $k = 1$, where k records the number of items in the itemset s currently being processed.

STEP 4: Generate the candidate k-itemsets and calculate their transaction-weighted utility $twu^d(s)$ from the deleted transactions as the summation of the utilities of the deleted transactions which include i_j. That is:

$$twu^d(s) = \sum_{i_j \in t_k} tu_k.$$

STEP 5: Check whether the $twu^d(s)$ of each k-itemset in the deleted transactions is larger than or equal to the minimum high utility count muc^d (= $TU^d \times \lambda$).

If s satisfied the condition, put it in the set of high transaction-weighted utilization k-itemsets for the deleted transactions, $HTWU_k^d$.

STEP 6: For each k-itemset s in the set of high transaction-weighted utilization k-itemsets ($HTWU_k^D$) from the original database, if it also appears in the set of high transaction-weighted utilization k-itemsets ($HTWU_k^d$) in the deleted transactions, do the following substeps (**Case 1**):

Substep 6-1:Set the updated transaction-weighted utility of itemset s as:

$$twu^U(s) = twu^D(s) - twu^d(s),$$

where $twu^D(s)$ is the high transaction-weighted utility of itemset s in the original database ($HTWU_k^D$) and $twu^d(s)$ is the high transaction-weighted utility of itemset s in the deleted transactions ($HTWU_k^d$), respectively.

Substep 6-2: Check whether the updated transaction-weighted utility $twu^U(s)$ is larger than or equal to the updated minimum high utility count muc^U. If s satisfied the above condition, put it in the set of updated high transaction-weighted utilization k-itemsets, $HTWU_k^U$. Otherwise, remove s from the set of high transaction-weighted utilization k-itemsets ($HTWU_k^D$) in the original database.

STEP 7: For each k-itemset s in the set of high transaction-weighted utilization k-itemsets ($HTWU_k^D$) from the original database, if it does not appear in the set of high transaction-weighted utilization k-itemsets ($HTWU_k^d$) in the deleted transactions, do the following substeps (**Case 2**):

Substep 7-1: Set the updated transaction-weighted utility of itemset s as:

$$twu^U(s) = twu^D(s) - twu^d(s).$$

Substep 7-2: Put s in the set of updated high transaction-weighted utilization k-itemsets, $HTWU_k^U$.

STEP 8: For each k-itemset s, if it does not appear both in the set of high transaction-weighted utilization k-itemsets ($HTWU_k^D$) from the original database and in the set of high transaction-weighted utilization k-itemsets ($HTWU_k^d$) in the deleted transactions, do the following substeps (**Case 4**):

Substep 8-1: Rescan the original database to determine the transaction-weighted utility $twu^D(s)$ of itemset s.

Substep 8-2: Set the updated transaction-weighted utility of itemset s as:

$$twu^U(S) = twu^D(S) - twu^d(S).$$

Substep 8-3: Check whether the updated transaction-weighted utility $twu^U(s)$ is larger than or equal to the updated minimum high utility count muc^U. If s satisfied the above condition, put it in the set of updated high transaction-weighted utilization k-itemsets, $HTWU_k^U$.

STEP 9: Generate the candidate ($k + 1$)-itemsets from the set of high transaction-weighted utilization k-itemsets ($HTWU_k^U$) in the updated database.

STEP 10: Set $k = k + 1$;

STEP 11: Repeat STEPs 4 to 10 until no new candidate itemsets are generated.

STEP 12: Process each itemset *s* in the set of $HTWU^U$; if $au^U(s) \geq muc^U$, itemset *s* is a high utility itemset. Put itemset s into the set of HU^U.

STEP 13: Set $HTWU^D = HTWU^U$ as the set of the original large (high) transaction-weighted utilization itemsets for the next transaction deletion in maintenance mining.

After STEP 13, the final high utility itemsets for the updated database can then be updated.

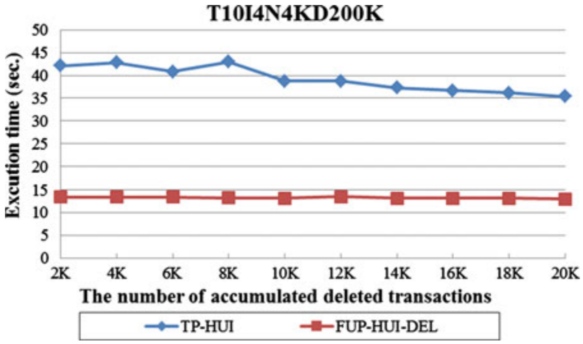
Experimental Results

In the experimental evaluation, the TP-HUI algorithm [9] and the proposed FUP-HUI-DEL algorithm are then compared to respectively show the performance. When transactions are deleted from the original database, the TP-HUI algorithm has to re-scan the updated database for extracting the updated high utility itemsets in a batch way. The proposed FUP-HUI-DEL algorithm divides the itemsets into four parts according to whether their transaction-weighted utilizations are large or small in the original database and in the deleted transactions. Each part is then processed on its own way to update the discovered knowledge.

The IBM data generator [17] is used to generate the simulation dataset called T10I4N4KD200K. Firstly, 200,000 transactions are used to initially mine the large (high) and pre-large transaction-weighted utilization itemsets with their actual utility values. Each 2,000 transactions are then sequential deleted from the original database bottom up at each time. The minimum high utility threshold is set at 0.2 % to evaluate the performance comparing to the TP-HUI and the proposed FUP-HUI-DEL algorithms. The results are then shown in Fig. 2.

In Fig. 2, the TP-HUI algorithm has to process the updated database in a batch way whenever transactions are deleted. The proposed FUP-HUI-DEL algorithm requires re-scanning the whole database if it is necessary to re-calculate the itemsets appearing in in case 4. Experiments are also made to evaluate the

Fig. 2 The comparisons of the execution time for transaction deletion



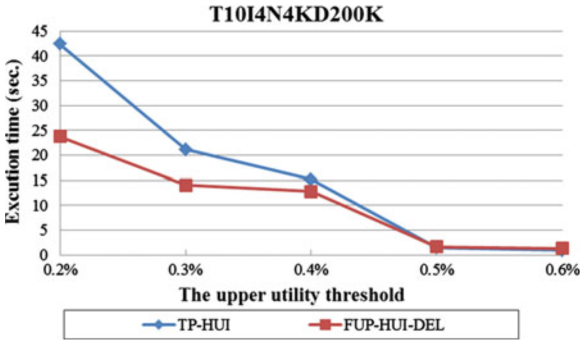


Fig. 3 The comparisons of the execution time in different minimum utility thresholds

efficiency of the proposed FUP-HUI-DEL algorithm in different minimum high utility threshold values shown in Fig. 3. The minimum high utility threshold is then set from 0.2 to 0.6 %, increases 0.1 % each time.

It can easily be observed from Fig. 3 that the execution time of the proposed FUP-HUI-DEL is much less than that by the TP-HUI algorithm for handling transaction deletion in different minimum high utility thresholds.

Conclusions

In this paper, a maintenance mining algorithm is thus proposed for transaction deletion based on the two-phase and FUP2 approaches. When transactions are deleted from the original database, the proposed maintenance algorithm processes the itemsets into four parts according to whether they are high (large) transaction-weighted utilization itemsets or small in the original database and in the deleted transactions to update the discovered high utility itemsets. Each part is then processed by its own procedure. Experimental results also show that the proposed maintenance algorithm runs faster than the batch approach for updating the high utility itemsets.

Acknowledgments This research was partially supported by Shenzhen peacock project, China, under contract No. KQC201109020055A, and Shenzhen Strategic Emerging Industries Program under Grants No. ZDSY20120613125016389.

References

1. Agrawal R, Imielinski T, Swami A (1993) Database Mining: a Performance Perspective. IEEE Trans Knowl Data Eng 5:914–925
2. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the international conference on very large data bases, pp 487–499

3. Hong TP, Lin CW, Wu YL (2008) Incrementally fast updated frequent pattern trees. Expert Syst Appl 34:2424–2435
4. Park JS, Chen MS, Yu PS (1997) Using a hash-based method with transaction trimming for mining association rules. IEEE Trans Knowl Data Eng 9(5):813–825
5. Liu Y, Liao WK, Choudhary A (2005) A fast high utility itemsets mining algorithm. In: Proceedings of the international workshop on utility-based data mining, pp 90–99
6. Yao H, Hamilton HJ, Butz CJ (2004) A foundational approach to mining itemset utilities from databases. In: Proceedings of the siam international conference on data mining. pp 211–225
7. Chan R, Yang Q, Shen YD (2003) Mining high utility itemsets. In: Proceedings of IEEE international conference on data mining, pp 19–26
8. Lan GC, Hong TP, Tseng VS (2011) Discovery of high utility itemsets from on-shelf time periods of products. Expert Syst Appl 38(5):5851–5857
9. Liu Y, Liao WK, Choudhary A (2005) A two-phase algorithm for fast discovery of high utility itemsets. Lect Notes Comput Sci 3518:689–695
10. Yao H, Hamilton HJ (2006) Mining itemset utilities from transaction databases. Data Knowl Eng 59(3):603–626
11. Lin CW, Lan GC, Hong TP (2012) An incremental mining algorithm for high utility itemsets. Expert Syst Appl 39(8):7173–7180
12. Cheung DW, Jiawei H, Ng VT, Wong CY (1996) Maintenance of discovered association rules in large databases: an incremental updating technique. In: Proceedings of the international conference on data engineering, pp 106–114
13. Cheung DW, Lee SD, Kao B (1997) A general incremental technique for maintaining discovered association rules. In: Proceedings of the international conference on database systems for advanced applications, pp 185–194
14. Hong TP, Wu CH (2010) An improved weighted clustering algorithm for determination of application nodes in heterogeneous sensor networks. J Inf Hiding Multimedia Sig Process 2(2):173–184
15. Lin CW, Hong TP, Chang CC, Wang SL (2013) A greedy-based approach for hiding sensitive itemsets by transaction insertion. J Inf Hiding Multimedia Sig Process 4(4):201–227
16. Lin CW, Hong TP (2013) A survey of fuzzy web mining. Wiley Interdisc Rev: Data Min Knowl Discovery 3:190–199
17. IBM: quest synthetic data generation code (1996) Available: http://www.almaden.ibm.com/cs/quest/syndata.html

Design and Implementation of a LBS General Website Content Extract System for Android

Yongbo Chen, Xin Zhou and Yun Liu

Abstract Every day more than 1 million new Android devices are activated worldwide [1]. This paper researched on the topic of design and implementation of a location based public opinion information system on android system, which provides personal or group users, such as people, enterprise, and government an easy access to information what they care about in specified location area with specified focus and interest. Those latest information would extract from some API-opened websites, such as a microblog website, a specified social network website, and take advantage of Google Geocoding API, we can extends the ability of the system to an infrastructure for other applications.

Keywords Content extract · LBS · Android

Introduction

Every day more than 1 million new Android devices are activated worldwide [1]. This paper researched on the topic of design and implementation of a location based public opinion information system on android system, which provides

Y. Chen · Y. Liu (✉)
School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
e-mail: liuyun@bjtu.edu.cn

Y. Chen
e-mail: 12120065@bjtu.edu.cn

Y. Chen · Y. Liu
Key Laboratory of Communication and Information Systems, Beijing Municipal
Commission of Education, Beijing Jiaotong University, Beijing 100044, China

X. Zhou
China Information Technology Security Evaluation Center, Beijing, China
e-mail: zhoux@itsec.gov.cn

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_113, © Springer Science+Business Media Dordrecht 2014

personal or group users, such as people, enterprise, and government an easy access to information what they care about in specified location area with specified focus and interest. Those latest information would extract from some API-opened websites, such as a microblog website, a specified social network website, and take advantage of Google Geocoding API, we can extends the ability of the system to an infrastructure for other applications.

Related Work

OAuth

OAuth is an open standard for authorization [2]. OAuth provides a method for clients and third-party applications to access server resources on behalf of a resource owner (such as a different client or an end-user). It takes an advantage of making resource sharing easily without provide the resource owner's, credentials (e.g. username and password pair), using user-agent redirections, which make end-users risk less security problem [3].

OAuth is required to take advantage of Open API [4]. Before OAuth standard, there are generally three traditional methods to verify an internet user:

a) Using username and password;
b) Using a temporary secret link address, such as a link in an account activation email;
c) Using cell phone verification code.

Those methods are either vulnerable structure from security point of view or cost extra resources, and would bring some security risk. OAuth was begun in November 2006 when Blaine Cook was developing the Twitter OpenID implementation [3]. By now OAuth 2.0 is the next evolution of the OAuth protocol.

By using OAuth protocol, the system would provide users a more secure permission control and data storage service. Generally, Open API system would take a validation with OAuth protocol for applications. So the system is designed with an OAuth Configuration Manage component, to get the corresponding OAuth access token when accessing the variety Open API system.

LBS

Location-based services are a general class of computer program-level services used to include specific controls for location and time data as control features in computer programs [5].

LBS is widely used in social network websites, IM software, indoor object search, entertainment and so on, and plays an important role in many people's daily life. Android system provides a well-organized location service. With Android Location package, it's efficient to use the location service, and by communicate with server side; the location information could be used to find out nearby events on internet or some spot activity on target location area. There are examples that are common LBS application type recommending social events in a city, requesting the nearest business or service, such as an ATM or restaurant, locating people on a map displayed on the mobile phone etc.

Android Location Listener interface provides an access to the device's location changing.

Webpage Content Extracting

Web pages often contain clutter (such as pop-up ads, unnecessary images and extraneous links) around the body of an article that distracts a user from actual content [6].

To make large scale of data fetching from website possible, automatically extract the content is an essential solution. By researching on extract the useful content of a website, further more application are given, such as make small screen devices efficient to visited content wrapped by order less advertisement in webpages.

Opera re-organized the page content with content analyzing, by extracting the main content of websites and make them user-friendly on small screen devices.

Currently, most used methods to extract web content is xpath-extract and regex-extract, both of them need a hand-tailored extractor. Suhit Gupta advised a method to generate content extractors based on DOM information of the html file [6].

Rahman et al. propose another technique that uses structural analysis, contextual analysis, and summarization. The structure of an HTML document is first analyzed with a definition of 'zones' and then properly decomposed into smaller subsections. The content of the each section well be extracted then summarized afterwards. However, this proposal still remains to be implemented yet [7]. Design and Prototype

Design of the System

When designing the system, there are two factors that mainly decided the architecture of it:

- As android system provides a variety of UI elements, it's convenient to build friendly GUI programmatically.
- The technical improvements of mobile devices and the growing mobile network bandwidth.

Mobile Client Structure and Responsibility

- Provides current location of the device;
- Calculate the distance between the device and a specified location;
- Maintenance of connection to server;
- Receive notification of server with an acceptable delay;
- Maintenance of location data synchronization;

Server Structure and Responsibility

- A OAuth Configuration Manage module, providing requests sent to different Open API websites with OAuth access tokens;
- A permission control module, between different users;
- A data providing and persistence module, containing client synchronizing location data. Composed of such sub modules:
  - A synchronizing configuration maintenance module;
  - A notification module;
- A Website information synchronizing module, which implemented by two cooperation sub-modules with different mechanisms:
  - A high performance multi-thread API-based information extract module which extracts abstract and partial of contents;
  - An aiding information extract sub module by analyzing html content with xpath and regex;
- A configuration maintenance module, manages Auto information of different API-opened website that requests OAuth;

Prototype of the System

OAuth Configuration Manage

OAuth access token should be applied and maintained manually, this module doesn't generate OAuth for client, but take responsibility of make each request sent to Open API website acceptable on target, and won't be denied due to OAuth token exception or error.

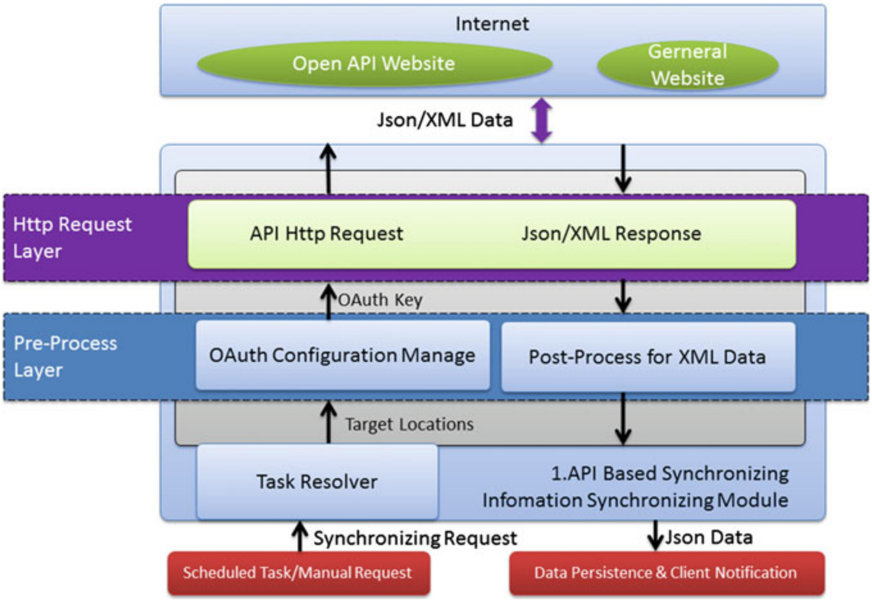


Fig. 1 Structure of API based sub-module

Information Synchronizing Module

In order to take use of the latest information on websites effectively, we need to make the system able to automatically synchronize content from some target websites. Reasonably, the system had been designed to be capable for both executing scheduled task and manual request once or repeat in a specified frequency.

There are two different methods to execute the information synchronizing, API-based and content analyzing based. This module composed of two sub-modules according to those two methods, shown in the following graph (Figs. 1, 2):

In the API-based case, an OAuth key is generally needed, the OAuth configuration manage component manages a variety keys that manually applied from target websites that provides Open API. The Http Request Layer uses a restful web service to proxy scheduled task or request, transform the synchronizing work detail into concrete URL request that can be easily processed through third party tools such as Apache Http Commons. The request result should be in Json or XML format, to make data consistent, we transform the XML data into Json data and processed in following up procedures.

In a synchronizing progress The API based sub-module provides more accurate results, while due to API restriction, less in amount and has a limited usage in rate. Meanwhile, the content analyzing sub-module support an unrestrained usage by mocking a common user requesting the website automatically, the actions containing searching keywords, retrieving the content, and automatically extract the

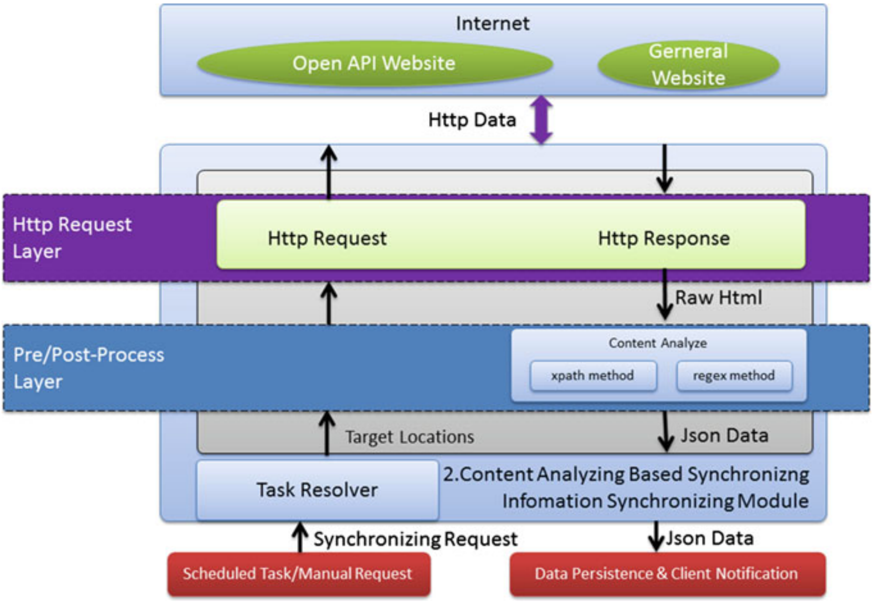


Fig. 2 Structure of content analyzing based synchronizing sub-module

key opinion of result page. Both of the sub-modules making the results into a consistency data structure presented as well formatted Json and they would be persisted into database for further usage such as client notification.

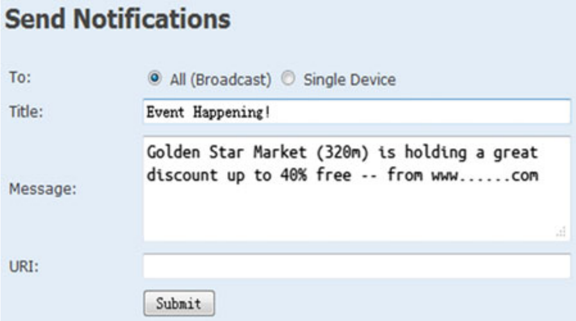
Notification Module

The server would notification the client with messages. Androidpn is used here to provide powerful and comprehensive notification service. As the server side sends a short message, target client or target client group would immediately response with a short ring or some vibration due to the cell phone hint mode configuration.

Screenshots

Figure 3 shows a graph of manually sending a notification, while running the system, this procedure would be called by notification module automatically by sending an http request posting the messages as parameters, and the procedure would be without GUI.

Fig. 3 A screen shot of sending manually notification



Send Notifications

To: ● All (Broadcast) ○ Single Device

Title: Event Happening!

Message: Golden Star Market (320m) is holding a great discount up to 40% free -- from www......com

URI:

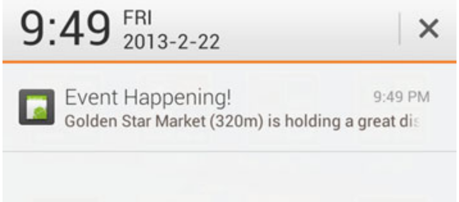
Figure 4 shows the client receiving the notification which was sent by server before. We use XMPP (Extensible Messaging and Presence Protocol) to implementation the notification. The XMPP tool here used is Androidpn which implements the XMPP based on asmack. The Androidpn framework contains a client side project and a sever side project. On the client side, it establishes a listener on port 5222, the XMPPConnection class of Androidpn is responsible for communication with the server side. On the server side, it's in fact a light weight Http container, which would send notification and handle the request sent by client. The top level of the server side consist of Session Manager that manages session between client and server, Auth Manager that authenticate client users, Presence Manager that manages the login state of client users, and Notification Manager that send notifications to client side. The scheme guarantees security, simplify, make the client side independent from Android operation system version. It also keeps the system extensible for further design.

Figure 5 shows that the application displayed notification location on map with an icon of broadcast.

Summary and Prospects

As price keep being lower and functions keep enhancing, smart cell phone popularize more every day, meanwhile, mobile communication technology is developing. In such an environment, LBS are greatly increasing. The combination of

Fig. 4 Client receiving notification



9:49 FRI 2013-2-22

Event Happening! 9:49 PM

Golden Star Market (320m) is holding a great di

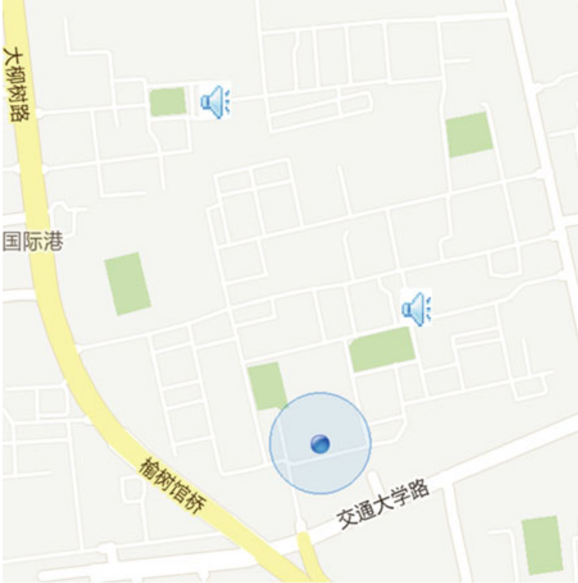


Fig. 5 Display location on map

LBS and Website information would be a promising trend. Some applications on smart phones already benefit from the combination of LBS and some social network, in future, there would be more applications take advantage of LBS and the combination of LBS with other service would provide powerful enhancement to daily life.

Acknowledgments This work has been supported by the National Natural Science Foundation of China under Grant 61172072, 61271308, the Beijing Natural Science Foundation under Grant 4112045, the Research Fund for the Doctoral Program of Higher Education of China under Grant W11C100030, the Beijing Science and Technology Program under Grant Z121100000312024.

References

1. http://developer.android.com/about/index.html
2. http://en.wikipedia.org/wiki/OAuth
3. http://tools.ietf.org/html/rfc5849
4. Lee D, Shin J, Lee S (2012) Design and implementation of user information sharing system using location-based services for social network services. J Meas Sci Instrum 3(2)
5. http://en.wikipedia.org/wiki/Location-based_service
6. Gupta S et al (2003) DOM-based content extraction of HTML documents. In: Proceedings of the 12th international conference on world wide web. ACM, 2003

MR. Eye: A Multi-hop Real-time Wireless Multimedia Sensor Network

Yanliang Jin, Yingxiong Song, Jian Chen, Yingchun Li, Junjie Zhang and Junni Zou

Abstract Simple data, such as temperature, humidity and luminary, gathered by nodes in wireless sensor networks (WSNs), can hardly meet the growing needs for real-time information. The availability of low-cost hardware is enabling the development of wireless multimedia sensor networks (WMSNs), i.e., networks of resource-constrained wireless devices that can retrieve multimedia content such as video and audio streams, still images, and scalar sensor data from the environment. However, multi-function means complex design, effective hardware, robust protocol and many other considerations. Comparing some of those current mainstream WMSN test beds, limits and shortcomings can be easily found. In this paper, MR. Eye is designed. Nodes, within MR. Eye, built on OMAP3530, are capable of coping with video, images and other real-time information. Moreover, the load balanced airtime with energy awareness (LBA-EA) is applied in this system and nodes can be easily controlled by SNMP gateway. The last but not the least, real-time video data can be transmitted fluently in multi-hop scenarios.

Keywords WMSNs · Mesh protocol · Sensor nodes · Multi-hop · Real-time

Introduction

Wireless sensor network (WSN) is currently a frontier research field, involving a high degree of interdisciplinary and highly integrated knowledge, drawing intense international attentions. The development of technologies, such as sensor technology, micro electro-mechanical systems, modern network and wireless communication, push forward the generation and development of modern wireless sensor

Y. Jin (✉) · Y. Song · J. Chen · Y. Li · J. Zhang · J. Zou
Key Laboratory of Specialty Fiber Optics, Shanghai University, Shanghai, China
e-mail: baizhsh@126.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_114, © Springer Science+Business Media Dordrecht 2014

networks. Nonetheless, along with the increasingly complex of environment, simple data gathered from conventional node devices cannot meet inspection needs, which require fine-gained and accurate information. Consequently, wireless multimedia sensor networks (WMSNs) come into being.

Comparing with conventional WSNs, WMSNs possesses apparent attributes in energy distribution, QoS requirements and sensor model, confronting huge challenges both in fundamental theory and on technical implementation level. As far as our best knowledge, those test beds listed above have their own limits more or less. Some of those have low power consumption, but they can only transmit still pictures with low resolution, such as Mica, MicaZ, Yale XYZ, Sun SPOT and Stargate in [1–5] respectively. Others, however, with high power consumption, can merely transmit real-time video within single-hop, such as Stargate and Embedded PC in [2] and [3] respectively. In this paper, we propose the MR. Eye, which incorporates wireless multimedia sensor nodes and a network management system. It possesses functions such as video and data acquisition, network traffic control, system management and configuration of wireless sensor node devices. Network management system mainly fulfill following tasks: real-time topology monitoring, viewing performance data, viewing alarm information, real-time video monitoring, user configuration management and device configuration management. Wireless multimedia sensor nodes are mainly in charge of video capturing, temperature and humidity data acquisition, GPS data acquisition, extracting data of battery state and current state. Nodes can also send data to network management system through agent-side software with UDP protocol. What's more, node devices have high-performance power management function.

Hardware Architecture

The hardware architecture of our proposed sensor node adopts next-generation mobile applications processor OMAP3530 as host processor, which is produced by TI Corporation for low-power portable applications such as smart phones, GPS system and notebook. OMAP3530 integrates an ARM Cortex-A8 core, a TMS320C64x+ DSP core, a graphics engine, a video accelerator and other multimedia peripherals on one single chip. We also realize USB extension in MR. Eye, which enables connection with 802.11 wireless modules, Ethernet ports and other USB devices. Each node device carries a digital and analog video capture port, connecting to a 2 GB SD memory card and various types of common sensor nodes through local bus and I^{2}C bus respectively. Besides, two separate power modules and clock modules provide the system with stable power supply and clock signal. Figure 1 plots the structure of a whole node device.

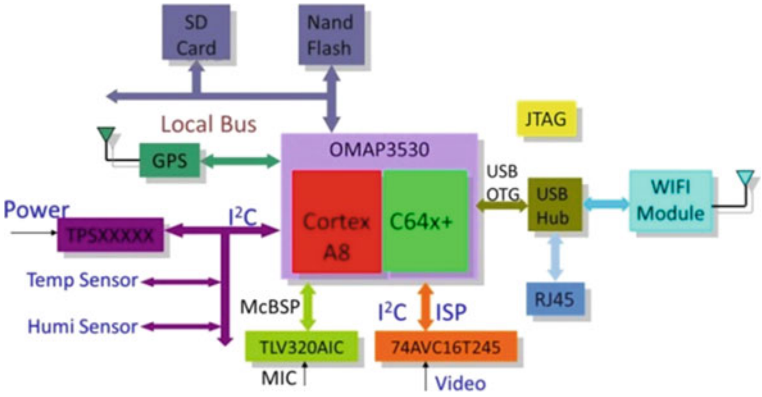


Fig. 1 Hardware architecture

Software Architecture

GStreamer

GStreamer, a framework used to build audio and video applications, is applied in our system. Video processing task is assigned on the DSP core by applying relevant video codec plug-in. The main tasks of GStreamer in our designed node devices are video acquisition, adjustment, coding, RTP package and transmission. Video is displayed in PC control system after the received video data is unpacked and decoded respectively.

Basing on the framework discussed above, we can build up multi-media tunnel basing on GStreamer. Real-time video data acquired by 'v4l2src' plug-in is adjusted by 'video scale' plug-in. Video data, after adjustment, is buffered and compressed through 'TIVidenc1', a plug-in where algorithm, bit rate and frame rate can be set. Thereafter, data stream flows into 'rtpmp4vpay', where data is packed with RTP to guarantee real-time video transmission. Finally, data enters 'udpsink' for transmitting data to destination. In PC system, the tunnel is just the opposite from the one in the node. PC asks for video data from nodes through 'udpsrc' and receives data packages. The header information of RTP package and payload data is unpacked by 'rtpmp4vdepay'. Then payload data is decoded by 'ffdec_mpeg4' to be played by 'dshowvideosink'. Figure 2 plots the structure of the GStreamer multi-media tunnel.

SNMP Gateway Control

SNMP gateway control [6] can be divided into two parts: network management and network agent. Network management part can be controlled through graphical user interface conveniently. What users need to do is just configuring the management

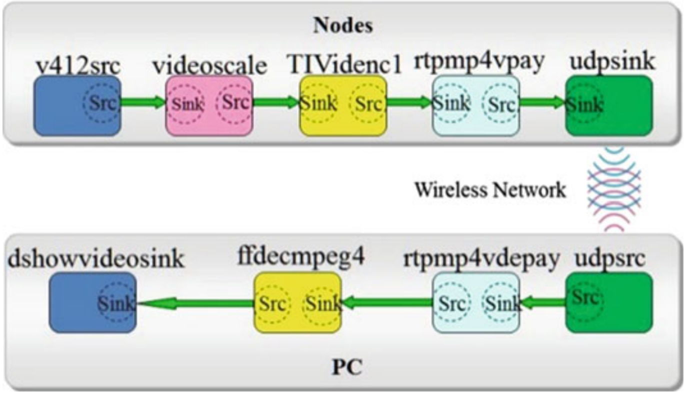


Fig. 2 GStreamer multi-media tunnel

interface. Configuration information is sent to network management module to be analyzed and displayed consequently. Network agent part mainly deals with exception handling, agent management and maintenance, MIB library management and message handling. Management part and agent part are connected by two communication modules. Figure 3 plots the SNMP gateway control framework.

LBA-EA and Mesh Protocol

In MR. Eye, we adopt Mesh topology and HWMP (hybrid wireless mesh protocol) [7] as routing protocol. We propose a new routing metric, load balanced airtime

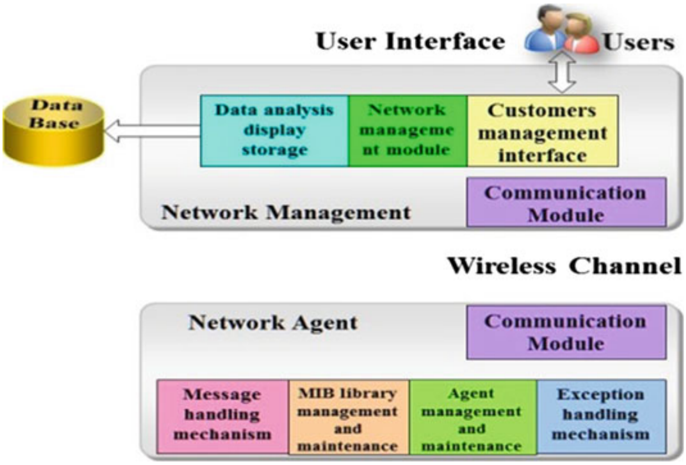

User Interface
Users
Data Base
Data analysis display storage
Network manageme nt module
Customers management interface
Communication Module
Network Management
Wireless Channel
Network Agent
Communication Module
Message handling mechanism
MIB library management and maintenance
Agent management and maintenance
Exception handling mechanism

Fig. 3 SNMP gateway control framework

with energy awareness (LBA-EA) [8], which is based on airtime link metric defined in IEEE 802.11s draft. The proposed metric is composed of two parts. One is a load-aware airtime factor, and the other is an energy-aware factor, with a tunable parameter. The weight of a path is calculated as follows:

$$W = (1 - \beta) \times \sum_{i=1}^{n_L} C_{lba}^i + \beta \times \sum_{j=1}^{n_N} \frac{E_{init}^j}{E_r^j} \tag{1}$$

where C_{lba}^i is an enhancement of airtime and represents a more accurate expected transmission time of link i. n_L is the number of links along the selected path. E_r^j is the residual value of node j's energy with its initial energy E_{init}^j, and n_N is the number of nodes along the selected path.

In (1),

$$C_{lba}^i = \left(T_o + \frac{L_t}{r_i} + d_q^i\right) \times \frac{f(\rho_i)}{1 - \rho_i} \times \frac{1}{1 - e_f^i} \tag{2}$$

where these parameters To, Lt, r_i and e_f^i are similar to those defined in airtime and d_q^i denotes the queuing delay of a forwarding node in the ith link, which is calculated as:

$$d_q^i = \frac{E\left(L_q^i\right)}{r_i} \tag{3}$$

where $E\left(L_q^i\right)$ is an average value of queue length in a given period of time.

And ρ_i denotes the factor of inter-flow interference which caused by contending neighbors, while $f(\rho_i)$ is a weight function that depends on ρ_i.

$$f(\rho_i) = \begin{cases} 1 & \rho_i < \rho_0 \\ K & \rho_0 \le \rho_i \le \rho_{\max} \\ \infty & \rho_i > \rho_{\max} \end{cases} \tag{4}$$

IEEE 802.11s protocol [9] is the most part of MAC80211 protocol framework, which contains network mgt, QoS rqt, security mgt, neighbor mgt, routing control and data forwarding. MAC80211 provides underlying drivers and user programs with uniform interface, facilitating the development.

Results

To evaluate MR. Eye, several node devices are manufactured to form a certain network on the campus. Figure 4 plots the hardware of the node and Fig. 5 shows the user interfaces in MR. Eye system. The details of hardware architecture have been elaborated in section Hardware Architecture. In this section, the simulation and the experiments of both indoor and outdoor scenarios are done.

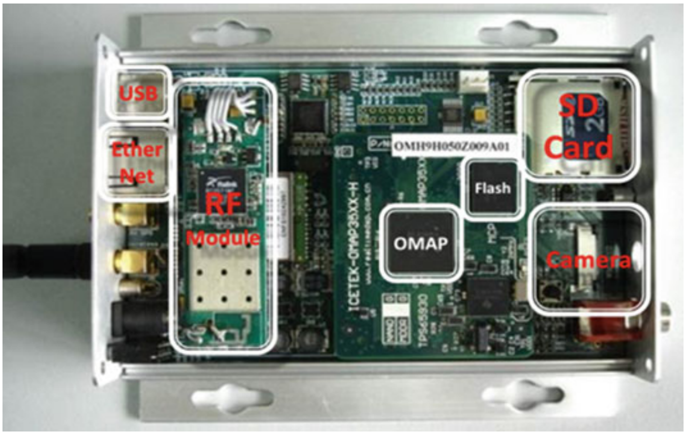


Fig. 4 Internal structure of a node

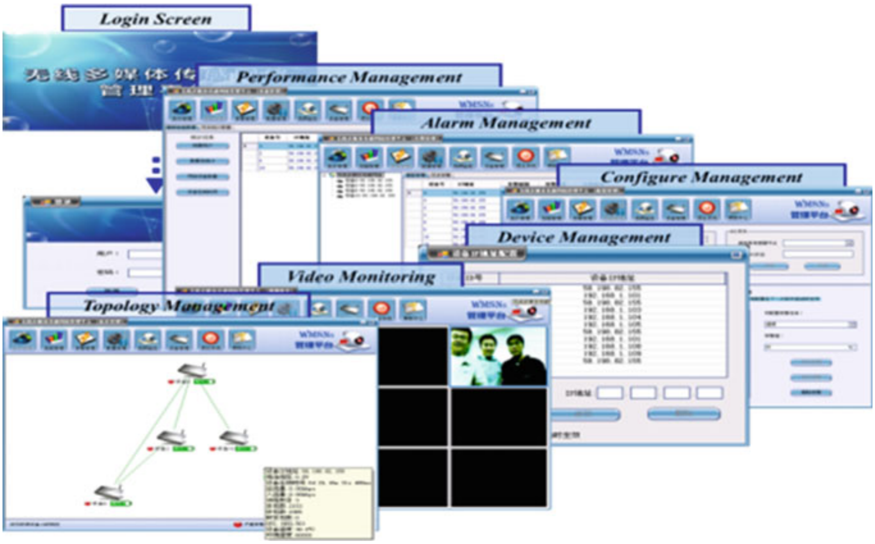


Fig. 5 User interfaces

Simulation

In this section, we implement our simulation in NS3 [10] to evaluate the performance of LBA-EA. we place 50 nodes randomly in an area for 600 × 600 m and the node 0 is the sink. We use 802.11e EDCA in Data/ACK mode for medium access and use RM-AODV defined in 802.11s as path selection protocol. We consider UDP to be the transport layer and assume that all 10 flows generate data at a constant rate. We consider the packet size of 1,024 bytes. The sources of the flows are randomly chosen and the destination is the sink node. Each simulation run has been executed 100 times and the average results are plotted in the graphs.

As shown in Fig. 6, LBA-EA HWMP has higher delivery rate than HWMP because LBA-EA can disperse flows into different paths thus reduce the probability of collision while Airtime introduces flows into a single path with high data rate. As is shown in Fig. 7, we can see that LBA-EA HWMP has less normalized routing load than HWMP when the number of source nodes increases from 5 to 30. The new scheme uses the load balanced to control the broadcasting information and decrease the overhead.

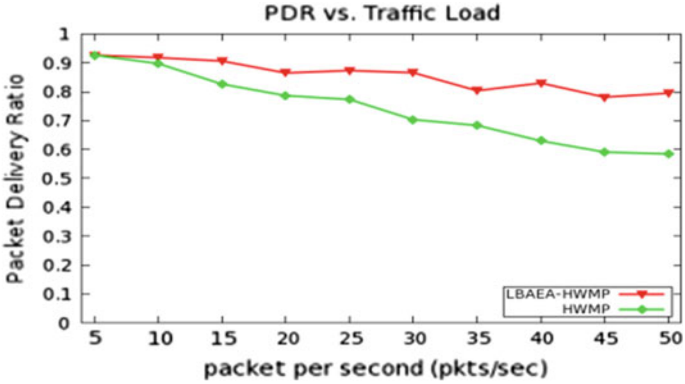


Fig. 6 Comparation of packet delivery ratio

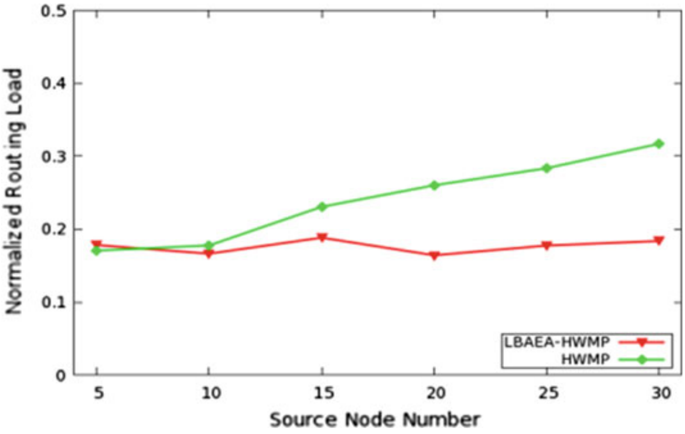


Fig. 7 Comparation of normalized routing load

Indoor Scenario

For indoor experiment, three major aspects are tested: generating dynamic topology, bandwidth and time-jitter respectively.

In bandwidth experiment, to form a three-hop chain topology, four nodes are put in linear shape in the corridor of the lab. The bandwidth of this scenario is plotted in Fig. 8. Same experiment condition with the bandwidth experiment, the time-jitter between the two ends of the chain topology is measured. Figure 9 plots the result of time-jitter. Clearly, the average time-jitter stays around 10 ms, which is acceptable for real-time video data.

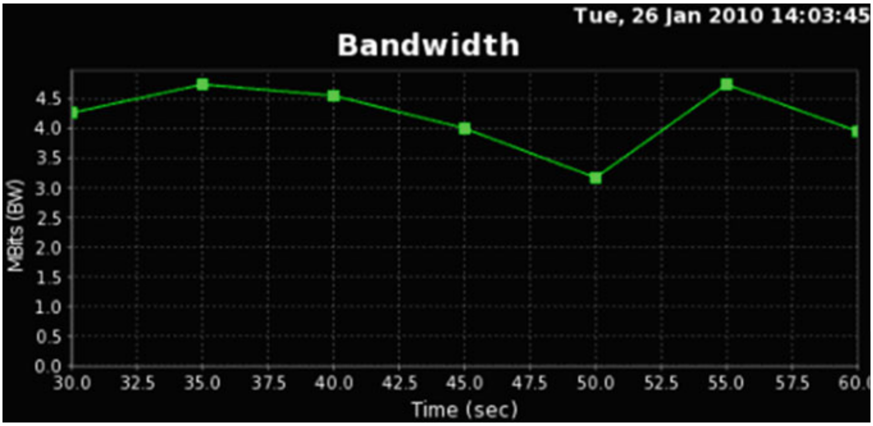


Fig. 8 Bandwidth experiment

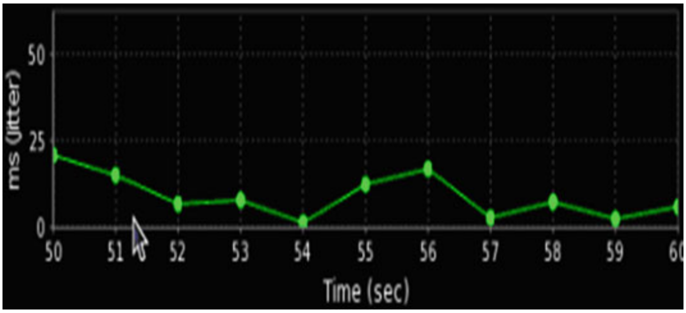


Fig. 9 Time-jitter experiment

Outdoor Scenario

As for outdoor scenario, we locate 8 nodes on the campus. In Fig. 10, we set node 4 as the sink node, which is connected to the SNMP gateway control system. Through the gateway system, real-time monitoring video from other distributed nodes can be easily displayed on the screen.

Figure 11 shows the video data gathered from node 8 and node 7 respectively. In both two-hop and three-hop scenarios, video can be played frequently.

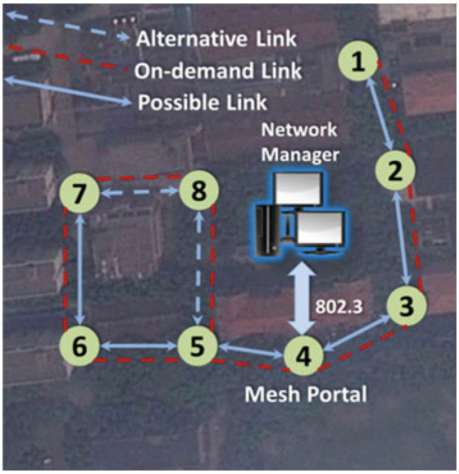


Fig. 10 Practical outdoor experiment

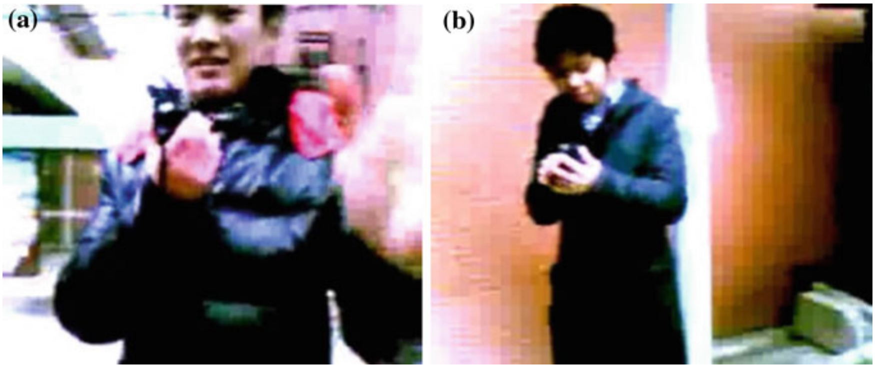


Fig. 11 **a** Video from a two-hop node 8. **b** Video from a three-hop node 7

Conclusion

A high performance MR. Eye is demonstrated in this paper. In MR. Eye, nodes with low power consumption have been designed. Meanwhile, highly effective mesh protocol, powerful SNMP gateway control system and friendly interface are developed. Multi-media nodes are able to collect and process signals like location, temperature, humidity, audio and video. Adopting HWMP routing protocol is suitable for the attributes of mesh topology. SNMP gateway control system facilitates the management of data gathered from nodes. Amiable graphical user interface realizes an easy access to configuration and monitoring of nodes. Indoor and outdoor experiments have proven the fact that MR. Eye is quite competent in practical applications. We will further expand the node number of the test-bed and enhance the system performance in the future.

Acknowledgments The authors would like to thank the reviewers for their detailed comments on earlier versions of this paper. This work was supported by Shanghai leading academic discipline project (S30108, 08DZ2231100), funding of key laboratory of wireless sensor network and communication, shanghai institute of micro system and information technology, Chinese academy of sciences (2012002) and Shanghai science committee (12511503303).

References

1. Gerla M, Xu K (2003) Multimedia streaming in large-scale sensor networks with mobile swarms. In: Papakonstantinou Y (ed) Proceedings of the ACM SIGMOD 2003. ACM Press, New York, pp 72–76
2. Kahn JM, Katz RH, Pister KSJ (1999) Next century challenges: mobile networking for "Smart dust". In: Proceedings of the 5th annual ACM/IEEE Int'l conference on mobile computing and networking. ACM Press, New York, pp 271–278
3. Kulkarni P, Ganesan D, Shenoy P, Lu QF (2005) SensEye: a multi-tier camera sensor network. In: Zhang HJ, Chua T-S (eds) Proceedings of the 13th annual ACM international conference on multimedia'05. ACM Press, New York, pp 229–238
4. Turning vision into reality (2005) http://www.research.sun.com/spotlight/SunSPOTSJune30.pdf
5. Feng W, Code B et al (2003) Panoptes: a scalable architecture for video sensor networking applications. In: Rowe L, Vin H (eds) Proceedings of the ACM Int'l conference on multimedia. ACM Press, New York, pp 151–167
6. Mauro D, Schmidt K (2001) Essential SNMP. O'Reilly Media, California
7. Cornils, M, Bahr M, Gamer T (2010) Simulative analysis of the hybrid wireless mesh protocol (HWMP), wireless conference (EW), European, 12–15, pp 536–543
8. Sun W-Z, Song Y-X, Chen M (2010) A load-balanced and energy-aware routing metric for wireless multimedia sensor network, the third IET international conference on wireless, mobile and multimedia networks, pp 21–24
9. Hiertz GR, Max S, Zhao R, Denteneer D, Lars B (2007) Principles of IEEE 802.11s. In: Computer communications and networks, ICCCN 2007. Proceedings of 16th international conference on Honolulu, HI, pp 1002–1007
10. Zhu Z, Wang P (2010) A multimedia system based on OMAP3530. In: applied mechanics and materials, 40–41, pp 506–509

Effects of the Online VOD Self-learning on English Ability of Taiwanese College Students: The ARCS Approach

Da-Fu Huang

Abstract This study investigated the effect of the Live DVD self-learning system on English learning motivation of students of a Taiwanese technological university. Modeled on the ARCS motivation model of Keller (1987), this study validated the ARCS model, examined the effects of Majors and English Levels on motivation, and tested the proposed structural model involving the effect of the ARCS model on self-assessed English skills. Five hundred and twenty-seven students of three English ability groups completed a survey questionnaire. Statistical procedures including confirmatory factor analysis, factorial ANOVA, and structural equation modelling were performed to address the research questions. The participants were found to have overall positive motivation toward the Live DVD System in terms of the ARCS model. English Levels had significant influence on Attention, with students of Mid and Low-level groups having higher attention to the Live DVD system than High-level group students. For Satisfaction, aside from the main effect of English Levels, the English Levels*Majors interaction effect was also found to be statistical; for non-engineering majors, students of mid and low English abilities were found to have higher satisfaction with the Live DVD System than high proficiency students, while for Level B groups, non-engineering had higher satisfaction with the self-learning system than engineering students. Only Confidence of the ARCS in the proposed structural model was found to have significant effect on Self-assessed Skills, students with stronger confidence tending therefore to have higher self-rated English abilities.

Keywords Live DVD · ARCS motivation model

D.-F. Huang (✉)
Southern Taiwan University of Science and Technology, Tainan, Taiwan, Republic of China
e-mail: dfjhuang@mail.stust.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_115, © Springer Science+Business Media Dordrecht 2014

Introduction

Lack of learning motivation of students, especially those of vocational and technological universities, has long been claimed to cause ineffectiveness of English education in Taiwan. Teaching approaches geared towards initiating self-directed learning via technology-mediated support with a view to enhancing learning motivation have hence been gaining ascendancy in English learning settings across different educational levels in Taiwan. Multimedia provides variety and excitement to a computer-supported teaching and learning environment, adapting instruction to the diverse learning preferences of students [1]. Advanced multimedia instruction heightens visual aspects of communication, provides dynamic learning experiences and raises learning results [2]. Multimedia materials such as DVDs and VODs as effective self-directed learning aids enhances students' comprehension and memory, increases their motivation and promotes their concentration on the content in a near natural environment [3–5]. Best multimedia, however, will be rendered useless unless they are utilized by students to the utmost effect, which should be ensured through some mechanism. This consideration made Southern Taiwan University of Science and Technology create an online VOD Self-learning System, brand named Live DVD and the first of its kind in Taiwan, where students viewed designated films as a requirement. Students were monitored of their use behaviors and tested on their learning achievement after viewing the target films and using the system's embedded language learning features, including repeated listening, vocabulary collection, dictionary, target expression searching, multiple language display mode, etc. This study aimed to understand to what extent the Live DVD System affected learning motivation, which in turn would affect self-assessed English abilities, and whether this relationship depends on majors and English levels. Specifically, this study intended to test a proposed structural equation model involving Keller's ARCS motivation model [6] as a framework to investigate student perceptions of the Live DVD System and its effects on self-assessed English abilities in terms of whether the system draws their *attention*, shows *relevance* to their learning goals, helps build *confidence* in realistic expectations and learning outcomes, and makes the learning *satisfying* (see Fig. 1).

Method

Twelve classes of first-year non-English major undergraduate students participated in the study by completing the motivation survey. The students were placed at classes of three English levels: high, mid, and low for better learning of the required English class. As a class requirement, the students did English learning by viewing the films at the Live DVD Platform, and were invited to complete the questionnaire at the semester end, with a return of 527 questionnaires. The

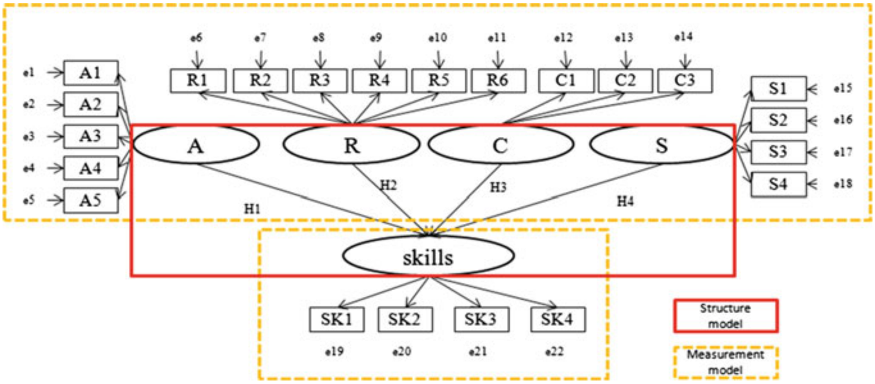


Fig. 1 The proposed structural equation model

questionnaire instrument on the Lickert scale was piloted on two classes of students and modified to insure a high reliability of the instrument for the formal data collection. After completion of data collection of 527 valid samples, confirmatory factor analysis, aside from reliability analysis, was conducted using SPSS v.18 and AMOS v.18 to check the convergent and discriminant validity of the instrument to validate the scale construct and the measurement model. Factorial ANOVA was then performed via SPSS v.18.0 to obtain the effects of Majors and English Levels on the ARCS motivation as well as the interaction effects. Structural equation modeling via AMOS was also conducted to test the hypotheses of the posited model featuring the structural model involving the effect of the ARCS motivation latent variables on Self-assessed Skills.

Result

Factor analysis employing principal component and Varimax extraction methods extracted five components including Attraction, Relevance, Confidence, Satisfaction, and Self-assessed Skills, cumutively accounting for ca. 72 % of total variation. A total of 22 question items were loaded on the 5 factors: Attraction (5 items), Relevance (6 items), Confidence (3 items), Satisfaction (4 items), and English Skills (4 items).

The reliability of the five factors turned out to be acceptably high, with Cronbach Alpha values of 0.91, 0.90, 0.74, 0.88, and 0.90 respectively for the preceding five factors. Confirmatory factor analysis supported the acceptable convergent validity of the scale [7–9]. In ARCS motivation model, the RMR was 0.039 (<0.05), the GFI, NFI, CFI being 0.904, 0.930, 0.947 respectively. In addition, the factor loading of each variable was significant and higher than 0.5, the CR of each dimension was over 0.7, and the AVE of each dimension was larger

Table 1 Two-way ANOVA for attention of ARCS model

Source	Type III sum of squares	df	Mean square	F	Sig.
Corrected model	10.362	5	2.072	3.014	0.011
Intercept	10,144.505	1	10,144.505	14,751.916	0.000
Major	0.511	1	0.511	0.744	0.389
English level	8.084	2	4.042	5.878	0.003**
Major * level	1.567	2	0.783	1.139	0.321
Error	358.278	521	0.688		
Total	10,571.360	527			
Corrected total	368.640	526			

than 0.5. As to the Self-assessed Skills measurement model, the RMR was 0.019 (<0.05), the GFI, NFI, CFI being 0.987, 0.988, 0.990 respectively. In addition, the variables' factor loadings were significant and higher than 0.5, the CR value over 0.7, and the AVE value larger than 0.5, indicating sufficient convergent validity and reliability of the model. The five components were also shown to have adequate discriminant validity complying with the rule that the correlation coefficient of two dimensions was less than the Cronbach's alpha reliability coefficients [10], and that the correlation coefficient of two dimensions was less than the square root of AVE [11].

The descriptive statistics of the ARCS model indicated a grand mean of 4.34, with separate means/SDs of 4.40/0.98, 4.29/0.84, 4.16/0.95, 4.49/0.94, respectively for each of the components, indicating a fairly positive motivation toward Live DVD self learning. Two-way ANOVA was performed to examine the effects of the variables of Group (of Majors) and Level (of English Ability) on the ARCS Motivation. As shown in Table 1, main effect of English Level ($F = 5.88$, $p < 0.01$) was found for Attention of ARCS Model. Post-hoc comparison showed Level A to be significantly lower in attention ($p < 0.05$) than Level B and Level C. Table 2 showed that main effect of English Level ($F = 7.96$, $p < 0.01$) and interaction effect ($F = 3.33$, $p < 0.05$) between Major and English Level were also found for Satisfcation of ARCS Model. İn Table 3, the simple main effect of Major showed a significant difference ($F = 10.346$, $p = 0.000$) among non-engineering majors, where both Level B and Level C significantly surpassed Level

Table 2 Two-way ANOVA for satisfaction of ARCS model

Source	Type III sum of squares	df	Mean square	F	Sig.
Corrected model	14.716	5	2.943	4.652	0.000
Intercept	10,562.019	1	10,562.019	16,691.946	0.000
Major	0.313	1	0.313	0.495	0.482
English level	10.075	2	5.037	7.961	0.000**
Major * level	4.216	2	2.108	3.332	0.036*
Error	329.669	521	0.633		
Total	10,968.938	527			
Corrected total	344.385	526			

Table 3 Test of simple main effect of major and english level on satisfaction of ARCS model

	Sum of squares	df	Mean square	F	Sig	Post hoc
Majors						
Engr	1.092	2	0.546	0.878	0.417	
Non-Engr	13.318	2	6.659	10.346	0.000**	Level B > Level A Level C > Level A
Levels						
A	1.156	1	1.156	1.531	0.218	
B	3.418	1	3.418	6.453	0.012*	Non-Engr > Engr
C	0.080	1	0.080	0.128	0.721	

A students in satisfaction. The simple main effect of English Level showed a significant difference (F = 6.453, $p < 0.05$) between non-engineering and engineering majors of Level B, the former significantly surpassing the latter in satisfaction.

Structural equation modelling was conducted to test the proposed structural model. The model fit analysis includes preliminary fit criteria, fit of internal structure of model and overall model fit [8]. The analysis of absolute fit measures (χ^2/d.f. = 2.282, GFI = 0.929, RMR = 0.041, RMSEA = 0.049), incremental fit measures (AGFI = 0.907, CFI = 0.97, NFI = 0.947) and parsimonious fit measures (PNFI = 0.791, PGFI = 0.709) suggested that the overall model fit of the proposal model after modest modification was good [12]. The regression coefficients of path analyses were used for testing the hypotheses of the proposed structural model, yielding the results as represented in Table 4 and Fig. 2.

Discussion

The participants' fairly positive attitude toward Live DVD System suggested that the online self-learning system is a worthwhile resource for students. However, Levels B and C tended to have higher attention to the system than Level A students, implying that while boosting the learning motivation of lower end students, the system needs to be modified in the designing of learning features to appeal to the interests and motivation of higher proficiency students. The

Table 4 Summary of hypotheses results

Hypothesized path		Standardized regression weights estimate	t-Value	Results
H1:	Attention → Eng skills	0.006	0.055	NH not rejected
H2:	Relevance → Eng skills	−0.098	−0.905	NH not rejected
H3:	Confidence → Eng skills	0.536***	6.045	NH rejected
H4:	Satisfaction → Eng skills	0.126	1.26	NH not rejected

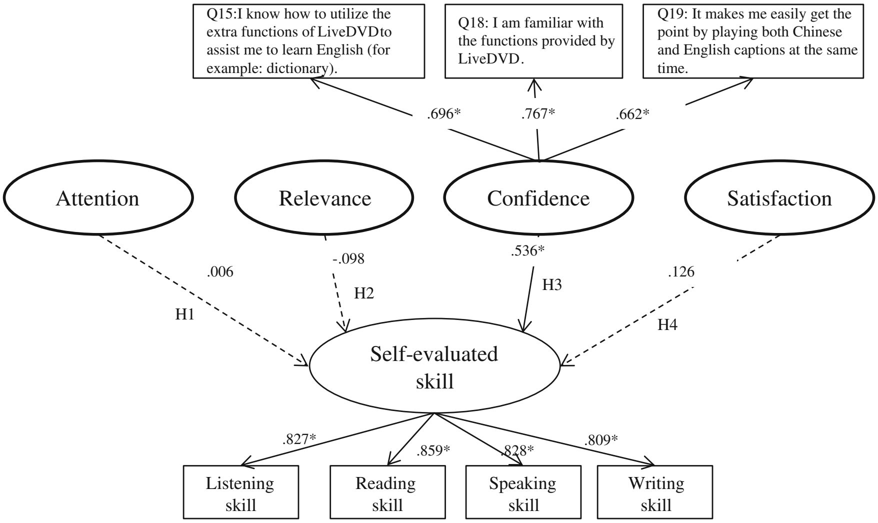

Q15:I know how to utilize the extra functions of LiveDVD to assist me to learn English (for example: dictionary).
Q18: I am familiar with the functions provided by LiveDVD.
Q19: It makes me easily get the point by playing both Chinese and English captions at the same time.
.696*
.767*
.662*
Attention
Relevance
Confidence
Satisfaction
.006
H1
-.098
H2
.536*
H3
.126
H4
Self-evaluated skill
.827*
.859*
.828*
.809*
Listening skill
Reading skill
Speaking skill
Writing skill

Fig. 2 The structural equation model with parameter estimates

precedence of Levels B and C over Level A in attention also raised the issue of the credibility of the placement test, which might have placed medium proficiency students into Group A, and reversely high proficiency students into Group B. The validity and reliability of the placement test for the incoming undergraduates needs therefore to be reviewed. Group A teachers are better advised to highlight the uses of the learning system and guide them to more advanced learning results. This implication seems to be particularly relevant for Non-engineering students given another result showing Levels B and C Non-engineering students to have higher satisfaction with the learning system than Level A counterparts.

In the ARCS model only Confidence was found to have direct effect on students' self-assessed skills, reflecting the important role of self-confidence in the learning process of college students, particularly of technological and vocational higher education, who tended to have lower learning motivation and English proficiency and hence lower self-confidence as well. Relative to the constructs of Attention, Relevance, and Satisfaction, Confidence, or students' perceived proficiency in using available learning features, linked to motivation most prominently in the EFL learning contexts. This research thus called attention to reinforcing of self-directed learning resources with user-friendly and practical features to build students' self-confidence which would directly affect their perceived English learning progress, and indirectly their learning motivation.

Conclusion

This study investigated the technological university students' perceptions of the Live DVD Self-learning System by testing a hypothetical structural equation model involving the ARCS measurement model and a structural model comprising the latent variables of ARCS and Self-assessed Skills. The results of the study suggested an English Level effect on Attention, and a Major effect as well as a Major*English Level interaction effect on Satisfaction, engendering implications for improvement of placement test validity and reliability and teaching practice of Level A classes. More important, Confidence was validated to be the sole construct of the ARCS motivations that affected student self-assessed skills, a result that highlights the critical need of constructing self-directed learning resources in a way that would facilitate student utilization and best build self-confidence. The results of this research are significant and applicable to technological and vocational higher education institutions in Taiwan with similar student backgrounds for purposes of English education reform and curriculum improvement.

References

1. Zaidel M, Luo XH (2010) Effectiveness of multimedia elements in computer supported instruction. J College Teach Learn 7(2):11–16
2. Wang L (2008) Developing and evaluating an interactive multimedia instructional tool. J Edu Multimedia Hypermedia 17(1):43–54
3. Astleitner H, Wiesner C (2004) An integrated model of multimedia learning and motivation. J Edu Multimedia and Hypermedia 13(1):3–22
4. Deimann M, Keller JM (2006) Volitional aspects of multimedia learning. J Edu Multimedia and Hypermedia 15(2):137
5. Guariento W, Morley J (2001) Text and task authenticity in the EFL classroom. ELT J 55(4):347–354
6. Keller JM (1987) Development and use of the ARCS model of motivational design. J Inst Dev 10(3):2–10
7. Anderson JC, Gerbing DW (1988) Structural equation modeling in practice. Psychol Bull 103:411–423
8. Bagozzi RP, Yi Y (1988) On the evaluation of structural equation models. J Acad Market Sci 16:74–94
9. Gefen D, Straub DW, Boudreau MC (2000) Structural equation modeling and regression. Commun Assoc Inf Syst 7(7):1–78
10. Gaski JF, Nevin JR (1985) The differential effects of exercised and unexercised power sources in a marketing channel. J Mark Res 22(5):130–142
11. Fornell C, Larcker D (1981) Structural equation models with unobservable variables and measurement error. J Mark Res 18(1):39–50
12. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2006) Multivariate data analysis, 6th edn. Prentice-Hall, New Jersey

Applying Firefly Synchronization Algorithm to Slot Synchronization

Yanliang Jin, Zhishu Bai, Lina Xu, Wei Ma, Xuqin Zhou and Muxin Wang

Abstract Firefly synchronization algorithm, in a bio-inspired way, has been proposed to replace conventional methods of solving the synchronization problem in wireless sensor networks. For the well-known M&S model and the RFA, a direct application of them can hardly lead to considerable results. This paper presents several vital improvements in order to make the RFA more applicable for a certain wireless sensor network. To better evaluate the results, the algorithm is modulated in a scenario with realistic parameters of a network. Moreover, differing network topologies have also been taken into consideration.

Keywords Firefly synchronization · RFA · WSNs

Introduction

In order to schedule a wireless sensor network in a sufficient way, a common notion of time slot should be the fundament. It is like the clock of a whole operating system that scheduling can only be achieved when every components work on a certain rhythm. To gain this goal, an accurate clock on wall would be necessary no more. Besides, dealing with a distributed network consists of sensor nodes, with limited capability of power and computation, conventional mechanism would be inappropriate and a nonhierarchical distributed algorithm may be preferable. Fireflies in South-East Asia, however, also distributed in trees with low intelligence, can flash in a perfect synchrony within a huge range from a chaotic situation. This striking phenomenon has been delved for a long run. Buck et al.

Y. Jin · Z. Bai (✉) · L. Xu · W. Ma · X. Zhou · M. Wang
Key Laboratory of Special Fiber Optics and Optical Access Networks of Ministry of Education, Shanghai University, Shanghai 200072, China
e-mail: baizhsh@126.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_116, © Springer Science+Business Media Dordrecht 2014

investigated fireflies' reaction to external flashings and studied their behaviors in more details.

Then, Mirollo and Strogatz thoroughly concluded a mathematical model for firefly synchronization based on pulse-coupled oscillators [1]. The rule governing this model is a quite simple one: (1) Each oscillator maintains a periodic time-phase function that acts as a build-in timer and when the phase counts to a certain threshold, the phase turns back to zero. (2) Once an oscillator receives a "flashing" from one of its neighbors, the function of whom updates its phase ϕ to a certain value ϕ' immediately. They have also proven that for an arbitrary number of entities and independent of the initial condition, the network shall always synchronize [2], so long as the constraints on the coupling between entities are met. The RFA is an algorithm based on the M&S model proposed in [3]. It differs slightly from the M&S model: the function is only adjusted once at the next beginning of cycle just after the firing instant. This difference, although small, makes the RFA more practical compared with the M&S model, for it alleviates the burden on computation ability of sensor nodes and neutralizes the impact of delays. Both the M&S model and the RFA serve as the way to force all sensor nodes in a network to achieve a rhythm, based on which scheduling can be implemented in order to gather or spread data in a well-organized way. However, direct application of the M&S model and the RFA in wireless communications can hardly achieve a satisfactory result. Either Topology impacts or realistic radio effects, such as propagation delays, channel attenuation and noise, place a huge effect on the performance of synchronization. What's more, in real wireless communication scenario, synchronization flash is composed of a sequence of pulses with several bytes or even more, which means that the reception and parse of synchronization words shall also bring in delays. Half-duplex transmission in the physical layer and random packet arrivals in the MAC layer also lead to a huge difference from experimental situations. As for the M&S model, lots of works have been done [4] to make it more applicable in WSNs.

This paper investigates how realistic radio effects, topology difference, half-duplex transmission and random packet arrivals affect the RFA. For practical applications of the RFA, several improvements were proposed in this paper. Simulation results show that with these improvements the RFA works well in a wireless sensor network. However, network topology variance still place a huge impact on the performance of the RFA. This paper is structured as follows: section Classic Mathematical Models of Firefly Synchronization reviews the M&S model and investigates two different situations of the RFA. Section Synchronization Metric and Blind Area gives the metric of synchronization and the principle of blind area. In section The RFA in Slot Synchronization, intertwining with CC1100, several details that may impact the RFA are pinpointed. Analysis of topology variance is given in section Topology Simulation and Analysis. Finally, section Conclusion and Future Work draws the conclusions.

Classic Mathematical Models of Firefly Synchronization

In this section, both the M&S model and the RFA are reviewed. Then, a divergence of the RFA is settled.

The M&S Model

According to the M&S model, pulse-coupled oscillators are governed by one simple rule:

1. Each oscillator maintains a periodic time-phase function that acts as a build-in timer. When the phase counts to a certain threshold, the phase turns back to zero.
2. Once an oscillator receives a "flashing" from one of its neighbors, the function of whom updates its phase ϕ to a certain value ϕ' immediately:

$$\phi' = f^{-1}(f(\phi) + \varepsilon) \tag{1}$$

where

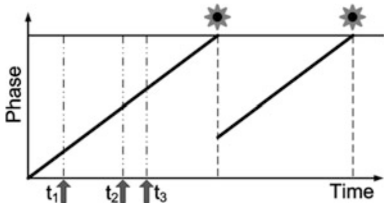
$$f(\phi) = \frac{1}{b}\ln\left(1 + [e^b - 1] \cdot \phi\right) \tag{2}$$

is the phase-state function and ε is the coupling strength. The phase-state function describes a curve that is smooth, monolithic increasing and concave down. The term b is the dissipation factor that decides the radian of the curve. Therefore, the later a flashing arrives, the larger the phase adjustment is.

The Reachback Firefly Algorithm

By the theory of the RFA [5], unlike the M&S model, oscillators use timestamps to record and queue up the firing messages from neighbors without responding in phase adjustment immediately. At the beginning of the next period, oscillators calculate the overall phase adjustment from the firing messages in queue and update the phase. The phase adjustment is also decided by phase-state function mentioned in section The M&S Model. Compared with the M&S model, the RFA avoids frequent phase adjustment to alleviate the computation burden. MAC-layer time stamping is also used in the RFA to avert two unpleasant conditions caused by propagation delays in WSNs: (1) The firing message of node A arrives just right at the firing instant of node B. (2) Unpredictable delays may cause disordered

Fig. 1 The RFA



firing message arrivals. Time stamping is a valid method that makes the RFA more applicable in a realistic situation (Fig. 1).

As for the RFA, there is still a cardinal divergence unsettled. For those messages arrive more close to the firing instant, it is more likely that the phase would jump to a value that exceeds the threshold according to function (1). How to deal with those irregular firing messages has not been settled yet. Moreover, how to determine the boundary has not been mentioned anywhere else so far as our best knowledge. Apparently, discarding all those irregular firing messages would be a better choice, for it at most prolongs the time to synchronization without causing instability. However, there still exists one particular situation. Suppose there are two sensor node clusters A and B, whose phases are almost the same and just meet the discarding condition. Consequently, the states of A and B are locked and shall never merge together until other nodes join into break this deadlock. To avoid this peculiar situation, nodes should keep the first firing message which leads to an over-jump and record the phase adjustment as $1 - \phi$. Then, the boundary is set at this phase, after which no more firing messages shall be recorded.

Synchronization Metric and Blind Area

Synchronization Metric

To evaluate the performance of the RFA, the Kuramoto synchronization metric [6] is used in this paper. Kuramoto metric is described as the following:

$$\begin{aligned} r \cdot \exp(j2\pi\bar{\phi}) &= \frac{1}{N}\sum_{n=1}^{N}\exp(j2\pi\phi_n) \\ \Rightarrow r &= \frac{1}{N}\cdot\exp(-j2\pi\overline{\phi})\cdot\sum_{n=1}^{N}\exp(j2\pi\phi_n) \end{aligned} \tag{3}$$

where r is the synchronization index, ϕ_n is the instant phase of oscillator n and $\bar{\phi}$ is the mean phase of all N oscillators. The synchronization index r equals to 1 when all nodes in a network are synchronized. When the system is in disorder, i.e. firing instants are randomly distributed within [0, T], the index r approaches 0.

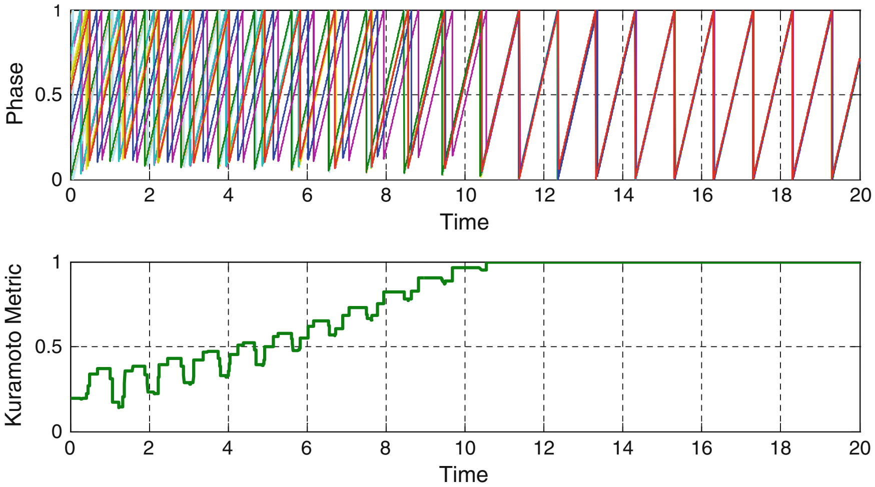


Fig. 2 Kuramoto Metric applied to the RFA

Figure 2 plots how the Kuramoto Metric evaluates the RFA, where 20 nodes are considered. At first, when the initial phases are randomly allocated, the Metric is pretty low. At about the 10th cycle, as the network achieves synchronization and finally stays steady, the Metric tends to 1.

Blind Area

Blind area is a fixed phase, below which nodes do not react to any firing messages received from neighbors. This specific phase is chosen as a value that is half of the phase threshold. The essence of blind area is similar with refractory period [7], proposed for the M&S model to avoid "echoes" caused by delays in networks. However, blind area still differs from refractory period for two aspects: (1) Blind area is a constant for a certain network. (2) Blind area is proposed to bring about a wait-and-chase situation in order to achieve the most efficient synchronization.

Refractory period T_{refr} is defined as the duration which satisfies that $T_{refr} \geq 2 \cdot v_{ij}$, where v_{ij} is the largest transmission delay between two nodes in a network. Apparently, for a wireless network consists of a number of nodes with multi-hop, finding out this delay is not an easy work. In addition, a refractory period is a wide range to choose from and it could vary a lot depending on the topology. Blind area is set as $\frac{1}{2}\Phi$th regardless of the topology which is easy to be located. Moreover, blind area always divides the nodes into two groups: a chasing group and a waiting group, where the chasing group always goes for the waiting group. $\frac{1}{2}\Phi$th is a peculiar value wiping off the possibility that two nodes chasing

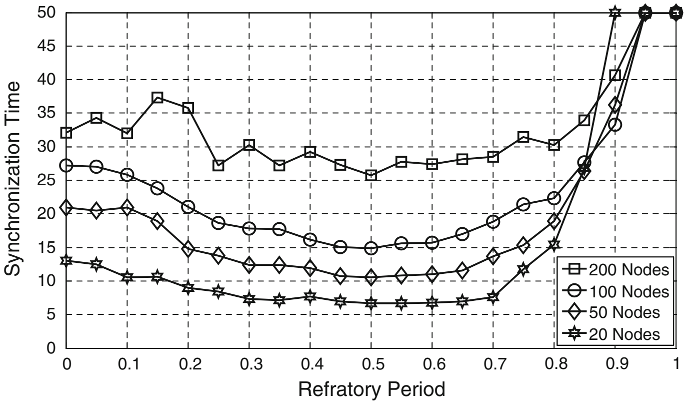


Fig. 3 Synchronization time with varying refractory period

for each other. Theoretically, blind area ensures both the efficiency and the stability of synchronization.

As for a network with certain node number, synchronization time varies with different refractory periods. Simulation result plotted in Fig. 3 shows that the shortest synchronization time can always be achieved when refractory period equals 0.5, which is a half of the phase threshold, for almost all situations. This has been explained above as the theory of blind area.

The RFA in Slot Synchronization

Figure 4 presents a wireless communication block diagram of realizing the RFA with chip CC1100. The diagram is divided into three parts: transmission, reception and arbitration. For every payload data $x(t)$ that is about to be transmitted shall be added with a header consists of several preamble bits and a sync-word $s(t)$ with 16 or 32 bits by CC1100. In the reception part, a sync-word is detected by CC1100 as the indicator of a valid payload data. The RFA is inserted as an arbitrator between the transmission part and the reception part to decide which part shall preempt the antenna. Node i can only transmit the data $y_i(t)$ when the result of its arbitration part is true. The RFA receives signals from the sync-word detector to queue up valid flashings and finally adjust the phase, which is elaborated in section Classic Mathematical Models of Firefly Synchronization. Once a sync-word is detected, the arbitration part should be informed no matter the whole data frame is successfully received. Because any two or more overlapping data frames could lead to a fail reception, namely collision which happens quite frequently when a network is in a chaotic state. From the transmitter side of node j, it takes T_{Tx} for a sync-word to be sent out completely, v_{ij} to go through the air to reach the receiver side of node i. Finally, sync-word detector takes T_{detec} to detect whether it is a valid

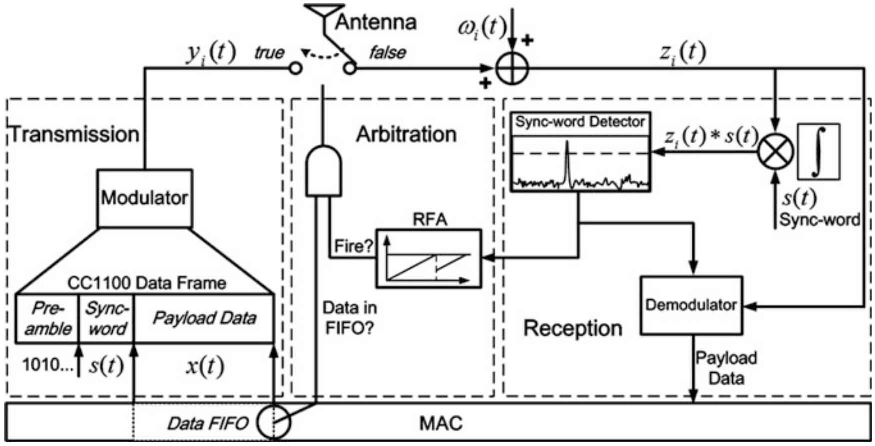


Fig. 4 Block diagram of wireless communication with the RFA applied as the arbitration part

sync-word or not. The total delay is written as $D = T_{Tx} + v_{ij} + T_{detec}$. Consider an additive white Gaussian noise $\omega_i(t)$ with zero-mean and variance σ^2, the incoming signal can be written as

$$z_i(t) = \sum_{j \in Z_i} A_{ij} \cdot y_j(t - \hat{D}) + \omega_i(t) \tag{4}$$

where Z_i is the set of nodes whose transmissions are observed by node i within its reception slot, $\hat{D}$ represents the delay that

$$\hat{D} = T_{Tx} + v_{ij} \tag{5}$$

and A_{ij} donates the amplitude

$$A_{ij} = \sqrt{d_{ij}^{-x} P} \tag{6}$$

in which d_{ij} is the distance between node i and node j, x is the path loss exponent and P is the transmit power. The incoming signal $z_i(t)$ directed to the reception part is scanned for the sync-word to decide the beginning of the data frame. The sync-word is detected by cross-correlating a known sync-word with $z_i(t)$:

$$\begin{aligned} z_i(t) * s_i(t) &= \int z_i(t - \tau) s_i(\tau) d\tau \\ &= \int \left[\sum_{j \in Z_i} A_{ij} \cdot y_j(t - \hat{D} - \tau) + \omega_i(t - \tau) \right] s_i(\tau) d\tau \\ &= \sum_{j \in Z_i} A_{ij} \int y_j(t - \hat{D} - \tau) s_i(\tau) d\tau + \omega_i(t) * s_i(t). \end{aligned} \tag{7}$$

Only a sync-word that is fully covered in the reception slot of node i can generate a wave peak with magnitude $A_{ij}S$, in which S is the length of sync-word. This wave peak is decided by the sync-word detector that whether it is a valid one by comparing the peak with a given threshold. Because of the existence of noise in the channel, false alarm and/or omission of a sync-word could happen. The likelihood of a successful detection is decided by both the threshold set by CC1100 and the length of sync-word.

Comparing with the ideal synchronization model, several cardinal differences should be taken into account:

1. The RFA alone cannot decide whether or when shall a data frame be transmitted. As a matter of the MAC strategy, nodes may not always transmit a data frame at their firing instant.
2. Multiple delays exist in the entire process of transmission and reception as the sync-word is not just an ideal pulse without length and processing delay.
3. To a certain node, at one time, the reaction to one node or a cluster of synchronized nodes is the same, for the sync-word detector can only detect a wave crest at one time no matter high or low.
4. Half-duplex mode makes the utilization of the RFA even more complicated, because nodes are "deaf" while transmitting and nodes which are receiving can only be scheduled to transmit mode after a whole data frame is completely received.

Apparently, although the MAC strategy macroscopically decrease the amount of firing messages a node could receive from its neighbor in one listen slot, it prolongs the time to synchronization without affecting the accuracy. Delays, however, would drastically compromise the attainable slot synchronization accuracy. The maximum transmission distance of CC1100 is about 150 meter for an indoor scenario. Then $v_{ij} \leq \frac{d}{c} = \frac{150}{3 \cdot 10^8} = 0.5\,\mu s$, where d is the distance between two nodes and c is the speed of light. The maximum transmission data rate of CC1100 is 500 kbps, when preamble is set as 8 bits and sync-word is set as 32 bits, then $T_{Tx} \geq \frac{32+8}{500 \cdot 10^3} = 80\,\mu s$, which is far greater than v_{ij}. Consequently, $D \approx T_{Tx} + T_{detec}$. For a certain hardware platform, D is a constant that can be easily obtained.

Figure 5 plots a general process of how the RFA works in a sensor node A. In the listen state, it takes $D \approx T_{Tx} + T_D$ for node A to completely receive and parse the sync-word ever since one of its neighbors has reached the firing instant. Once node A detects a valid sync-word for node B, it could calculate the right time at which node B fired by subtracting the constant delay D. For nodes which fires at $t1$, $t2$ and $t3$, node A only react to the node firing at $t3$ for the firing messages of the other two nodes are blocked by blind area. At the firing instant of node A, only when there is payload data in the MAC layer can it turn to the transmit mode to emit the data frame. It is should be mentioned that there may exist nodes which fire at the previous cycle of node A, but their sync-words could not be detected until the next cycle. This specific situation is like the node C which fires at $t4$ in Fig. 5.

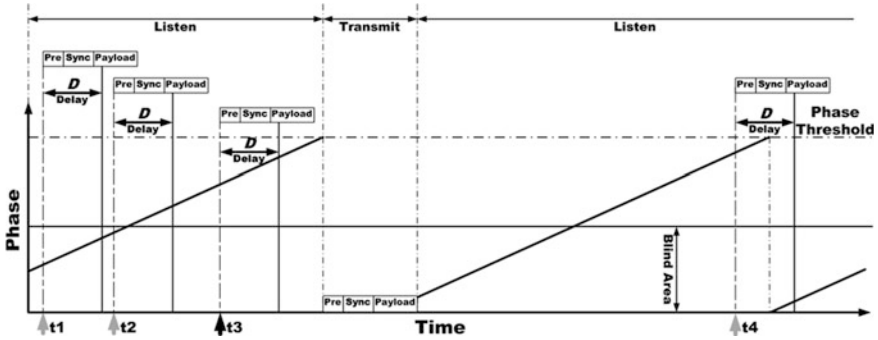


Fig. 5 Listen and transmit slot with the RFA applied

Even though the right firing instant of node C can be obtained, it is too late to add it to the previous queue. Given this kind of situation, the firing message at *t*4 should not be discarded and shall be added into the next queue to be calculated. In the diagram that Fig. 4 plots, the randomly scheduled data transmission at the firing instant could vastly compromise the synchronization rhythm, for the sensor nodes can only work in half-duplex mode. In order to regain the stability of synchronization, a vital modification is proposed to the diagram to overcome the drawback of half-duplex, shown in Fig. 6.

Now that nodes do not react to any flashings received within a blind area slot, the slot could be assigned to transmit data frames. During the transmission period, the time-phase function is still maintained and synchronization rhythm is neither violated. To make sure that the data frame could be sent out completely within the blind area T_{BA}, it should fulfill:

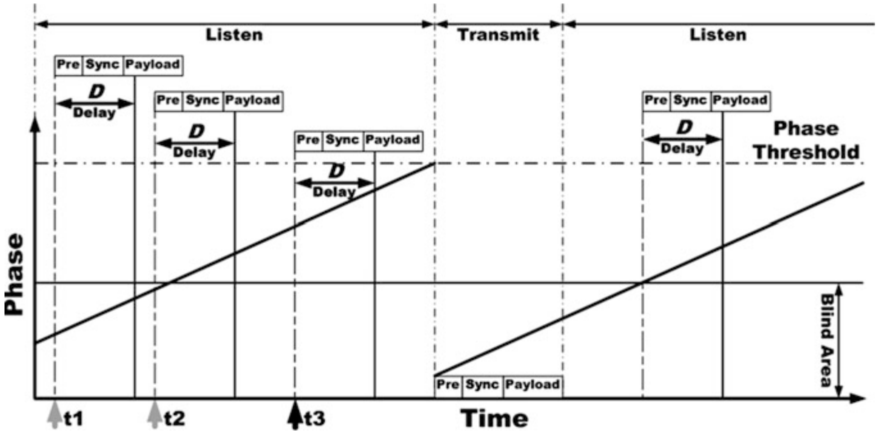


Fig. 6 Modified listen and transmit slot

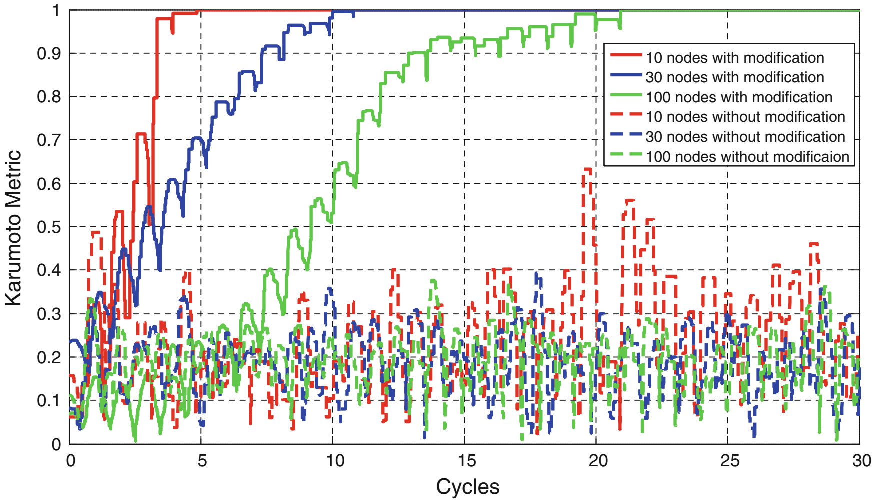


Fig. 7 The performances of the RFA in modified slot and unmodified slot with varying node number

$$T_{BA} > (N - 1) \cdot \varepsilon + T_{DF} \tag{8}$$

where T_{DF} is the time it takes to completely transmit a data frame and the term N is number of sensor nodes in a network. Figure 7 shows the simulation results of the RFA applied in slot synchronization within a fully-connected network. Simulation parameters are given in Table 1. It is clear that the RFA works well with the modification but the dash lines show that networks can hardly get synchronized with the effect of half-duplex mode. Compared Fig. 3, the synchronization time is apparently prolonged, for the influence from MAC strategy.

Topology Simulation and Analysis

In this section, the performance of the RFA in slot synchronization, elaborated in section The RFA in Slot Synchronization, is analyzed when different topologies

Table 1 Slot synchronization simulation parameters

Data frame length	25 bytes
Sync-word length	32 bits
Preamble length	32 bits
Cycle duration	1 s
Total delay D	0.66 ms
Baud rate	400 Kbps
Repeat	100 times

Table 2 Topology simulation parameters

Node number	30
Coupling strength	0.00239
Dissipation factor	10
Data frame length	25 bytes
Sync-word length	32 bits
Preamble length	16 bits
Cycle duration	1 s
Total delay D	0.12 ms
Baud rate	400 Kbps
Repeat	100 times

are applied. Similar works have been done based on the M&S model. However, a crucial discrepancy between this section and the result in [8] is found and a sound explanation is drawn finally.

In this paper, a concept named connection rate is utilized to characterize topological properties of a network. Given a certain graphic:

$$G = G(V, E).$$

Connection rate L_G is defined with the number of existing links E and the number of nodes N in a network:

$$L_G = \frac{2|E|}{|V| \cdot (|V| - 1)} \tag{9}$$

where $|E|$ and $|V|$ is the number of edges and nodes respectively in the graphic. Connection rate is a simple way to characterize the density of a network. Moreover, with a given L_G and certain number of nodes, it is easy to generate and handle a network. Especially, $L_G = 1$ means a fully-connected network.

The simulation is evaluated with MATLAB accounting for topologies like line, ring, star and mesh with varying connection rate. Simulation parameters are given in Table 2 and the result is shown in Fig. 8. Mesh-n means this topology is generated with a connection rate that equals n and Mesh-fully means the connection rate is 1.

Mesh Topology. For mesh topology, three different connection rates are taken into consideration. It is easy to tell from Fig. 8 that the synchronization time is inversely proportional to the connection rate, which is reasonable. A higher connection rate means that a node could have a larger neighbor set and phase adjustment could happen more frequently in one cycle. In mesh topology, therefore, the performance of the RFA, applied to slot synchronization, is mainly affected by the connection rate. However, the same connection rate does not lead to the same topology shape, and there still exist some special topology shapes that worth the attention, such as star, line and ring.

Star Topology. Star topology, in which every node is connected to a center node, may lead to the worst performance when the RFA is applied, shown in

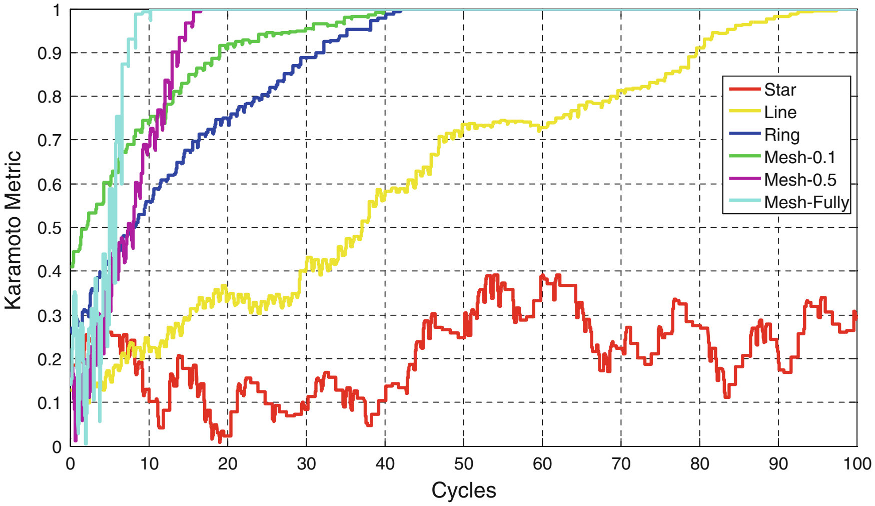


Fig. 8 Simulation results for different topologies

Fig. 8. The reason that causes such a bad condition is that the center node is always dragged around by its neighbors and its neighbors cannot communicate each other. This situation, however, is not permanent. After a number of experiments, it is interesting to find out that a network with star topology could be synchronized when the number of nodes in it decreases below about 10. The conclusion is just opposite to [9], in which star topology leads to the best performance.

Line Topology. In a line topology, nodes are connected one by one but the head and tail are not connected. Except for the head and tail, each node has only two neighbors. Clearly, this kind of shape does no good to synchronization. A longer line makes it harder to achieve an overall rhythm. Given a certain connection rate, when the number of nodes if greater than 3, a mesh topology always leads to a better performance than what a line topology does.

Ring Topology. By connecting the head and the tail of a line topology a ring topology is generated. Although the connection rate increases a little, as shown in Fig. 8, the performance is improved vastly comparing with the line topology.

Conclusion and Future Work

In this paper the RFA has been modified to be more feasible in wireless communication with realistic radio effects from delay, half-duplex mode and so on. Simulation results show that blind area works well in slot synchronization. Besides, half-duplex mode effect is neutralized by inserting the transmit mode in the blind area. Different Topologies have also been taken into account. For certain connection rates, a topology that is randomly generated is always more suitable than those special ones, such as line, ring and star especially. In future work, it will be interesting to test the theory of this paper on a realistic test bed. Other factors that may affect synchronization accuracy shall also be studied.

Acknowledgments The authors would like to thank the reviewers for their detailed comments on earlier versions of this paper. This work was supported by Shanghai leading academic discipline project (S30108, 08DZ2231100), funding of key laboratory of wireless sensor network and communication, shanghai institute of micro system and information technology, Chinese academy of sciences (2012002) and Shanghai science committee (12511503303).

References

1. Mirollo R, Strogatz S (1990) Synchronization of pulse-coupled biological oscillators. SIAM 50(6):1645–1662
2. Leidenfrost R, Elmenreich W, Bettstetter C (2010) Fault-tolerant averaging for self-organizing synchronization in wireless ad hoc networks. Proceedings of the international symposium on wireless communication systems (ISWCS), New York
3. Wieser S, Montessoro PL, Loghi M (2013) Firefly-inspired synchronization of sensor networks with variable period lengths. In: Adaptive and natural computing algorithms, pp 376–385, Springer, Berlin Heidelberg
4. Tyrell A, Auer G, Bettstetter C (2007) Biologically inspired synchronization for wireless networks. In: Dressler F, Carreras I (eds) Studies in computational intelligence, 69:47–62, Springer, Berlin, Germany
5. Tyrrell A, Auer G, Bettstetter C (2010) Emergent slot synchronization in wireless networks. IEEE Trans Mob Comput 9:719–732
6. Kuramoto Y (1984) Chemical oscillations, waves, and turbulence. Springer, Verlag
7. Mathar R, Mattfleldt J (1996) Pulsed-coupled decentral synchronization. SIAM J on applied math 56(4):1094–1106
8. Leidentfrost R, Elmenreich W (2009) Firefly clock synchronization in an 802.15.4 wireless network. EURASIP J Embed Syst, 17
9. Moreno Y, Pacheco AF (2004) Synchronization of Kuramoto oscillators in scale-free networks

Novel Mutual Information Analysis of Attentive Motion Entropy Algorithm for Sports Video Summarization

Bo-Wei Chen, Karunanithi Bharanitharan, Jia-Ching Wang, Zhounghua Fu and Jhing-Fa Wang

Abstract This study presents a novel summarization method, which utilizes attentive motion analysis, mutual information, and segmental spectro-temporal subtraction, for generating sports video abstracts. The proposed attentive motion entropy and mutual information are both based on an attentive model. To capture and detect significant segments among a video, this work uses color contrast, intensity contrast, and orientation contrast of frames to calculate saliency maps. Regional histograms of oriented gradients based on human shapes are also adopted at the preliminary stage. In the next step, a new algorithm based on mutual information is proposed to improve the smoothness problem when the system selects the boundaries of motion segments. Meanwhile, differential salient motions and oriented gradients are merged to mutual information analysis, subsequently generating an attentive curve. Furthermore, to remove non-motion boundaries, a smoothing technique based on segmental spectro-temporal subtraction is also used for selecting favorable event boundaries. The experiment results show that our proposed algorithm can detect highlights effectively and generate smooth playable clips. Compared with existing systems, the precision and recall rates of our system outperform their results by 8.6 and 11.1 %, respectively. Besides, smoothness is enhanced by 0.7 on average, which also verified feasibility of our system.

B.-W. Chen (✉) · J.-F. Wang
Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China
e-mail: dennisbwc@gmail.com

K. Bharanitharan
Department of Electrical Engineering, Feng Chia University, Taichung, Taiwan, Republic of China

J.-C. Wang
Department of Computer Science and Information Engineering, National Central University, Jhongli, Taiwan, Republic of China

Z. Fu
School of Computer Science, Northwestern Polytechnical University, Xi'an, China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_117, © Springer Science+Business Media Dordrecht 2014

Keywords Video summarization · Attentive motion entropy · Mutual information analysis · Segmental spectro-temporal subtraction

Introduction

Video summarization techniques have been proposed for years to offer people comprehensive understanding of the whole story in a video. To date, a great deal of effort has been devoted to providing people with more friendly interfaces and better concept interpretation [1–13]. Traditional video summarization approaches can be roughly classified into two categories: One is the static storyboard [1, 10, 13], which is composed of still images extracted from the original video; the other is the dynamic skimming [2–9, 11, 12], which concatenates several shorter clips. Both of them aim to offer users a compact view of a video. This work mainly focuses on the study of dynamic skimming approaches because generating playable clips is suitable for users to navigate sports videos. So far, a large amount of related research has been proposed to analyze sports videos, including colors, motion vectors [6, 7, 12, 14], saliency maps [15, 16], information theory-based features [8, 11], and so forth.

Although these systems are capable of capturing salient shots among videos, it is still difficult to detect genuine events and remove false alarms. For example, Walther et al. [15, 16] attempted to highlight objects in an image by using low-level visual features and generating saliency maps. Like many other color-based approaches, their method required heuristic rules for filtering unnecessary frames. On the other hand, [7] focused on another features and balanced this problem via estimating motion vectors. Cernekova et al. [8] adopted an information theory-based model and estimated motion events by detecting mutual information between frames. Nevertheless, such methods were susceptible to global motion changes, such as camera panning and zooming, which may cause false positive results. Duan et al. [3] then developed an enhanced video skimming system by extracting video objects and converting them into motion vector fields. Although precision rates increased, Duan's approach needed to map quite many visual features to the semantic level. A more sophistic classifier was required in the training phase.

To solve such problems, a compromised way, which combines merits of the above algorithms for summarizing sports video events, is proposed in this study. Therefore, this work provides:

1. A novel approach based on attentive motion entropy analysis, which joins salient motions and regional histograms of oriented gradients;
2. A new mutual information estimator that calculates differential salient motions and differential oriented gradients for detecting coherent frames and the boundaries of a motion event;

3. A boundary-smoothing scheme based on segmental spectro-temporal subtraction for removing obscure motion segments.

With the use of the proposed techniques, the system can decrease false positives caused by conventional motion analyses and create a friendly navigation interface for users.

The rest of this paper is organized as follows. Section Related Work introduces the related approaches about dynamic video skimming. Section Proposed Video Summarization System then describes details of the proposed video summarization method. Next, section Experimental Results summarizes the performance of the proposed method and the analysis results. Conclusions are finally drawn in section Conclusion.

Related Work

To date, a great amount of research has devoted to generating dynamic video abstracts, using various approaches. In [12], the authors employed two types of models, "global motion model" and "local motion model," to distinguish camera events from sports events. Ma et al. [6, 17] made use of different features, such as motions, color, and other media descriptors, and fuse them into a curve. Such curves are often referred to as attentive curves; they can offer developers numeric conversion of images. In 2004, Cernekova et al. [8] adopted another approach, modeling shot and scene changes by detecting mutual information between frames. There were also some approaches [5, 6] that exploited a more sophisticated way to extract highlights. For example, the authors [6] utilized content-based parsing techniques to detect face regions. However, the major drawback is that it requires a recognition system and involves a training phase. In addition to the aforementioned methods, heuristic rules are often used as a criterion for identifying event patterns. In [5], the authors presented several feature extraction methods, including wavelet-based motion-trend analyses, hybrid field-color models, and prior knowledge-driven line detection. Liu et al. [2] has devised a perceived motion model for key-frame extraction. They came up with a triangular modeling rule to reshape the generated attentive curve. Duan et al. [3] developed a semantic video skimming system by extracting video objects and converting them into motion vector fields. Instead of proposing visual features as the aforementioned research did, Lu et al. [18] made use of a transition graph model to calculate skimming length of shots. Like Lu's approach, Ngo et al. [19] modeled a temporal graph for scene classification using N-Cut algorithm. Although video abstracts can be presented to users in forms of scenes and shots, both of Lu's and Ngo's approaches required scene and shot detection, which are inapplicable in sports videos due to obscure scene changes.

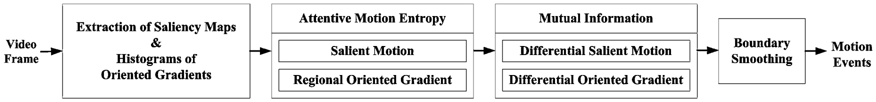

Video Frame
Extraction of Saliency Maps & Histograms of Oriented Gradients
Attentive Motion Entropy
Salient Motion
Regional Oriented Gradient
Mutual Information
Differential Salient Motion
Differential Oriented Gradient
Boundary Smoothing
Motion Events

Fig. 1 System Overview

Generally speaking, finding an effective way to transform video data into a numeric curve is the major challenge in dynamic skimming. Accordingly, this study aims to provide another feasible solution to this issue.

Proposed Video Summarization System

The proposed system consists of four stages, as shown in Fig. 1. The first stage of the workflow is the saliency map extraction along with computation of histograms of oriented gradients. The second calculates attentive motion entropy according to salient motions and oriented gradients; potential motion events are preliminarily marked in this step. Then, the third stage calculates mutual information between frames using differential salient motions and differential oriented gradients; the objective is to find out the boundaries of a motion event. Finally, the last stage is to detect sports activities from mutual information curves and refine the results using boundary smoothing. The outputs of the system are concatenated playable clips.

Extraction of Saliency Maps and Histograms of Oriented Gradients

The following description gives a short introduction to saliency map extraction. A saliency map is an image that can point out the visual regions, which have the most perceptual impact on human brains. For these years, this technique has been intensively and broadly studied, and it has a wide range of applications [6, 15, 16]. The idea behind the algorithm is to provide visual perception analysis by using color contrast, intensity contrast, and orientation contrast in an image. In this study, the main reason to adopt such techniques in our system is that unnecessary global motions could be removed by employing saliency map extraction. Besides, such techniques can avoid detecting false highlights in videos.

The following are the key-points of the saliency map according to the authors' research [15]: After the system extracts each frame, the image is sub sampled into a Gaussian pyramid and then decomposed into several channels for red (R), green (G), blue (B), yellow (Y), intensity (I), and local orientation (O_θ). From these channels, two feature maps, "center c" and "surround s" are constructed and normalized. A saliency map can be obtained by averaging the functions proposed in [15] at this stage.

To capture moving athletes, this study follows [20], which extracted intensity-based and gradient-based features, to model human shapes. An Adaboost classifier is also used to detect humans in video frames. After the classifier finds human contours or silhouettes, a regional histogram of oriented gradients (HOGs) based on the labeled areas is created. Let $f(x,y)$ represent the pixel of coordinate x and y, ∇ denote gradients, W refer to the weight of a coordinate, and ϕ be an edge direction. The histogram of oriented gradients can be expressed as follows.

$$\begin{aligned}\nabla_{x,y}^{\text{Horizon}} = {} & 2f(x+1,y) - 2f(x-1,y) + f(x+1,y+1) \\ & - f(x-1,y+1) + f(x+1,y-1) - f(x-1,y-1)\end{aligned} \tag{1}$$

$$\begin{aligned}\nabla_{x,y}^{\text{Vertical}} = {} & 2f(x,y+1) - 2f(x,y-1) + f(x+1,y+1) \\ & - f(x+1,y-1) + f(x-1,y+1) - f(x-1,y-1)\end{aligned} \tag{2}$$

$$W_{x,y} = \left(\left(\nabla_{x,y}^{\text{Horizon}}\right)^2 + \left(\nabla_{x,y}^{\text{Vertical}}\right)^2\right)^{1/2} \tag{3}$$

$$\phi_{x,y} = \tan^{-1}\left(\nabla_{x,y}^{\text{Horizon}} \Big/ \nabla_{x,y}^{\text{Vertical}}\right). \tag{4}$$

When gradients are computed, a histogram of edge directions is then created to record the number of pixels that belongs a direction.

Attentive Motion Entropy

After extracting saliency maps and regional gradients of video frames, this step uses those data to calculate motion entropy. As mentioned in the previous section, although dominant motion features were widely used for key-frame identification in sports videos, we observed that such techniques were subject to global camera motion effects. Therefore, motion entropy based on saliency maps and regional gradients is proposed herein to alleviate those problems in video skimming. Motion entropy [11] has been proved effective in rejecting false alarms and efficient computation. The major difference between motion entropy and attentive motion entropy is that the latter is operated in the attentive regions of frames, including saliency maps and regional gradients. Motion entropy of saliency maps is given as follows:

$$H_{\text{SaliencyMap}} = \sum_{k} w_k \times h_k \tag{5}$$

where k is the index of the sector after the polar axis is partitioned into equal segments; h_k (motion directivity entropy) is the total entropy that belongs to the corresponding sector, and w_k is the total weights in the sector. Let p_k represent the proportion of motion vectors, whose angles fall into the kth sector, and h_k can be defined as

$$h_k = -p_k \times \log p_k. \tag{6}$$

It describes the activity rate of a frame when it is converted to a saliency map; the strength of directivity entropy is calculated by

$$w_k = \sum_i r_{i,k} \Big/ \sum_i r_i \tag{7}$$

where $\sum_i r_{i,k}$ is the sum of the lengths of the motion vectors in the kth region, and the denominator means the sum of the total lengths.

Based on the same concept in [11], motion entropy of regional gradients can be obtained by using histograms of oriented gradients. Thus, attentive motion entropy can be expressed by

$$H_{\text{Attentive}} = H_{\text{SaliencyMap}} \times H_{\text{RegionalHOG}} \tag{8}$$

where $H_{\text{SaliencyMap}}$ calculates inter frame saliency, and $H_{\text{RegionalHOG}}$ represents intra frame saliency. The following figures compare the proposed attentive motion entropy and the other approaches.

Mutual Information Based on Differential Salient Motions and Oriented Gradients

Once a highlighted frame is detected in a video, the system then determines the boundaries of an event. Thus, playback would become smooth when users watch it. A simple way to estimate the boundaries of an event involves using color information between frames, such as [8]. The authors presented a novel concept of applying mutual information to detecting similar images. Although it may work in movies, some problems still arise in analyzing sports videos. This is because color features in different scenes among a sports video slightly change. Therefore, this work modifies the original equation to compute motion data.

$$\begin{aligned} MI_{t,t+1}^{\text{SaliencyMap}} &= MI_{t,t+1}^{\text{Active}} - MI_{t,t+1}^{\text{Inactive}} \\ &= \mathbb{C}_{t,t+1} \times \log \frac{\mathbb{C}_{t,t+1}}{\mathbb{C}_t^{\text{Active}} \mathbb{C}_{t+1}^{\text{Active}}} - \mathbb{C}'_{t,t+1} \times \log \frac{\mathbb{C}'_{t,t+1}}{\mathbb{C}_t^{\text{Inactive}} \mathbb{C}_{t+1}^{\text{Inactive}}} \end{aligned} \tag{9}$$

where $\mathbb{C}_{t,t+1}$ denotes the proportion of the motion vectors, which remain as active (directional) states from one frame to the next, whereas $\mathbb{C}'_{t,t+1}$ represents inactive ones; $\mathbb{C}_t^{\text{Active}}$ or $\mathbb{C}_t^{\text{Inactive}}$ calculates the percentage of the directional or unidirectional vectors in a single frame, respectively. When multiplied with $MI_{t,t+1}^{\text{RegionalHOG}}$

which calculates mutual information of differential oriented gradients (i.e., $MI_{t,t+1}^{\text{Active}} - MI_{t,t+1}^{\text{Inactive}}$), [21] becomes

$$MI_{t,t+1}^{\text{Saliency}} = MI_{t,t+1}^{\text{SaliencyMap}} \times MI_{t,t+1}^{\text{RegionalHOG}}. \tag{10}$$

Equation 10 measures the coherent status between consecutive frames. The higher mutual information is, the more coherent two frames are. With the use of 10, the system can find the start and the end of a motional event efficiently because a sports event usually comes along with significant motion activity changes.

Boundary Smoothing Based on Segmental Spectro-Temporal Subtraction

After mutual information is computed, the browsing system needs to output meaningful parts of a video by choosing "salient peaks" from the curve. A simple way is to set up a predefined threshold, and then the system can determine sports events based on the value. Nevertheless, many local minimums appear in a segment, and they may cause the system to misclassify false boundaries. Consequently, removing those local minimums becomes an important issue.

Activity detection technology has been widely used in signal processing. It involves detecting silence segments and recognizing whether an audio or visual signal of interest begins or not. Zero-crossing rates [22] are considered as the simplest way to separate silence from an unknown signal. However, in our case, the value of mutual information is always larger than zero, which makes zero-crossing rates inapplicable. Other research, such as spectral entropy [23] and spectral flatness [24], also offers the same functionality

In this work, segmental spectro-temporal subtraction, which is derived from spectral subtraction [21], is proposed at this stage. Spectral subtraction was originally developed to suppress noise from voice signals in the frequency domain. Let X represent the noiseless mutual information and N denote noise signals in a video. The spectro-temporal representation of MI, X, and N can be given as follows.

$$MI(f,t) = X(f,t) + N(f,t) \tag{11}$$

Taking power spectrum on both sides yields

$$|MI(f,t)|^2 = |X(f,t)|^2 + |N(f,t)|^2. \tag{12}$$

The above function can be rewritten in terms of segments, where η is the index of segments. The estimated X is calculated by subtracting averaged local minimum of noise near η.

$$\left|\hat{X}_{\lambda}(f,t)\right|^{2}=\left|MI_{\lambda}(f,t)\right|^{2}-\left|N_{\lambda}(f,t)\right|^{2}=\left|MI_{\lambda}(f,t)\right|^{2}-E\left(\sum_{\tau}\left|N_{\lambda-\tau}(f,t)\right|^{2}\right). \tag{13}$$

Experimental Results

To assess the performance of our system, six soccer videos (videos 1–4) were used in the test. These video dataset contained 152 sports highlights, including corner kicks, goalkeeper shots, block tackles, goal-and-cheer shots, etc., which were manually labeled and segmented by humans. All of these sports events were motion events. Besides, unnecessary segments, such as commercial parts, anchor shots, audience scenes, and captions, were removed from the videos. Motion events in the same video were concatenated together, each of which is separated with at least one non motion event. To evaluate whether the proposed system can distinguish motion events from non highlights in videos, average precision and recall rates are used as the criteria in the experiment.

Performance test between the parameters of the proposed method is listed in Table 1. Table 1 compares average precision and recall rates of the parameters, where the first column of the table is the parameter number, the second denotes the parameter type, and the rest columns list the test results. Closely examining precision and recall rates revealed that motion entropy-based parameters (parameters 1 and 2) could detect more accurate motion events, compared with mutual information-based ones (parameters 3 and 4). However, the recall rates of mutual information-based parameters rose to more than 75 %, owing to selecting longer segments. Clearly, when combined with parameters 1–4, the performance could achieve as high a precision rate as 82.1 %. Also, the recall rate was as high as 83.7 %. Such experimental results indicate that integration of the two types of parameters can offer higher discriminability than that of individual one.

Table 2 lists the experimental results of different approaches. As shown in the table, our results outperformed the other methods. The precision and recall scores were respectively increased to 85.9 and 85.5 % on average. In comparison with the other baselines, the precision difference between our methods and the other

Table 1 Performance test between the parameters of the proposed method

Number	Parameter	Precision (%)	Recall (%)
1	Motion entropy based on saliency maps	80.3	72.2
2	Motion entropy based on histograms of oriented gradients	80.0	71.3
3	Mutual information based on differential salient motions	72.0	80.5
4	Mutual information based on differential oriented gradients	70.8	75.5
5	Motion entropy (1 and 2) +mutual information (3 and 4)	82.1	83.7

Table 2 Evaluation results of different approaches

Method	Precision (%)	Recall (%)
Saliency map [16]	69.5	62.2
Motion activity [7]	78.0	74.9
Motion entropy [11]	81.9	78.5
Color-based mutual information [8]	79.6	81.7
Proposed method without boundary smoothing	82.1	83.7
Proposed method with boundary smoothing	85.5	86.1

Table 3 Smoothness test. After watching video summaries, each user can give a score (integer), which ranges from 1 (worse) to 3 (better), to each system

Video number	Motion activity	Motion entropy	Color-based mutual information	Proposed
I	1.2	1.7	2.0	2.1
II	1.4	2.2	2.0	2.0
III	1.4	1.6	2.1	2.8
IV	1.0	2.1	1.8	2.6
V	1.4	2.2	2.0	2.2
VI	1.4	1.5	1.1	2.5
Average	1.3	1.9	1.8	2.4

approaches was enhanced by 8.6 %. The recall difference of our results also found an average increase of 11.1 %. Notably, when the proposed boundary smoothing technique was added to the system, the precision and recall rates can reach 86.5 and 86.1 % because inappropriate motions were filtered. In contrast to the other methods, the proposed one has demonstrated its efficacy of detecting highlights.

In order to benchmark smoothness of selected segments, a common subjective measurement, mean opinion scores (MOS), was employed in the test. Total ten persons were invited to the test, and each of them was asked to give a score ranging from 1 (worse) to 3 (better). The evaluation results are shown in Table 3. We observed that the average scores of our system were above 2.0, whereas those of the baselines were smaller than 2.0. Due to lack of mutual information for event boundary checking, the average score of Yeo and Liu's method reached only 1.3. In contrast to Yeo and Liu's system, the MOS score of Cernekova's approach, which focuses on only color mutual information, was increased to 1.8 although Cernekova's approach could determine motion boundaries. The score of the proposed method could reach as high as 2.4, which was larger than that of motion entropy system by 0.6. Such results imply that our algorithm has better ability to generate smooth segments than the baselines.

The following figures list the snapshots of the testing results for demonstration. Notably, for clarity of presentation, the total testing frames are 39 (13 different keyframes), and each snapshot is manually extracted. As shown in Fig. 2a, some undesired frames (all the bottom snapshots) are selected due to their color saliency. Like the result of Cernekova et al., although motion entropy (Fig. 2b) can remove

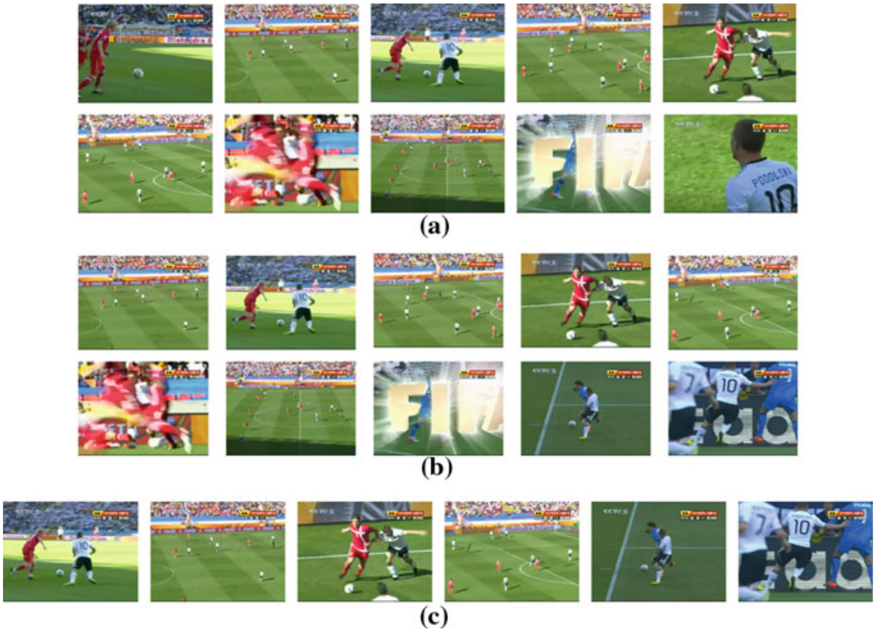


Fig. 2 Snapshots of the video summaries, where the time order of images is from the top-left to the bottom-right. **a** Color-based mutual information [8]; **b** Motion entropy [11]; **c** Proposed system. The duration of sample video is around one minute. Notably, for clarity of presentation, the total testing frames are 39 (13 different key frames), and each snapshot is manually extracted

frames that contain slight motions, such as the first and last frames in Fig. 2a, it still detects false highlights (the bottom-left four frames). Compared with the above results, the proposed system (Fig. 2c) is less susceptible to sudden movement and color saliency.

Conclusion

This work introduces a new approach for summarizing sports videos. In order to detect the highlights in a video, attentive motion entropy is employed by combining salient motions with regional histograms of oriented gradients. Another event-boundary detection algorithm is also proposed in this work to search similar frames in a video, which is based on mutual information analysis. To identify the boundaries of motion events more accurately, differential salient motions and differential oriented gradients are analyzed according to the directivity changing rates between frames. Moreover, segmental spectro-temporal subtraction is used to smooth mutual information curves and to remove unnecessary highlight boundaries that have smaller mutual information. With the above-mentioned methods, our system is capable of generating video abstracts efficiently. Experiments on a 6-video database indicate that the proposed approach can achieve a precision and recall rate of 85.5 and 86.1 %, respectively. Besides, smoothness is also enhanced by 0.7 on average. Furthermore, a comparison reveals that the proposed approach is superior to the other baselines.

References

1. Bagga A, Hu J, Zhong J, Ramesh G (2002) Multi-source combined-media video tracking for summarization. In Proceedings of the 16th IEEE international conference pattern recognition, Quebec, Canada, Aug 11–15. IEEE computer society, Washington, pp 818–821
2. Liu T, Zhang H-J, Qi F (2003) A novel video key-frame-extraction algorithm based on perceived motion energy model. IEEE trans. circuits and systems for video technology 13(10):1006–1013
3. Duan L-Y, Xu M, Tian Q, Xu C-S, Jin JS (2005) A unified framework for semantic shot classification in sports video. IEEE Trans Multimedia 7(6):1066–1083
4. Li Z, Schuster GM, Katsaggelos AK (2005) MINMAX optimal video summarization. IEEE trans. circuits and systems for video technology, 15(10):1245–1256
5. Liu T-Y, Ma W-Y, Zhang H-J (2005) Effective feature extraction for play detection in American football video. In: Proceedings of the 11th international multimedia modeling conference (Melbourne, Australia, Jan. 12–14). IEEE computer society, Washington, pp 164–171
6. Ma Y-F, Hua X-S, Lu L, Zhang H-J (2005) A generic framework of user attention model and its application in video summarization. IEEE Trans Multimedia 7(5):907–919
7. Yeo B-L, Liu B (2005) Rapid scene analysis on compressed video. IEEE trans circuits and systems for video technology, 5(6):533–544
8. Cernekova Z, Pitas I, Nikou C (2006) Information theory-based shot cut/fade detection and video summarization. IEEE transactions circuits and systems for video technology, 16(1):82–91
9. Li Y, Lee S-H, Yeh C-H, Kuo C-CJ (2006) Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. IEEE Signal Process Mag 23(2):79–89
10. Taskiran CM, Pizlo Z, Amir A, Ponceleon D, Delp EJ (2006) Automated video program summarization using speech transcripts. IEEE Trans Multimedia 8(4):775–791
11. Chen C-Y, Wang J-C, Wang J-F, Hu Y-H (2007) Event-based segmentation of sports video using motion entropy. In: Proceedings of the 9th IEEE international symposium multimedia (Taichung, Taiwan, 10–12). IEEE computer society, Washington, pp 107–111
12. You J, Liu G, Sun L, Li H (2007) A multiple visual models based perceptive analysis framework for multilevel video summarization. IEEE trans. circuits and systems for video technology, 17(3):273–285
13. Chen B-W, Wang J-C, Wang J-F (2009) A novel video summarization based on mining the story-structure and semantic relations among concept entities. IEEE Trans Multimedia 11(2):295–312
14. Black MJ (1996) The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Comput Vis Image Underst 63(1):75–104
15. Walther D, Rutishauser U, Koch C, Perona P (2005) Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Comput Vis Image Underst 100(1–2):41–63
16. Walther D, Koch C (2006) Modeling attention to salient proto-objects. Neural Networks 19(9):1395–1407

17. Ma Y-F, Lu L, Zhang H-J, Li M (2002) A user attention model for video summarization. In: Proceedings of the 10th ACM international conference multimedia (Juan-les-Pins, France, Dec. 01–06). ACM Press, New York, pp 533–542
18. Lu S, King I, Lyu MR (2005) A novel video summarization framework for document preparation and archival applications. In: Proceedings of the 2005 IEEE aerospace conference (Big Sky, Montana, United States, Mar. 05–12). IEEE computer society, Washington, 1–10
19. Ngo C-W, Ma Y-F, Zhang H-J (2005) Video summarization and scene detection by graph modeling. IEEE transactions circuits and systems for video technology, 15(2):296–305
20. Chen Y-T, Chen C-S (2008) Fast human detection using a novel boosted cascading structure with meta stages. IEEE Trans Image Proc 17(8):1452–1464
21. Kamath SD, Loizou PC (2002) A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Proceedings of the IEEE international conference acoustics, speech, and signal processing (Orlando, Florida, United States, May 13–17). IEEE computer society, Washington, pp 4164–4167
22. Zhang T, Kuo C-CJ (2001) Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans Speech Audio Proc 9(4):441–457
23. Misra H, Vepa J, Bourlard H (2006) Multi-stream ASR: an oracle perspective. In: Proceedings of the ISCA international conference spoken language processing (Pittsburgh, Pennsylvania, United States, Sep. 17–21)
24. Gray AH, Markel JD (1974) A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. IEEE Trans Acoustics, Speech and Signal Processing 22(3):207–217

A Framework Design for Human-Robot Interaction

Yu-Hao Chin, Hsiao-Ping Lee, Chih-Wei Su, Jyun-Hong Li, Chang-Hong Lin, Jhing-Fa Wang and Jia-Ching Wang

Abstract Multimodal human-robot interaction integrates various physical communication channels for face-to-face interaction. However, face-to-face interaction is not the only communication method. The proposed framework achieves more flexible communications such as communication to others at a distance. Intimate and loose interactions are categorized as ubiquitous multimodal human-robot interaction. Therefore, this work presents a framework design for human-robot interaction using ubiquitous voice control and face localization and authentication implemented for intimate and loose interactions. The simulation results demonstrate the practicality of the proposed framework.

Introduction

Humanoid and human-friendly robots are useful in daily life owing to their ability to assist individuals in executing some tasks by easily communicating with each other. Face-to-face interaction is almost adopted for human-robot interaction. Additionally, individuals can simply, naturally, and efficiently interact with robots without location and time constrains. Such communication is called "ubiquitous human-robot interaction".

Many investigations on multimodal human-robot interaction focus on various communication channels in social interaction like speech and visual communication [1–4]. Unfortunately, the only communication in most of these systems is face-to-face interaction. Ubiquitous human-robot interaction has been proposed [5]

Y.-H. Chin · C.-W. Su · J.-H. Li · Chang-HongLin · J.-C. Wang (✉)
Department of Computer Science and Information Engineering,
National Central University, Jhongli, Taiwan, Republic of China
e-mail: jcw@csie.ncu.edu.tw

H.-P. Lee · J.-F. Wang
Department of Electrical Engineering, National Cheng-Kung University,
Tainan, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_118, © Springer Science+Business Media Dordrecht 2014

to exploit gesture and voice to realize different interactions such as intimate and loose interactions. For loose interaction, author of [5] employed a camera to look over the room to detect whether someone is requesting interaction through gesturing. However, such an approach is not the most natural and efficient means of interacting with robots because individuals need to wave their hands to call robots to come close to them in order for them to instruct the robots. Therefore, speech may be a better choice for interaction between human and robots. The proposed framework design for human-robot interaction developed ubiquitous voice control subsystem to construct six microphones in the room to detect whether someone is giving a request by voice commands for loose interaction. Furthermore, one microphone installed in each robot is utilized to record close talking for intimate interaction. As well as speech, face localization and authentication subsystem is applied to recognize a user face for intimate interaction, and to localize a user in a different distance for loose interaction.

The target scenario of the proposed framework is broad. For example, the framework can be employed in a ward or a hotel room, in which robots recognize user faces, then pay attention to their voice command like ‘take tea to me’ everywhere, and localize people to accomplish this task.

The remainder of this paper is organized as follows. Section The Framework Design for Human-Robot Interaction describes the framework design for human-robot interaction. Section Ubiquitous Voice Control then presents the ubiquitous voice control subsystem in detail. Section Experimental Results summarizes the experimental results. Conclusions are finally drawn in Section Conclusion.

The Framework Design for Human-Robot Interaction

The Framework Design for Intimate Interaction

Figure 1 describes the framework design. For the close interaction between the user and the robots, the face is first recorded by a camera on a robot and then authenticated by the face recognition module. If the authentication result is acceptance, then the audio signal recorded by a microphone embedded on a robot is processed. The recorded audio signal is usually distorted by distributed background noises in the real world. Thus, the speech enhancement module has to improve the noisy recorded audio signal. The enhanced audio signals are then recognized via a recognition module. If the enhanced audio signals are recognized as voice commands like “take tea to me”, then the robot executes the associate actions. The robot needs to execute two actions, one is taking tea and the other is localizing and recognizing the user. For the face localization and authentication subsystem, six audio signals recorded by six microphones embedded on ceiling are processed by the preprocessing module, and the strongest signal is selected to determine the region in which the user is localized. Additionally, the face detection module detects the user from the region size and position of the face. The face

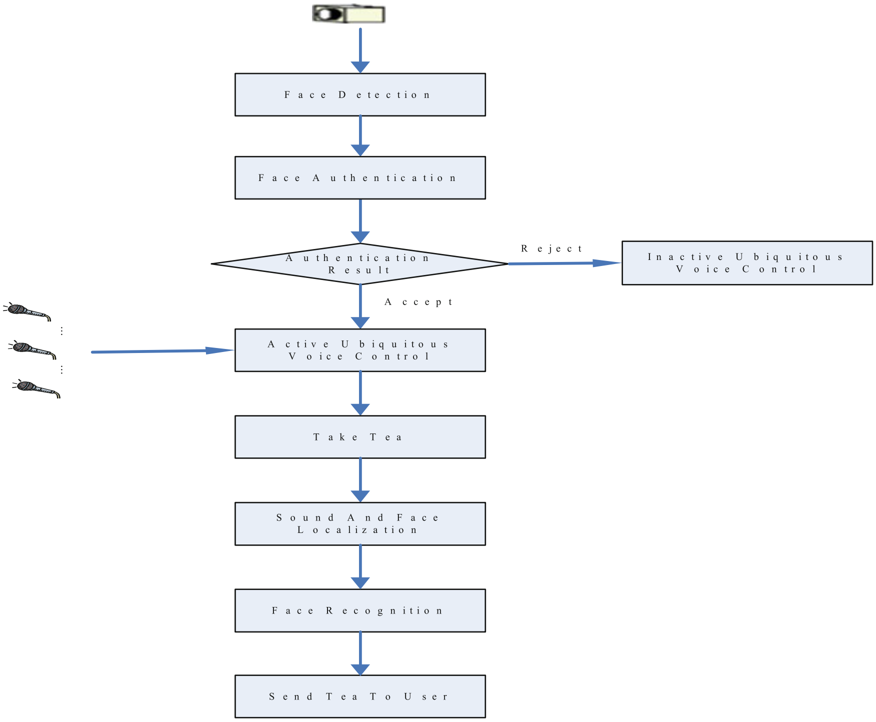


Fig. 1 The framework design

recognition module then recognizes the facial image. If the result is acceptance, then the robot sends tea to user.

The Framework Design for Loose Interaction

If users and robots interact loosely, then the robots locate the user and turn toward the user's location to detect and recognize the user's face. If the recognition result is acceptance, then the audio signals recorded by six microphones is transmitted to the preprocessing module which normalizes the signals and selects some stronger signals for enlarging the signals to be efficiently recognized and reducing computation time in the speech enhancement and speech recognition modules. The speech enhancement module is then run to enhance some recorded audio signals. Finally, the enhanced audio signals are then recognized by the speech recognition module. If the enhanced audio signals are recognized as voice commands, then the robots will execute the associate actions. The robots react as described in the previous section.

In the next section, we will address on the ubiquitous voice control subsystem.

Ubiquitous Voice Control

Our ubiquitous voice control subsystem mainly includes a speech enhancement module and a speech recognition module. To apply speech recognition in ubiquitous computing environment, a speech enhancement module is essential.

A well-known difficulty in speech recognition is recording audio signal with remote microphones at varying distances. Therefore, in many cases, head-mounted close-taking microphones such as Bluetooth headset are often used for speech recognition. This approach is called a wearable computing environment [6, 7]. However, this approach is not very comfortable and convenient for users.

Therefore, we need to overcome this limitation by using ubiquitous computing environment. The ubiquitous computing environment [6, 7] which utilizes multiple microphones provides an intuitive solution to this limitation. In a ubiquitous computing environment, microphones are embedded everywhere in the environment. Figure 2 illustrates the ubiquitous computing environment in the proposed design. Microphones are installed on the ceiling, and arranged separately by unique-distance. Therefore, microphones can record any audio signal irrespective of its origin. However, due to the restriction of the microphone number and distance, the recorded audio sources away from the microphones are noisy and weak. To overcome this problem, besides using sensitive and anti-noise microphones, speech enhancement and normalization are adopted. In the proposed

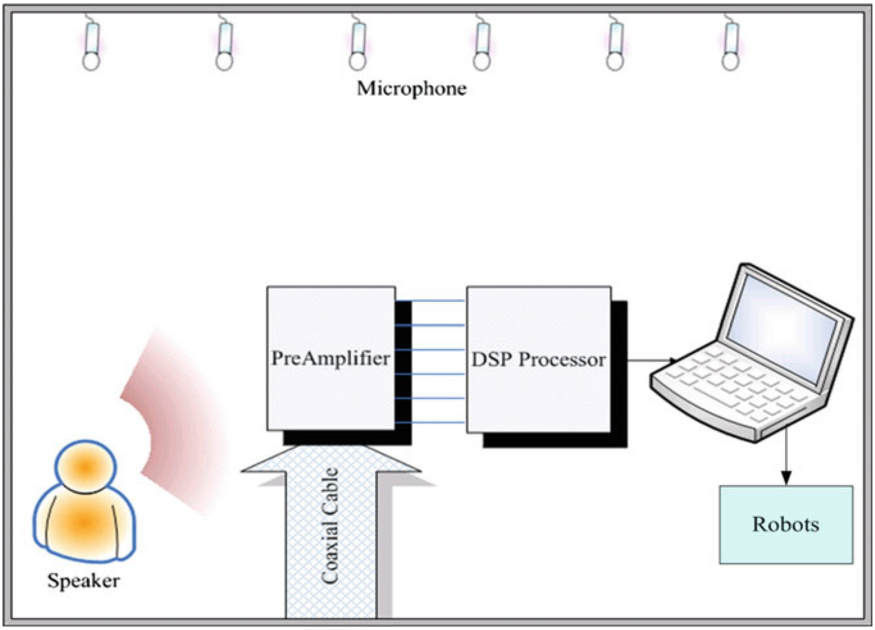


Fig. 2 Illustration of the ubiquitous computing environment in our design

ubiquitous computing environment, six noisy audio signals recorded by six microphones have to be pre-processed to reduce the number of noisy audio signals and speed up the computing of the system. This work uses voting strategy to select the most accurate recognition result. Furthermore, the maximum amplitude of the sounds also gives the hints to robots for sound localization.

Experimental Results

Six microphones were placed on the ceiling of an room for recording pervasive audio signals. Passing through the silver link microphone cables, the audio signals were delivered to a preamplifier card with six input ports. The parallel amplified streams were converted to a serial stream by the 16-bit PCMCIA analog input card, and then received by a notebook.

In our experiments, 19 keywords were adopted as voice commands. The testing database contained 57 spoken sentences. The recognition rate in the proposed ubiquitous computing environment was around 70 %. We are currently working on adapting the placement of microphones and survey the appropriate microphone suit for far distance recording.

Conclusion

This work presents a framework design for human-robot interaction which is being developed by building technologies for ubiquitous voice control subsystem and face localization and authentication subsystem. This multimodal human-robot interaction consists of two communication methods, intimate and loose interaction. These interactions are realized by combining ubiquitous voice control subsystem, including preprocessing, speech enhancement and speech recognition modules with face localization and authentication subsystem, including face detection and face recognition modules. Hence, the proposed framework design demonstrates a ubiquitous, convenient and natural environment, which is appropriate for human-robot interaction.

References

1. Breazeal C, Aryananda L (2002) Recognizing affective intent in robot directed speech. Auton Robots 12(1):83–104
2. Sidner C, Kidd C, Lee C, Lesh N (2004) Where to look: a study of human-robot engagement. In: Proceedings of intelligent user interfaces Madeira, Island of Funchal, Portugal, pp 78–84
3. Hanafiah ZM, Yamazaki C, Nakamura A, Kuno Y (2004) Human-robot speech interface understanding inexplicit utterances using vision. In: Proceedings conference on human factors in computing systems, Vienna, Austria, April pp 1321–1324

4. Yoshizaki M, Kuno Y, Nakamura A (2001) Human-robot interface based on the mutual assistance between speech and vision. In: Proceedings workshop on perceptive user interfaces, CD–ROM
5. Takeda H, Kobayashi N, Matsubara Y, Nishida T (1997) Towards ubiquitous human-robot interaction. In: Workiing Notes for *IJCAI*-97 Workshop Intell Multimodal Syst, pp 1–8
6. Furui S (2000) Speech recognition technology in the ubiquitous/wearable computing environment. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing Istanbul, vol 6, pp 3735–3738
7. Rhodes BJ, Minar N, Weaver J (1999) Wearable computing meets ubiquitous computing: reaping the best of both worlds. In: Proceedings of 3rd international symposium on wearable computers (ISWC'99), Oct

Blind Signal Separation with Speech Enhancement

Chang-Hong Lin, Hsiao-Ping Lee, Jyun-Hong Li, Chih-Wei Su, Yu-Hao Chin, Jhing-Fa Wang and Jia-Ching Wang

Abstract A new speech enhancement architecture using convolutive blind signal separation (CBSS) and subspace-based speech enhancement is presented. The spatial and spectral information are integrated to enhance the target speech signal and suppress both interference noise and background noise. Real-world experiments were carried out in a noisy office room. Experimental results demonstrate the superiority of the proposed architecture.

Introduction

Many multiple- microphone speech enhancement methods have been proposed to exploit spatial information to extract the single source signal [1–4]. To significantly reduce the number of microphones and do not require a priori information about the sources, blind signal separation (BSS) methods are adopted to effectively separate interfering noise signals from the desired source signal. The second-order decorrelation based convolutive blind signal separation (CBSS) algorithm was recently developed [5]. Since estimating second-order statistics is numerically robust and the criteria leads to simple algorithms [6].

We proposed a novel architecture, in which critical-band filterbank is utilized as a preprocessor to provide improved performance and further savings on convergence time and computational cost. However, only critical-band CBSS does not work for removing background noise, which originates from a complex combination of a large number of spatially distributed sources. Therefore, a subspace-

C.-H. Lin · J.-H. Li · C.-W. Su · Y.-H. Chin · J.-C. Wang (✉)
Department of Computer Science and Information Engineering,
National Central University, Jhongli, Taiwan, Republic of China
e-mail: jcw@csie.ncu.edu.tw

H.-P. Lee · J.-F. Wang
Department of Electrical Engineering, National Cheng-Kung University,
Tainan, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_119, © Springer Science+Business Media Dordrecht 2014

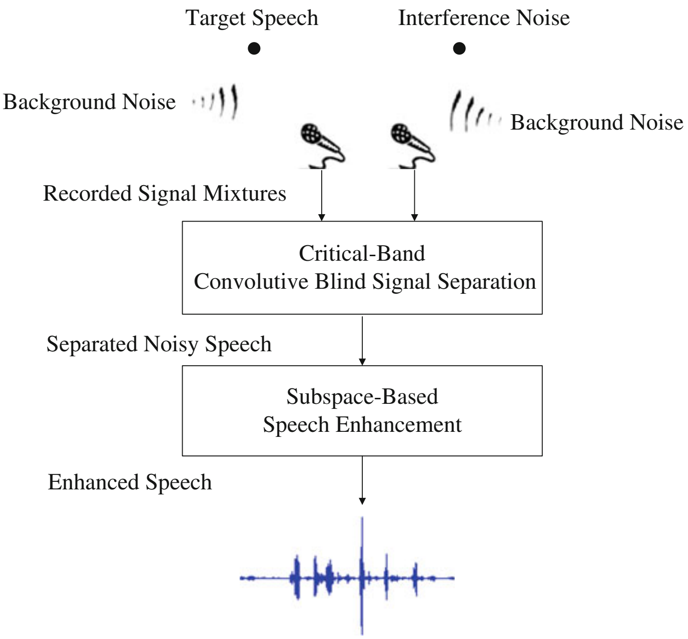


Fig. 1 Block diagram of proposed speech enhancement system

based speech enhancement method is utilized to reduce the background noise by exploiting additional spectral information [7].

Proposed Architecture

Figure 1 schematically depicts the block diagram of the proposed speech enhancement system. This architecture comprises a critical-band CBSS module and a subspace-based speech enhancement module. The input mixed signals are first processed by using the critical-band CBSS to separate the target speech from the interference noise. Next, the extracted target speech is fed into the subspace-based speech enhancement module to reduce the residual interferences and background noise. The proposed architecture adopts both spatial and spectral processing, and needs only two microphones.

Critical-Band Convolutive Blind Signal Separation

First, a critical-band filterbank based on the perceptual wavelet transform (PWT) is built from the psycho-acoustic model. The recorded signal mixtures are decimated into critical band time series by PWT. The CBSS is performed to separate the noisy speech and the interference noise in each critical band. A signal selection strategy based on high order statistics is then adopted to extract the target speeches. Finally, the inverse perceptual wavelet transform (IPWT) is applied to the critical-band extracted speeches to reconstruct the full-band separated noisy speech.

Perceptual auditory modeling is very popular for speech analysis and recognition. The wavelet packet decomposition is designed to adjust the partitioning of the frequency axis into critical bands which are widely used in perceptual auditory modeling. Within the 4 kHz bandwidth, this work uses 5-level wavelet tree structure to approximate the 17 critical bands derived based on the measurement [8, 9].

Convolutive Blind Signal Separation

This work assumes that two mixture signals $\bar{x}(t) = [x_1(t), x_2(t)]^{\mathrm{T}}$ composed of two point source signals $\bar{s}(t) = [s_1(t), s_2(t)]^{\mathrm{T}}$ and additive background noise $\bar{n}(t)$ are recorded at two different microphone locations:

$$\bar{x}(t) = \sum_{\tau=0}^{P} \mathbf{A}(\tau)\bar{s}(t-\tau) + \bar{n}(\tau). \tag{1}$$

The mixing matrix $\mathbf{A}$ is a 2×2 matrix and P represents the convolution order. Passing through the critical-band filterbank, PWT separates mixture signals into 17 critical-band wavelet packet coefficients. In each critical band, using an M-point windowed discrete Fourier transformation (DFT), the time-domain equation (1) can be converted into frequency-domain. The convolutive BSS is then performed in each critical band.

Signal Selection

In each critical-band, the CBSS has separated the mixed signals as speech dominant and interference dominant signals. Next, we should identify the target speech from the two separated outputs. Nongaussianity can be considered as a measure for discriminating the target speech and the interference noise by using kurtosis [3].

The last stage simply synthesize the enhanced speech using the inverse perceptual wavelet transform (IPWT).

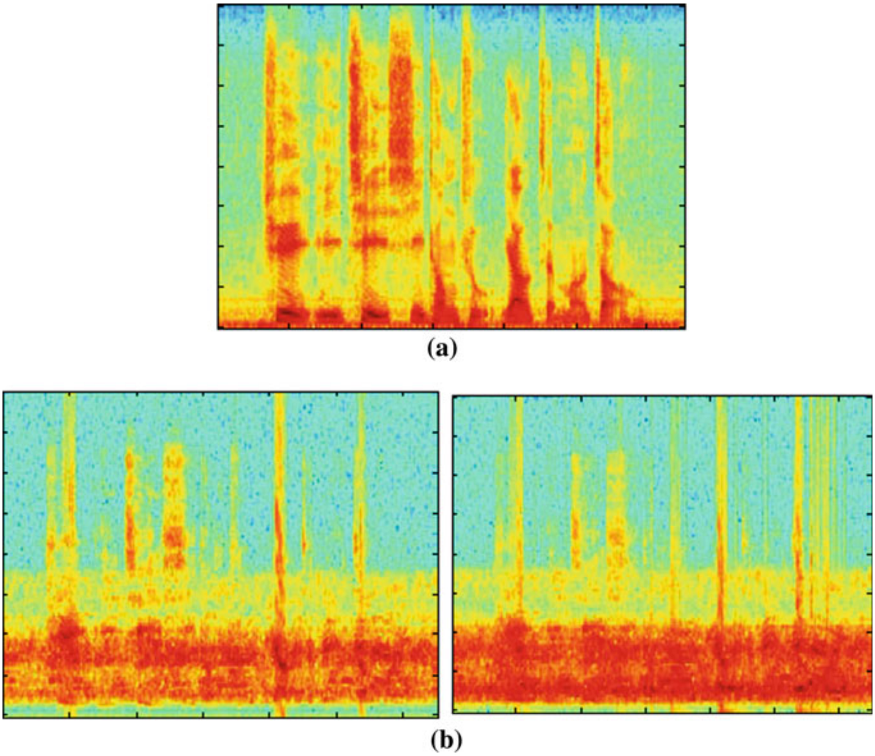


Fig. 2 **a** The original clean speech signals; **b** the 2-channel corrupted speeches under babble noise

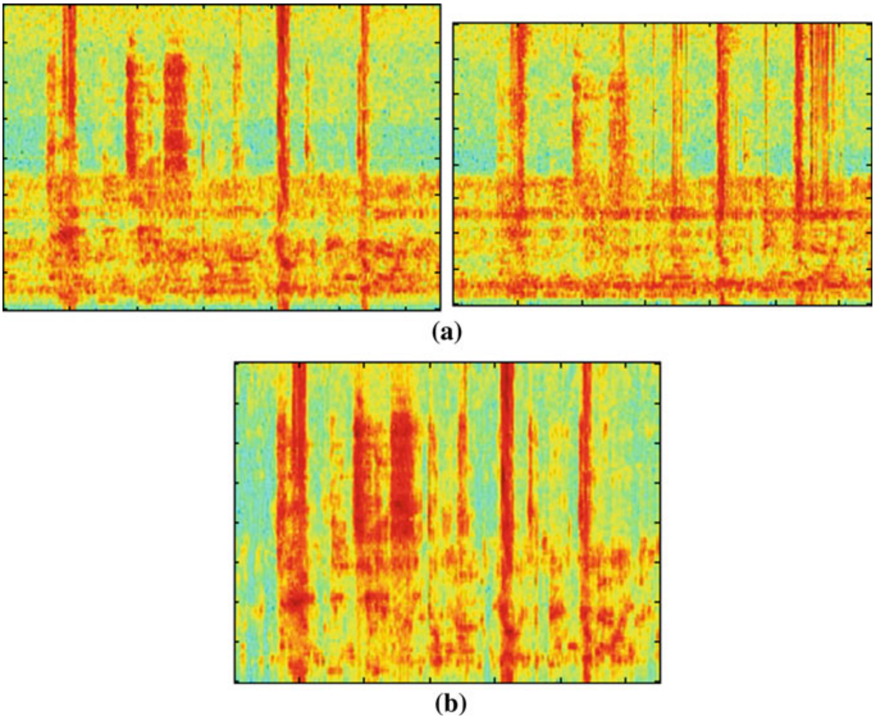
Subspace-Based Speech Enhancement

The subspace-based speech enhancement is used to enhance the separated noisy speech by minimizing the background noise. The additive noise removal problem can be described as a clean signal $\bar{s}$ being corrupted by additive noise $\bar{n}$. The resulting noisy signal $\bar{u}$ can be expressed as

$$\bar{u} = \bar{s} + \bar{n}, \qquad (2)$$

where $\bar{s}$= $[s(1), s(2), \ldots, s(L)]^{\mathrm{T}}$, $\bar{n} = [n(1), n(2), \ldots, n(L)]^{\mathrm{T}}$, and $\bar{u} = [u(1), u(2), \ldots, u(L)]^{\mathrm{T}}$. The observation period has been denoted as L. Henceforth, the vectors $\bar{s}$, $\bar{n}$, and $\bar{u}$ will be considered as part of real space R^L.

Ephraim and Van Trees proposed a subspace-based speech enhancement method [7]. The goal of this method is to find an optimal estimator that would minimize the speech distortion by adopting the constraint that the residual noise fell below a preset threshold.



(a)

(b)

Fig. 3 **a** The two-channel critical-band CBSS outputs; **b** the selected enhanced speech

Experiment Results

The experiment was performed with a speech source and a babble interference noise at an angle of 150° and a distance of 40 cm from the center of the microphone array. Twenty different spoken sentences were played, each with about 50,000 samples and babble noise in AURORA database was employed as interference noise.

For objective evaluation, the SNR measure was adopted to evaluate these speech enhancement algorithms. Additionally, the modified Bark spectral distortion (MBSD) was also applied to assess speech quality. Since MBSD measure, presented by Yang et al. [10], is a perceptually motivated objective measure for mimicking human performance in speech quality rating. In both measures, the proposed architecture significantly outperforms conventional subspace enhancement method.

Figure 2 shows the spectrograms of original clean speech and two speeches corrupted by babble noise. Figure 3 illustrates the spectrograms of the critical-based CBSS outputs and the enhanced result. Figure 3a clearly reveals that one output is target speech dominant, while the other is interference dominant.

Conclusion

This work develops a spatio-spectral architecture for speech enhancement. The architecture consists of a critical-band CBSS module and a subspace-based speech enhancement module. The spatial and spectral information are exploited to enhance the target speech, and to suppress strong interference noise and background noise using two microphones. Kurtosis analysis is then adopted to select the target CBSS output. The enhancement performance is improved significantly.

References

1. VanVeen BD, Buckley KM (1988) Beamforming: a versatile approach to spatial filtering. IEEE Acoust, Speech Sig Process Mag 5:4–24
2. Kellermann W (1991) A self-steering digital microphone array. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, vol 5, April pp 3581–3584
3. Low SY, Nordholm S, Togneri R (2004) Convolutive blind signal separation with post-processing. IEEE Trans Speech Audio Process 12(5):539–548
4. Visser E, Otsuka M, Lee TW (2003) A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments. Speech Commun 41(2):393–407
5. Parra L, Spence C (2000) Convolutive blind source separation of nonstationary sources. IEEE Trans Speech Audio Process 8(3):320–327
6. Parra L, Fancourt C (2002) An adaptive beamforming perspective on convolutive blind source separation. Davis G (ed) In: noise reduction in speech applications, CRC Press LLC
7. Ephraim Y, Van Trees HL (1995) A signal subspace approach for speech enhancement. IEEE Trans Speech Audio Process 3(4):251–266
8. Zwicker E, Terhardt E (1980) Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. J Acoust Soc Am 68:1523–1525
9. Chen SH, Wang JF (2004) Speech enhancement using perceptual wavelet packet decomposition and Teager energy operator. J VLSI Sig Process Syst 36(2–3):125–139
10. Yang W, Benbouchta M, Yantorno R (1998) Performance of the modified bark spectral distortion as an objective speech quality measure. In: Proceedings of IEEE international conference on acoust, speech, signal process pp 541–544

A Robust Face Detection System for 3D Display System

Yu Zhang and Yuanqing Wang

Abstract Face detection is a kind of extremely useful technology in many areas, such as security surveillance, electronic commerce and human–computer interaction and so on. Face detection can be viewed as a two-class classification problem in which an image region can be classified as being a "face" or "non-face". Detection and locating the position of the observers' face exactly play a critical role in stereoscopic display system. Accuracy, speed and stability are some main standards to evaluate an object-tracking system. The face detection system presented in the paper with classifiers trained by AdaBoost algorithm can meet the specific requirements of stereoscopic display in detecting speed, accuracy and stability. After accurate face detection, we utilize a certain method to detect the pupil in the area of face which is obtained in the above process. At last, the active 3D display equipment will project corresponding images of the same scene to users' pupil respectively to make sure the viewer can obtain the sense of depth. According to the experimental results, this system is highly accurate, stable and users can get well experience through this 3D display system.

Keywords Face detection · 3D display · AdaBoost

Introduction

Face detection is a kind of extremely useful technology in many areas, such as security surveillance, video search, electronic commerce, human–computer interaction and so on [1–3]. Face detection can be viewed as a two-class classification problem in which an image region could be classified as being a "face" or "non-face." Face detection also provides interesting challenges to the pattern recognition and machine learning area.

Y. Zhang (✉) · Y. Wang
Department of Electronic Science and Engineering, NanJing University, NanJing, China
e-mail: yxu828@sina.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_120, © Springer Science+Business Media Dordrecht 2014

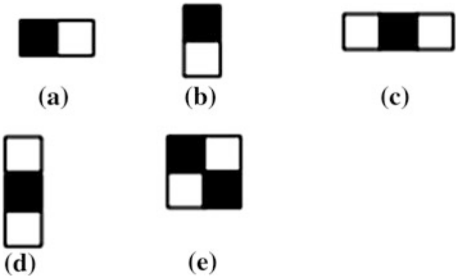
After several decades of research, the existing face detection algorithms can be generally divided into four categories: knowledge-based methods [4], feature invariant approaches [5], template matching methods [6] and appearance-based methods [7, 8]. Accuracy, speed and stability are some main standards of evaluating an object-tracking system. The face detection system presented in the paper with classifiers trained by AdaBoost algorithm [7, 8] can meet the specific requirements of stereoscopic display in detecting speed, accuracy and stability. Afterwards, the accurate location of pupil will be detected in the next process with reference to the face area in the image which could reduce the computational complexity effectively and improve the detection accuracy in pupil tracking process. Subsequently, the stereoscopic display system will accurately project the corresponding images to the pupils of observers to achieve good user stereoscopic experiences without any accessory equipment.

Haar Features

Haar-like features [7, 8] are a kind of simple rectangle features which has five basic feature templates to describe an area feature in an image as shown in Fig. 1. The object can be detected efficiently and quickly through these features than basic pixels. The value of haar feature can be defined as the sum of the white pixels subtracts the sum of the black pixels in grayscale area which reflects the local gray level change in an image. Some selected haar feature with well performance to distinguish face and non-face are shown in Fig. 2 and each haar feature is described collectively by the parameters of feature type, size and location in the image.

The five basic haar features could be placed at any position of the images in any size, so there will be lots of features in the training process and it will cost considerable time if we calculate these feature values directly. In order to enhance the computing speed, we utilized integral image [7, 8] to avoid plenty of repetitive computation. The value of every point in integral image could be defined as the sum of grayscale value of the pixels located on the left and upper the point which is shown in (1). The integral image of a whole image can be quickly calculated by some simple certain rules. Subsequently, we can calculate the haar feature value by some ordinary addition and subtraction operation with the integral image only.

Fig. 1 Some haar features



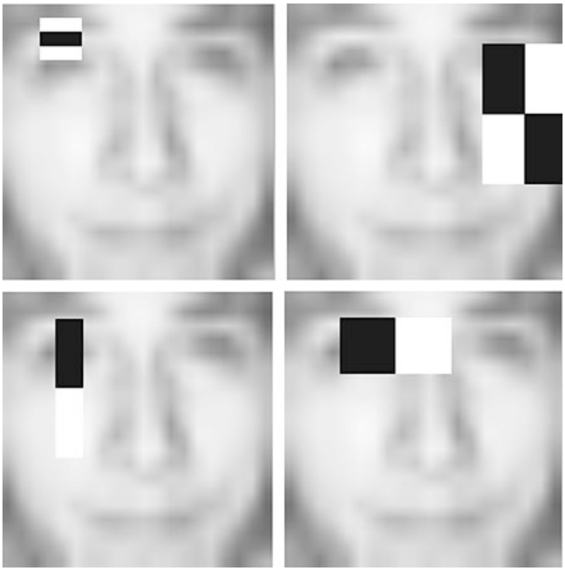


Fig. 2 Some haar features with well performance to distinguish face and non-face

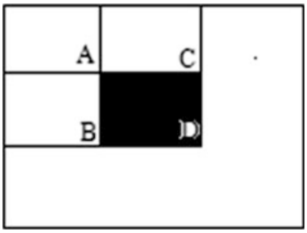


Fig. 3 Integral image

For instance, the sum of grayscales of black area in Fig. 3 could be calculated quickly only with reference to the values of the four points A, B, C, D in the integral image and the specific operation can be seen in (2). By this means, the haar feature can be calculated in constant time which could greatly reduce the computation complexity.

$$\mathrm{I}(\mathrm{x},\ \mathrm{y}) = \sum_{\substack{x\prime \le x \\ y\prime \le y}} \mathrm{i}(\mathrm{x}\prime, \mathrm{y}\prime) \tag{1}$$

$$\mathrm{I}_{area} = \mathrm{I}_A + \mathrm{I}_D - \mathrm{I}_C - \mathrm{I}_B \tag{2}$$

AdaBoost Algorithm

The AdaBoost algorithm [7, 8] is one kind of machine learning algorithm, which greatly improves the original boosting algorithm by dynamically adjusting the weight of weak classifiers to automatically meet the basic algorithm requirement.

The basic idea of Adaboost algorithm is to select some weak classifiers from weak classification space and integrate these weak classifiers to form a strong classifier according to certain rules. AdaBoost algorithm usually could be divided into two categories: discrete AdaBoost and real AdaBoost. In early discrete Adaboost algorithm, the classifier output are only two cases no matter it is a weak classifier or a strong classifier, which limits the performance of the weak classifier. Afterwards, in real AdaBoost, the weak classifier is improved to output real number which preferably depicts the confidence level meantime. The real Adaboost algorithm can make better performance than the discrete AdaBoost algorithm with the consideration from the experimental result so we choose the real Adaboost algorithm as our core method in the face detection system.

Experimental Results and Conclusions

The face detection system can work well in the test environment with single user and multiple observers. After accurate face detection, we utilize corresponding method to detect the pupil in the area of face which is obtained in the above process. After obtaining the coordinate of the pupils' location in the face, the active 3D display equipment will project a pair of images obtained in the same scene from different perspective to users' pupil respectively to make sure the viewer can obtain the sense of depth in order to form vivid stereoscopic feelings. Now this technology is utilized in our active 3D display system as a module to provide well stereoscopic experience for the observers without any attachment such as glasses and so on (Fig. 4).

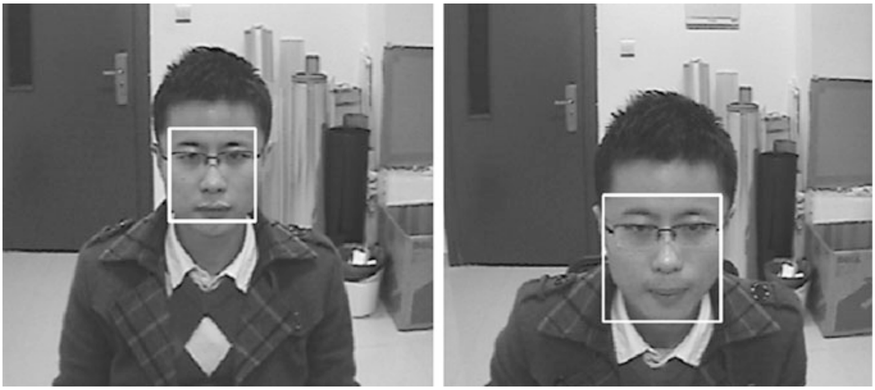


Fig. 4 some experimental results

Acknowledgments The project was supported by the program (Grant No. AHJ2011Z001).

References

1. Hsu RL, Abdel-Mottaleb M, Jain AK (2002) Face detection in color images. Pattern Analysis Mach Intell 24(5), 696–706. IEEE Press
2. Gorbenko A, Popov V (2013) Face detection and visual landmarks approach to monitoring of the environment. Int J Math Anal 7:213–217
3. Chen ZX, Liu CY, Chang FL, Han XZ (2013) Fast face detection algorithm based on improved skin-color model. Arabian J Sci Eng 38(3):629–635
4. Lin C, Fan KC (2001) Triangle-based approach to the detection of human face. Pattern Recogn 34(6):1271–1284
5. Li G, Xu Y, Wang J (2010) An improved adaboost face detection algorithm based on optimizing skin color model. In: 2010 sixth international conference on natural computation, vol 4, pp 2013–2015
6. Aiping C, Lian P, Yaobin T, Ning N (2010) Face detection technology based on skin color segmentation and template matching. In: 2010 second international workshop on education technology and computer science, vol 2, pp 708–711. IEEE
7. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vision 57(2):137–154
8. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, vol 1, pp 511–518 IEEE

Part XI
Virtual Reality for Medical Applications

Using Inertia Measurement Unit to Develop Assessment Instrument for Self-Measurement of the Mobility of Shoulder Joint and to Analyze Its Reliability and Validity

Shih-Ching Yeh, Si-Huei Lee and Yi-Hang Gong

Abstract Frozen shoulder is one type of shoulder disease that is commonly seen clinically, and its main symptom is the limit in the mobility of the shoulder joint of the patient and the shoulder pain. Since its therapeutic process of rehabilitation takes a long period of time, the patient usually abandons the therapy. In addition, under the current medical situation, physical therapist usually does not have extra time and space to help each patient densely to measure the mobility of the joint and to evaluate the progress of rehabilitation, hence, the patient usually is lack of in-time feedback to understand clearly the current rehabilitation progress, which in turn results in future low willingness of the patient to take rehabilitation, eventually, the goal of early and continuous therapy cannot be reached. Therefore, the objective of this research is to develop a set of "Assessment instrument for self-measurement of the mobility of shoulder joint", meanwhile, its reliability and validity is tested too. The system has associated wireless sensor technology and virtual reality, and the patient only has to follow the instruction and teaching on the screen to finish all kinds of standard shoulder joint actions, the progress in shoulder joint can then be evaluated at any time, eventually, the goal of real-time and self-assessment of the effectiveness of rehabilitation can be achieved.

Keywords Wireless IMU sensor · Virtual reality · Frozen shoulder

S.-C. Yeh · Y.-H. Gong (✉)
National Central University, Zhongli, Taiwan, Republic of China
e-mail: sam77917sam@hotmail.com

S.-C. Yeh
e-mail: shihchiy@csie.ncu.edu.tw

S.-H. Lee
Taipei Veterans General Hospital, Taipei, Taiwan, Republic of China
e-mail: sihuei.lee@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_121, © Springer Science+Business Media Dordrecht 2014

Introduction

Frozen shoulder is a general name for the injury of the soft tissue of shoulder and articular capsule, which is commonly seen in people around the age of 50. It is a commonly seen shoulder disease clinically, and its symptom includes the limit of active or passive mobility of shoulder joint and shoulder pain. Codman [1] was the first one to use the term frozen shoulder to describe such patient, later on, Neviaser [2] has observed the change of glenohumeral joint synovium and proposed adhesive capsulitis. The general population prevalence of frozen shoulder is usually in the range of 2–5 %, and the incidence among females is higher than that among males (with ratio of 58:42) [3]. The incidence for people with diabetes or hyperthyroidism is also higher than that of general people.

Lundberg [4] was the first one to propose the classification of frozen shoulder, namely, idiopathic or primary and secondary frozen shoulder, and the cause of disease of secondary frozen shoulder can also be divided into systemic, extrinsic and intrinsic. The systematic cause of disease includes diabetes, thyroid disease, and the extrinsic cause of disease includes stroke, humerus fracture and Parkinson's disease, etc. The intrinsic cause of disease includes rotator cuff injury, tendinitis or tendon calcination.

According to the disease process of frozen shoulder and histological check, Nevaiser and Nevaiser [5] has proposed four stages for frozen shoulder: (1) Pre-adhesive stage: It usually has only slight capsulitis, clinically, only slight pain can be felt at the end angle of the joint, and there is almost no problem such as the limit of mobility of the joint. (2) Freezing stage: Histologically, red and thickened joint capsule can be seen, and clinically, serious pain and limit on the joint is seen. (3) Frozen stage: Histologically, only slight inflammation can be seen in bursa synovialis, but the adhesive tissue becomes mature and thickening, meanwhile, the clinical pain symptom is also greatly reduced, and joint ankylosis becomes more significant, furthermore, active and passive mobility of joint becomes close. (4) Thawing stage: In this stage, almost no pain will be felt, the mobility of the joint will be significantly due to the completion of remodeling reaction.

Two major goals for the therapy of frozen shoulder are: (1) Reduce the pain in the disease part so that the patient can cooperate with the subsequent therapy; (2) Keep and improve the mobility of the joint so that the patient can get back to normal life as soon as possible. The most important thing in therapy is to use appropriate joint exercise, for example, stretching exercise (Including Codman's exercise, pulley therapy, front and side finger wall-climbing exercise and towel exercise, etc.) and joint mobilization, to extend the adhesive joint capsule so as to improve the mobility and function of the joint. In addition, appropriate muscle strengthening exercise should also be given to prevent muscle atrophy due to long term exercise reduction.

For the rehabilitation effectiveness of frozen shoulder, in addition to patient's subjective reduction in pain and progress in function, objectively, it has to rely on the test on active and passive mobility of joint implemented by the therapist, however, the implementation from the therapist is very time-consuming, and it takes an appropriate space too. Hence, if a good and simple joint assessment method for the patient to make self-test at home can be set up, then not only the patient can make self-test on the mobility of the joint, the home-stay rehabilitation patient can also get real time feedback too. In addition, since the patient can make the test at home, hence, only through appropriate learning, even without the existence of the therapist, the patient can, without spending too much time, complete the assessment of the active mobility of the shoulder joint.

Related Work

Virtual reality technology has fast development in recent years. Due to the progress in computer software and hardware technology and great enhancement in computer graphic capability, its application field becomes also much larger than before. Virtual reality is a technique using computer graphic and image composition technology in association with voice signal processing to simulate real environment in existence. The user, through all kinds of stimulations on the sense organs provided by the virtual reality system, can own the feeling like he is personally on the scene. Another feature is that the virtual reality environment allows the user to interact with the environment and to provide real time feedback response.

Today, there are lots of researches in the physical therapy industry using virtual reality in association with medical rehabilitation, and currently, most of the researches are focusing on stroke patient. Fischer et al. [6] has announced a rehabilitation system for the patient with upper extremity stroke. In the system, the specifically designed gloves developed by them together with sensor is used to let the patient make finger stretching rehabilitation, meanwhile, helmet display is also used to display the entire virtual reality scene. The user has to use the glove, and through the detection by the sensor, to catch and release all kinds of objects in the virtual reality environment to achieve the task, and the final data analysis result also proves that the patient gets significant progress after using this rehabilitation system. Kuttuva et al. [7] proposed another type of rehabilitation system for upper extremity stroke patient. The user sits beside a chair of the same height as his waist and faces the computer, and the hand on the disease side is put on the table, then a detector device is installed in the neighborhood of the wrist, through the detector, the moving track of the wrist of the user can then be detected, meanwhile, together with two games developed by them, rehabilitation exercise for upper extremity can be carried out. Camerao et al. [8] similarly used the stroke patient as the main target, together with the self-developed upper extremity sensor, the traditional rehabilitation therapeutic process can be put on virtual reality, and information such as hand moving speed and time lag of seizing the object can be more accurately assessed.

Hauschild et al. [9], in 2007 has published a paper regarding research using virtual reality to help patient with spinal cord injury for rehabilitation. In the research, electric stimulus and the power-supplied artificial limb for upper arm developed by them are used to put the traditional rehabilitation method on virtual reality. Through the helmet display, the patient can receive rehabilitation training using virtual reality to simulate the catching of object by upper extremity.

In 2000, scholars such as Rose et al. [10] has proposed a project for 210 persons under test to receive training in different groups, it was found that in the group tested with sense action training by stability tester in virtual reality shows significant progress in the within-the-group performance after therapy. After finishing the therapy, these people under test show significant progress when the same action is made by them in the real environment.

The above researches have explained the feasibility of applying virtual reality technology in medical rehabilitation, and they have also shown that virtual reality technology has become one very useful assisted tool in the medical rehabilitation field.

Goal

The main objective of this research is to apply wireless sensor and virtual reality technology to develop a set of system which can objectively assess the active mobility of the shoulder joint of the patient, meanwhile, the correlation between the measurement result of "Assessment instrument for self-measurement of the mobility of the shoulder joint" and the result of the traditional measurement is measured to assess the reliability and validity of the system.

System Architecture

This system architecture mainly associates wireless sensor IMU and virtual reality technology (Fig. 1). Wireless sensor is mainly used to detect the posture of joint on human body, through the installation of wireless sensor on limb and body part corresponding to each rehabilitation action, the acquisition of joint angle of that part is then performed, meanwhile, the acquired posture value is sent into the system for further processing. The system will, based on the received sensor posture value, display it on the screen according to human body animation reconstruction way so that the patient can clearly know the current body and joint posture; meanwhile, the system has also designed multi-representations, which include text, graph, image and audio, to remind the patient the current angle measurement result and the historical information, through these methods, real time information can be provided to the patient.

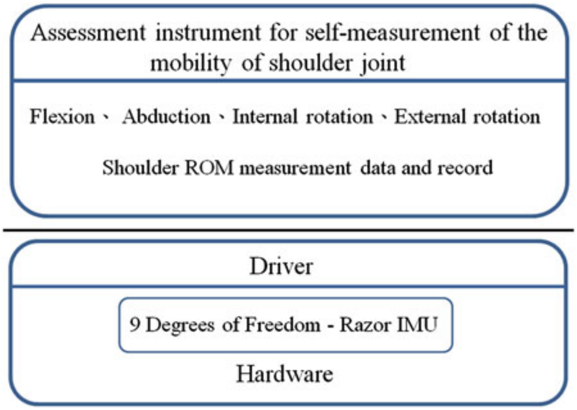

Assessment instrument for self-measurement of the mobility of shoulder joint
Flexion、Abduction、Internal rotation、External rotation
Shoulder ROM measurement data and record
Driver
9 Degrees of Freedom - Razor IMU
Hardware

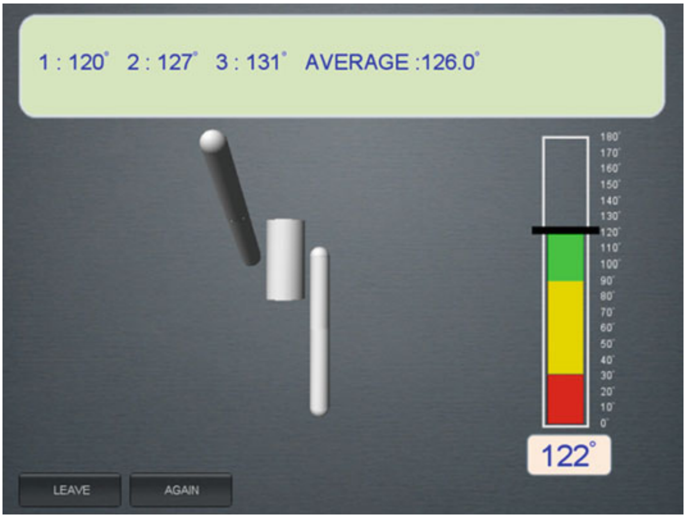
Fig. 1 System architecture drawing

Wireless IMU Sensor

This system mainly uses wireless sensor technology to measure the angle of the shoulder joint of the patient. Wireless sensor can be free from the binding of cable mainly through low power wireless transmission, meanwhile, using the IMU (Inertial measurement unit), the posture values of an object can be measured (pitch, yaw and row). 9DOF Razor IMU has associated three sensors such as gyroscope, accelerator and magnetic meter to provide three-axis and nine degrees of freedom of measurement, meanwhile, after measuring the posture value of an object, the wireless transfer module will be used to transfer the measurement data, and the angle received by the sensor will be converted and transferred to PC in the form of 0–255 digital signals through Zigbee.

Software

The main development environment of the system software is in game engine 3D Unity, and four actions are mainly measured in the angle measurement of the shoulder joint: shoulder abduction, shoulder flexion, shoulder internal rotation and shoulder external rotation. In the formal measurement process, on the right side of the screen, there will be angle bar chart to show the angle measured by the sensor so that the patient can know at any time the current bending angle of the shoulder joint, meanwhile, three recorded angles and average values made by the system will all be listed as reference used by the patient and the physical therapist. In addition, the system will also compare this record with the last record, finally, whether progress can be seen when this measurement result is compared with the previous measurement result (Fig. 2) will be listed. During the measurement process, the system will process the received posture value, and the posture of



1 : 120° 2 : 127° 3 : 131° AVERAGE :126.0°

122°

Fig. 2 Angle measurement screen

human body can then be displayed through animation, in other words, real time visual feedback can be sent to the patient.

Experiment Design

Participants

This research is foresighted, interfering, random-distributed and single-blind clinical test. Person under test meeting the condition will be assigned randomly according to the random number table generated by the computer into the experiment group and reference group. There will be 25 persons under test in each group.

Experimental Procedure

The shoulder joint angle to be measured in the experiment can be mainly divided into four actions: shoulder flexion, shoulder abduction, shoulder internal rotation and shoulder external rotation. Person under test in the experiment group and reference group will finish sequentially angle measurement for shoulder joint for four actions, and each action will be measured three times respectively, finally, the

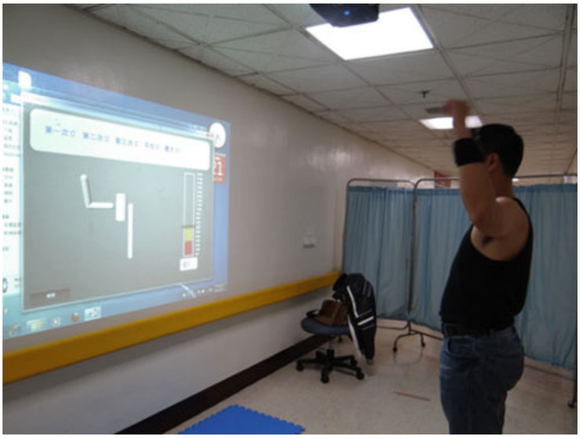


Fig. 3 The situation of real measurement of the angle of shoulder joint of person under test in the experiment group

results of three shoulder joint angle measurements are averaged, and then the values are recorded by the physical therapist.

In the experiment group, the person under test will follow the sequence of the action to attach the sensor on specific location, before the formal angle measurement, the person under test can, through the teaching film provided by the system, get to understand how to perform the action and the steps to be watched during the angle measurement stage, during the measurement process, the person under test can, through the system, know the real time information of the exercise of his own shoulder joint. In addition, through the display of animation, person under test can judge correctly the current exercise posture of the shoulder joint, hence, the goal of real time feedback can be achieved (Fig. 3). In addition, the system will also display each measurement result and the finished average value on the screen, not only it can draw the attention of the person under test on the entire experimental process, but also it can save lots of time for the entire set of experiment.

In the reference group, manual measurement method is used. The measurement of then angle of shoulder joint is mainly performed by the physical therapist using traditional angle measurement tool. During the experimental process, there is no assistance from any software system (see Fig. 4), and the experimental step is the same as that of the experimental group, that is, for each action, the physical therapist will use traditional measurement method to do three measurements, then the average value will be calculated and recorded.

Measurement and Analysis

In the experiment group, wireless sensor is mainly used for the measurement, but in the reference group, the physical therapist will use traditional angle measurement tool to do the measurement, for each action in both groups, three times of

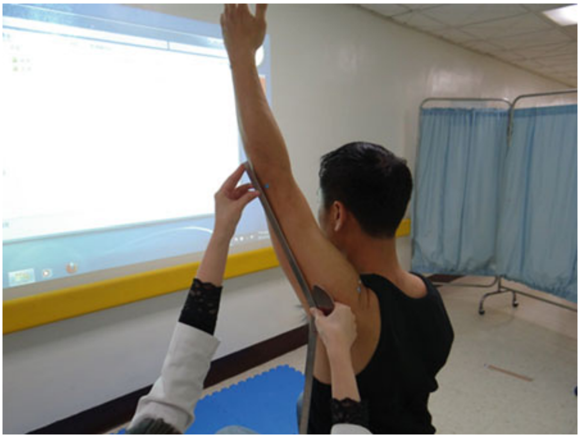


Fig. 4 The real situation of the measurement of shoulder joint angle of person under test in the reference group

angle measurement have to be done and the values then averaged, finally, the data will be analyzed by SPSS. The analysis is done through a comparison of correlation coefficient for the difference between the experiment group and reference group, meanwhile, the reliability and validity of the system is inspected too.

In addition, survey questionnaire is used to evaluate patient's feeling on five items such as Presence, Usefulness, Playfulness, Intention to Use and Flow Theory. In the scoring of the questionnaire, 7 selection items of each question are given with scores, very disagree is given with a score of 1, disagree is given with a score of 2, slightly disagree is given with a score of 3, no comment is given with a score of 4, slightly agree is given with a score of 5, agree is given with a score of 6, very agree is given with a score of 7, and the calculated average result is called susceptibility.

Discussion

In this study, correlation coefficient is mainly used to compare the correlation between experiment group and reference group, and the main task of correlation coefficient is to, aiming at two variables suspicious of the existence of linear correlation (in random), calculate a number to reflect the strength of the linear correlation, hence, when the absolute value of overall correlation coefficient get closer to 1, the linear correlation between two variables will get stronger.

The analysis result is as shown in Table 1. It can be seen that the correlation coefficient values of four actions are very close to 1, which means that the

Table 1 Comparison of correlation coefficient between experiment group and reference group

Flexion	Abduction	External rotation	Internal rotation
0.997^{**}	0.978^{**}	0.897^{**}	0.984^{**}

Table 2 The average susceptibilities of five scopes

Presence	Usefulness	Playfulness	Intention to use	Flow theory
5.8	5.7	6.1	5.4	5.7

measurement results between experiment group and reference group do not have significant difference but have very high correlation, and it means that the results measured by the system are very close to the results measured using traditional measurement method.

In addition, in the feedback of the survey questionnaire, 17 questionnaires have been collected. Table 2 shows the average susceptibility of five scopes of the survey questionnaire.

The result shows that the average susceptibilities of five scopes are all close to or even exceed a score of 5.5, which means that person under test has very good susceptibility on this system.

Conclusion

In this research, a new set of system has been proposed to measure the mobility of the shoulder joint of the patient with frozen shoulder. The system has mainly associated wireless sensor IMU and virtual reality technology. The final experiment result shows that the result measured by system sensor has very high correlation to the result measured by traditional angle measurement tool, moreover, there is no significant difference too. In addition, through questionnaire survey, the patient has very good susceptibility to the system, hence, it is proved that this system can effectively and reliably measure the mobility of the shoulder joint. Through the assistance of this system, in the future, it will save a lot of time and labor when the patient is going to measure the mobility of shoulder joint, meanwhile, the patient will be able to self-measure and self-assess the progress of his own rehabilitation.

References

1. Codman EA (1934) The shoulder: rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa. T. Todd Co, Boston, MA
2. Neviaser JS (1945) Adhesive capsulitis of the shoulder, a study of the pathological findings in periarthritis of the shoulder. J Bone Joint Surg Am 27(2):211–222
3. Thierry D (2005) Adhesive capsulitis. Emedicine 11:7
4. Lundberg BJ (1969) The frozen shoulder. Clinical and radiographical observations. The effect of manipulation under general anesthesia. Structure and glycosaminoglycan content of the joint capsule. Local bone metabolism. Acta Orthop Scand Suppl 119:1–59

5. Neviaser RJ, Neviaser TJ (1987) The frozen shoulder. Diagnosis and management. Clin Orthop Relat Res 223:59–64
6. Fischer HC, Stubblefield K, Kline T, Luo X, Kenyon RV, Kamper DG (2007) Hand rehabilitation following stroke: a pilot study of assisted finger extension training in a virtual environment. Top Stroke Rehabil 14(1):1–12
7. Kuttuva M et al (2006) The Rutgers Arm, a rehabilitation system in virtual reality: a pilot study. Cyberpsychol Behav 9(2):148–151
8. Cameirao MS, i Badia SB, Zimmerli L (2007) The rehabilitation gaming system: a virtual reality based system for the evaluation and rehabilitation of motor deficits. IEEE of Virtual Rehabilitation, Venice, Italy, pp 29–33
9. Hauschild M, Davoodi R, Loeb GE (2007) A virtual reality environment for designing and fitting neural prosthetic limbs. IEEE Trans Neural Syst Rehabil Eng 15(1):9–15
10. Rose FD et al (2000) Training in virtual environments: transfer to real world tasks and equivalence to real task training. Ergonomics 43(4):494–511

The Development of Interactive Shoulder Joint Rehabilitation System Using Virtual Reality in Association with Motion-Sensing Technology

Shih-Ching Yeh, Si-Huei Lee and Yao-Chung Fan

Abstract Virtual Reality combined with medical rehabilitation has become the modern trend. In the clinical therapy, frozen shoulder is a common shoulder disease. They usually use proper rehabilitation exercise to pole adhesive joint capsule. In these long time process, patients usually can't continue. Our research is using virtual reality to combine with motion tracking in order to develop interactive shoulder joint rehabilitation system. We decide to use interactive game-based oriented way and to give real-time visual and auditory feedback to improve user's motivation and willingness of doing rehabilitation. In this way, patients can achieve better result of rehabilitation compared with traditional therapy.

Keywords Frozen shoulder · Virtual reality · Microsoft kinect

Introduction

Shoulder joint is the joint with largest mobility among all human's joints. It is supported by complicated soft tissue, hence, once shoulder joint is injured, lots of daily life functions will be limited. Frozen shoulder is a general name for the injury of the soft tissue of shoulder and articular capsule, which is commonly seen in people around the age of 50. It is a commonly seen shoulder disease clinically,

S.-C. Yeh (✉) · Y.-C. Fan
National Central University, Zhongli, Taiwan, Republic of China
e-mail: shihchiy@csie.ncu.edu.tw

Y.-C. Fan
e-mail: 100522070@cc.ncu.edu.twl

S.-H. Lee
Taipei Veterans General Hospital, Taipei, Taiwan, Republic of China
e-mail: sihuei.lee@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_122, © Springer Science+Business Media Dordrecht 2014

and its symptom includes the limit of active or passive mobility of shoulder joint and shoulder pain. Codman [1] was the first one to use the term frozen shoulder to describe such patient, later on, Neviaser [2] has observed the change of glenohumeral joint synovium and proposed adhesive capsulitis. The general population prevalence of frozen shoulder is usually in the range of 2–5 %, and the incidence among females is higher than that among males (with ratio of 58:42) [3]. The incidence for people with diabetes or hyperthyroidism is also higher than that of general people.

Lundberg [4] was the first one to propose the classification of frozen shoulder, namely, idiopathic or primary and secondary frozen shoulder, and the cause of disease of secondary frozen shoulder can also be divided into systemic, extrinsic and intrinsic. The systematic cause of disease includes diabetes, thyroid disease, and the extrinsic cause of disease includes stroke, humerus fracture and Parkinson's disease, etc. The intrinsic cause of disease includes rotator cuff injury, tendinitis or tendon calcination.

Two major goals for the therapy of frozen shoulder are: (1) Reduce the pain in the disease part so that the patient can cooperate with the subsequent therapy; (2) Keep and improve the mobility of the joint so that the patient can get back to normal life as soon as possible. The most important thing in therapy is to use appropriate joint exercise, for example, stretching exercise (Including Codman's exercise, pulley therapy, front and side finger wall-climbing exercise and towel exercise, etc.) and joint mobilization, to extend the adhesive joint capsule so as to improve the mobility and function of the joint. In addition, appropriate muscle strengthening exercise should also be given to prevent muscle atrophy due to long term exercise reduction. Before implementing exercise therapy, hot compression, short wave, ultrasonic wave or interference wave, etc. can be used first to improve the extensibility of joint capsule, muscle, tendon and ligament, etc.

For the rehabilitation of frozen shoulder, the patient has to go between the home and hospital in a long term, physical and psychological depletion usually lets the patient give up, and the effectiveness of rehabilitation is thus reduced. In recent years, virtual reality is gradually applied in the rehabilitation therapy, and computer aided virtual reality can accurately create task of different strength and difficulty, through digital interface control, the doctor or therapist can easily set up and adjust the environment stimulations. This makes the design of the therapeutic process of rehabilitation richer, not only the patient can be provided with more chances to practice, but also real time sound and image feedback can be provided so that the patient can have better limb and body practice result within real-time and interactive operation environment design. Together with the game design, the rehabilitation therapy will become more interesting, and the patient's motivation for rehabilitation will be enhanced, in other words, through the giving of certain feedback and feeling of achievement to the patient, the maximal effectiveness of the rehabilitation can be achieved.

Related Work

Along with technological progress, we gradually are capable of using some high tech instrument and technology to help patients. Today, lots of scholars performing medical rehabilitation related researches and first-line doctor and therapist around the world start to use these technologies to make clinical therapy, in other words, they perform all kinds of new therapeutic process of rehabilitation and development of rehabilitation technology based on Virtual Reality and Augmented Reality, the design concept of user centered, consideration of the perception of the user on the system, usability and immersion, use of the interactive mode and strategy provided by the human machine interface [5, 6].

Virtual reality does not take too much expense to get rehabilitation situation close to the real world. Moreover, virtual reality can be adjusted on the schedule table according to the level of injury and rehabilitation of the patient. Currently, researches associating virtual reality with medical rehabilitation are mostly on patients with stroke [7, 8].

There is a research [9] used virtual reality to construct interactive game for rehabilitation training. In the research, Microsoft Kinect Sensor is also used to make action control of the full body, and the user, under the sensing of the sensor, has completed a series of rehabilitation balance tasks.

These researches have pointed out that the association of virtual reality with medical rehabilitation is feasible, in addition to that, motion-sensing detection can also be added to increase interactive element and to enhance the effectiveness of rehabilitation.

Goal

The major goals of this research are:

1. The situation design and system analysis of interactive shoulder joint rehabilitation system.

This research has associated the application of Kinect sensor to design a set of interactive virtual reality rehabilitation system, and innovative technology has been applied to construct new and interactive rehabilitation system. For the selected rehabilitation exercise mode, life situation simulation is performed, and all kinds of perceptual stimulations are designed. In addition, since the assisted equipment for today's rehabilitation action training only has single direction functional purpose, there is lack of signal feedback design, hence, for the rehabilitation effectiveness aspect, in addition to the subjective assessment from the rehabilitation doctor and the patient himself, there is still no quantified objective analysis. The interactive virtual reality rehabilitation system designed in this research not only provides interactive, virtual and real rehabilitation assisted tool, but also can provide quantified index of the effectiveness of rehabilitation to the patient, and all these can enhance the motivation of continuous rehabilitation of the patient.

2. Initial investigation was made on the effectiveness of interactive shoulder joint rehabilitation system for the improvement of frozen shoulder.

Interactive and signal feedback medical assisted training simulator in association with Kinect sensor is used together with perspective and interventional research method to cure frozen shoulder patient and to enhance therapeutic motivation and to quantify therapeutic effectiveness. Empirical medical method is used to assess the subjective and objective effectiveness difference before and after the therapy. Moreover, the overall effectiveness is investigated so as to set up the clinical rehabilitation new model for frozen shoulder patient.

System Architecture

This system uses Unity3D as the development platform to develop five traditional actions of rehabilitation such as flexion, abduction, internal rotation, external rotation and circumduction. The association of virtual reality and medical rehabilitation has resulted in the design of seven rehabilitation games. Kinect is used together to prepare interactive shoulder joint rehabilitation system so that during the rehabilitation process, training and assessment can be performed. Based on the requirement, system construction is done by aiming at the following two goals:

1. Different training posture is given according to the requirement of severity at different quadrant for the patient's frozen shoulder:

For example, if the patient has serious flexion angle problem, then the patient should be made to face the screen to train the flexion extension of the shoulder joint;
If the patient has more serious flexion abduction angle problem, then the patient should be made to face screen to train the abduction extension of the shoulder joint.

2. Task of different level of difficulty should be given according to different lifting capability of the arm of the patient:

Since the severity and speed of progress of each frozen shoulder patient is quite different, hence, each patient will have different capability during each time of rehabilitation (Fig. 1).

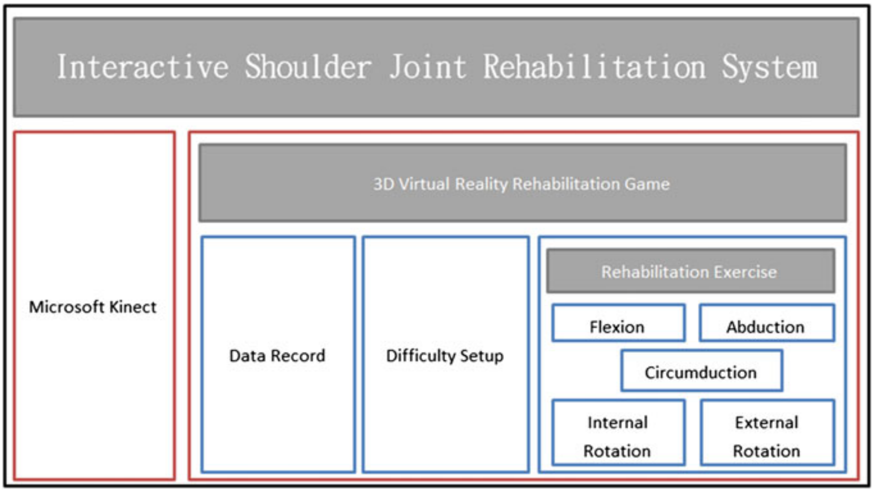


Fig. 1 System architecture

Hardware

In this interactive shoulder joint rehabilitation system, Kinect is used to acquire node coordinate of the skeleton of human body, then these information are transferred back to the computer, and in Unity3D, program is written to receive these skeleton information so as to achieve the interaction among scenes.

Kinect itself has two lenses and one infrared transmitter, and a total of three cameras (RGB color camera in the middle, infrared CMOS camera in the right side and infrared transmitter in the left side). Through the camera, object moving in front of the lens is caught; in addition, there is also one set of array type microphones. In the data acquisition part, Kinect has provided more than 20 skeleton nodes for the entire body to record the location information of the body skeleton in the space when the user is moving.

Software

The main development environment of this system software is based on game engine Unity3D. Unity3D is interactive game engine for development, after simple installation, it can be easily operated by the user. It has advantages such as user friendly interface and simple operation; hence, it plays a very critical role in the development of virtual reality scenes. The software has been used to develop task goal, and these game based tasks have brought brand new feeling to the patient in the rehabilitation process. In addition to that, using the image, voice and video to provide instruction for rehabilitation action for the patient can bring deeper impression and memory to the patient as compared to that by oral explanation. Such task oriented rehabilitation method not only can enhance patient's usage motivation but also can enhance the rehabilitation efficiency. In addition, some interactive feedback designs are added to transfer patient's attention on the pain during the rehabilitation process and to remove their discomfort in their mind. Software design can be divided into three directions:

1. **Situational rehabilitation task design**

In the traditional therapeutic process of rehabilitation, real task is used to lead the patient to complete the task goal and to achieve the goal of exercise, training and rehabilitation. Each game of this system has very clear task instruction to tell the patient to achieve the goal of the game as much as possible.

2. **Human machine interactive design**

The interaction between the user and virtual reality environment includes the tracking of hand location and the visual and audio feedback.

The tracking technology of hand location uses Kinect to acquire the skeleton information of the upper extremity, in other words, the program will read these information into the game scene of Unity3D, and skeleton information will be corresponded to the setup target object, hence, during the game process, the patient can operate on scene interactive object to achieve the task goal.

Visual feedback part is displayed in score. During the game process, the upper side of the screen will display the accumulated score or the remaining time; after the game is ended, the upper side of the screen will display total score or total time spent; in the entire process, the edge of the screen will display bar chart, that is, it will display the goal set up this time and the highest angle arm can be lifted at that moment. Moreover, sound feedback is displayed in different sound effect, in the beginning of the game, there will be instruction voice, beside, during the game process, when the patient has achieved the goal, feedback will be given immediately too. Such design can reflect the effectiveness of rehabilitation of the patient in real time, in other words, the patient can see real performance and level of progress exactly, and the rehabilitation motivation can then be greatly enhanced.

3. **Task performance**

The system itself, after the finish of the game, will display the total score acquired and total time spent by the patient so that the patient can know the game result.

Such situational rehabilitation task design process needs a professional therapist to lead and instruct 40 patients to try for several times, and the feedback and suggestion need to be recorded at any time so as to revise the software into a situational mode that can fit to most of the patients (Fig. 2).

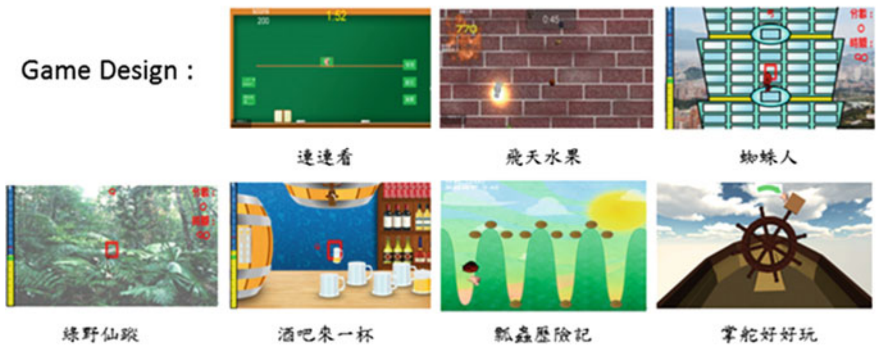


Fig. 2 Unity3D game design

Experiment Design

Participants

This research is perspective, interventional, randomized controlled and single-blind clinical test. Persons meeting the condition will be assigned randomly according to random table into experiment group and reference group. 40 persons are expected to be used in this research.

Experimental Procedure

The shoulder joint angle to be measured in the experiment can be mainly divided into four actions: shoulder flexion, shoulder abduction, shoulder internal rotation and shoulder external rotation. Person under test in the experiment group and reference group will finish sequentially angle measurement for shoulder joint for four actions, and each action will be measured three times respectively, finally, the results of three shoulder joint angle measurements are averaged, and then the values are recorded by the physical therapist.

Before the start of the experiment, the experiment process and purpose of this experiment will be explained first to the person under test, and letter of consent will be singed under the agreed condition. This experiment is designed into two times each week for a total period of four weeks. Before and after the experiment, professional medical personnel will make Constant-Murley Score (CMS) assessment, then the experiment group and reference group will follow the following flow to do rehabilitation:

Experiment group: The shoulder joint of the disease side will receive hot compression and electric therapy of interfering wave, and each time will last 20 min, for two times a week and a total of 4 weeks. Moreover, it will receive rehabilitation from interactive shoulder joint rehabilitation system, for 20 min each time, two times a week and a total of 4 weeks.

Reference group: The shoulder joint at the disease side will receive hot compression and electric therapy from interfering wave, 20 min each time, two times a week and a total of four weeks. Person under test in this group will receive homestray rehabilitation exercise for general frozen shoulder, and the rehabilitation action is similar to that in the experiment group, but there is no situational design (Codman's exercise and front and side finger wall-climbing exercise).

When rehabilitation from interactive shoulder joint rehabilitation system is received, the implementation sequence each time is fixed, and each person each time will finish these seven rehabilitation games.

Measurement and Analysis

If division is made based on the measurement time, then one respective measurement will be done before the start of the therapeutic process and at the end of the entire therapeutic process, and the measurement will be done aiming at the shoulder joint angle and the overall function of the upper extremity of the patient. The part-time assistant, without knowing the experiment group or reference group, will be responsible for the measurement of the angle of shoulder joint of the patient before and after the therapy.

1. Active and passive shoulder joint angle: Standardized angle measurement tool is used to measure the angle of shoulder joint at the disease side using standard posture, and it includes flexion, abduction, internal rotation and external rotation.
2. The overall function of upper extremity: Using Constant-Murley Score (CMS) for the assessment.

Before and after the experiment on the patient, professional medical and nursing personnel should ask the patient the severity of pain, then together with the extendable angle of the patient, it will be quantified into scores with full score of 100.

For the comparison of differences among numerical variables, non-parameteric statistics (Kruskal–Wallis Test) is used to carry out statistical analysis. In the statistics, $p < 0.05$ is used to represent that there is significant statistical difference. Moreover, survey questionnaire is prepared to assess patient's perception on Presence, Playfulness, Intention to Use and Flow Theory, etc. In the scoring of the survey questionnaire, 7 selection items of each question are given with score respectively, when it is very disagree, score of 1 will be given, when it is disagree, score of 2 will be given, when it is slightly disagree, score of 3 will be given, when it is no comment, score of 4 will be given, when it is slightly agree, score of 5 will be given, when it is agree, score of 6 will be given, when it is very agree, score of 7 will be given, and the calculated average result is called perception.

Table 1 *P*-value comparison between experiment group and reference group

Median(IQR) (degree)	Experiment group (before)	Reference group (before)	P-value	Experiment group (after)	Reference group (after)	*P*-value
Flexion	149.1 (129.2–162.7)	149.0 (131.7–157.5)	0.904	171.6 (153.2–174.0)	166.0 (154.7–171.2)	0.050
Abduction	145.8 (104.4–163.5)	144.0 (100.7–163.9)	0.947	168.0 (150.3–171.3)	157.7 (127.7–169.9)	0.040
External rotation	62.1 (55.7–83.9)	60.0 (39.7–71.6)	0.076	83.65 (71.5–88.2)	74.3 (56.5–81.2)	0.017
Internal rotation	40.5 (36.4–64.2)	37.1 (21.9–53.7)	0.051	65.5 (54.0–71.9)	50.65 (40.2–59.0)	0.003

Discussion

Moreover, non-parametric statistics of Wilcoxon Rank Sum Test is used for statistical analysis, and it was found that *P* value <0.05, which means that significant difference does exist.

The analysis results are as shown in Tables 1 and 2. The *P*-values of four actions are all <0.05 and the result shows that there is significant difference between the experiment group and the reference group. It means that our designed rehabilitation system can make greater progress to the angle of shoulder joint of the patient as compared to that of the traditional therapy (Table 3).

Table 2 Comparison of *P*-value of CMS score between experiment group and reference group

Constarit-Muriey score(CMS)	Experiment group (before)	Reference group (before)	P-value	Experiment group (after)	Reference group (after)	*P*-value
Median(IQE)	63.5 (43.5–70.7)	63.0 (50.0–71.2)	0.738	85.0 (72.5–89.0)	76 (68.2–84.7)	0.046

Table 3 Average perception of four scopes

Presence	Playfulness	Intention to use	Flow theory
5.4	5.7	5.5	5.7

Conclusion

The therapeutic process of rehabilitation is very tedious. As technology advances continuously, we can associate virtual reality with medical rehabilitation model to bring brand new feeling and perception to the patient, and we can also create the experience of playing game and performing rehabilitation at the same time.

Through the turning of the patient's attention onto playing games, the patient can forget the pain perceived in doing rehabilitation and the boring during tedious rehabilitation process. Interactive shoulder joint rehabilitation system indeed shows its effectiveness in therapy from the clinical test result, when it is compared with the current frozen shoulder rehabilitation, during the entire rehabilitation process, the patient not only has objective and quantified data to be referred to, but also can use interesting content of situational task as well as visual and audio real time feedback, hence, the rehabilitation motivation of the patient is greatly enhanced. Eventually, optimal therapeutic effectiveness can be reached continuously.

This study is a pilot study, which has used very new technology and tool to design and develop "interactive shoulder joint rehabilitation system" that has put together computer aided tool virtual reality, and the clinical therapeutic effectiveness is only an initial test. In the future, it is hoped that it can be developed and extended into software with more shoulder joint rehabilitation games, and it is hoped that deeper, more perfect and larger scale of clinical test can be performed. It is hoped that cloud technology can be associated in the future to create a new model of remote health care.

References

1. Codman EA (1934) The shoulder: rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa. T. Todd Co, Boston, MA
2. Neviaser JS (1945) Adhesive capsulitis of the shoulder, a study of the pathological findings in periarthritis of the shoulder. J Bone Joint Surg Am 27:211–222
3. Thierry D (2005) Adhesive capsulitis. Emedicine 11:7
4. Lundberg BJ (1969) The frozen shoulder. Clinical and radiographical observations. The effect of manipulation under general anesthesia. Structure and glycosaminoglycan content of the joint capsule. Local bone metabolism. Acta Orthop Scand Suppl 119:1–59
5. Rheingold H (1991) Virtual reality (1st edn). Simon and Schuster, New York
6. Rizzo AA et al (2006) A virtual reality scenario for all seasons: the virtual classroom. CNS Spectr 11(1):35–44
7. Sveistrup H (2004) Motor rehabilitation using virtual reality. J Neuroeng Rehabil 1(1):10
8. Meldrum D et al (2012) Virtual reality rehabilitation of balance: assessment of the usability of the Nintendo Wii((R)) Fit Plus. Disabil Rehabil Assist Technol 7(3):205–210
9. Lange B et al (2011) Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor. Conf Proc IEEE Eng Med Biol Soc 2011:1831–1834

Development of a Virtual Reality-Based Pinch Task for Rehabilitation in Chronic Hemiparesis

Shuya Chen, Shih-Ching Yeh, Margaret McLaughlin, Albert Rizzo and Carolee Winstein

Abstract Impaired pinch performance affects dexterity function after stroke. Virtual reality-based training may be beneficial for improving dexterity function. This study aimed to develop a virtual reality-based pinch task and to investigate its feasibility for chronic hemiparesis. The pinch task in the virtual environment was accomplished by coordinating two PHANTOM devices that provide haptic feedback. Participants grasped and lifted a virtual cube with 30-sec time limit for 10 trials. Cube size, cube mass and lift height were systematically varied. The participant poststroke attempted an average of 38 trials per session with a 60 % success rate and without complaint of fatigue or pain. After the training, the participant poststroke decreased the total time. However, the peak pinch force did not change. The results suggest that the virtual reality-based pinch task was

S. Chen (✉)
Department of Physical Therapy, China Medical University, Taichung, Taiwan
e-mail: sychen@mail.cmu.edu.tw

S. Chen · C. Winstein
Division of Biokinesiology and Physical Therapy at the School of Dentistry, University of Southern California, Los Angeles, CA, USA
e-mail: winstein@usc.edu

S.-C. Yeh
Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan
e-mail: shihchiy@csie.ncu.edu.tw

M. McLaughlin
Annenburg School for Communication and Integrated Media Systems Center, University of Southern California, Los Angeles, CA, USA
e-mail: mmclaugh@usc.edu

A. Rizzo
Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA
e-mail: arizzo@usc.edu

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_123, © Springer Science+Business Media Dordrecht 2014

feasible for chronic hemiparesis. Further investigation is warranted to better understand the effect of pinch force regulation using hepatic feedback.

Keywords Virtual reality · Pinch · Rehabilitation · Hemiparesis

Introduction

Stroke is a leading cause of long-term disability [1]. Recently, evidence show that stroke survivors have not used their affected arm to full extent in daily life even the arms are able to do so [2, 3]. While many patients regain most of the reaching and grasp capabilities for their upper extremity function, recovery of the pinch skill remains incomplete for most of the patients. Pinch movement is one the important skills of upper extremity [4, 5]. Impaired pinch skill significantly affects dexterity function after stroke. With practicing various interesting exercise tasks for considerable amount of time, the regained pinch capability would significantly improve the upper extremity function after stroke.

Virtual reality-based training has been shown to be beneficial for the enhancement of hand function poststroke [6–9], because game-like exercises in virtual environment are motivating and allow therapists to easily modify training parameters [8, 10]. This study aimed to develop a new virtual reality-based task for pinch function and to investigate its feasibility for chronic hemiparesis.

Methods and Materials

Participants

One participant with poststroke hemiparesis and one age-matched control participant were recruited. The inclusion criteria for the participants poststroke were: (1) stroke at least 1 month prior, (2) more than 18 years of age, (3) Mini-Mental Status Exam score $\geq$ 24, (4) no significant range of motion limitations in the hemiparetic upper extremity, (5) ability to perform the VR-based pinch task. Tested hand and the total time on pinch task were also matched for both participants.

Virtual Reality-Based Pinch Task

The pinch task was one of four upper extremity virtual reality-based tasks from our pilot project [8]. To develop the pinch task (specifically for the opposite movement

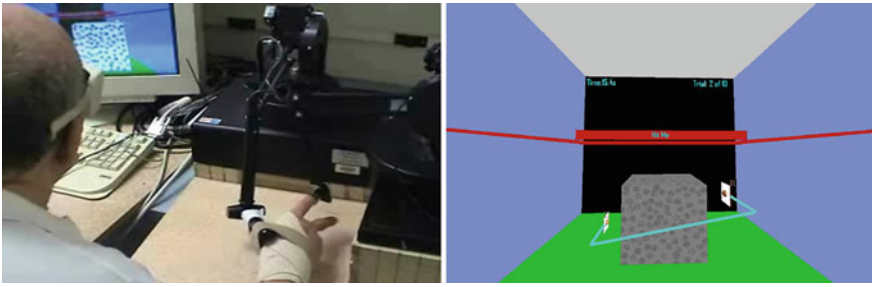


Fig. 1 Actual training with two PHANToM devices (*left*) and the display (*right*)

of thumb and index finger) in virtual environment, there were two PHANToM devices (SensAble Technologies) placed perpendicular to each other for the pinch movement (Fig. 1) and reconfigured to provide hepatic feedback for the pinch task. The goal of pinch task was to successfully pick the virtual cube up from the floor, lift it to a specified height, and place it back on the floor without dropping. Haptic feedback was provided for the thumb and index finger through the PHANToM devices, so that the participants felt they were lifting a real cube with mass. There were 10 trials per block. Each trial was configured using eight parameters: cube width (20–40 mm), cube height (20–40 mm), cube length (20–40 mm), mass (50–150 gw), dynamic friction (0.5–1.0), static friction (0.5–1.0), stiffness (0.5–1.0), and lift height (20–80 mm). Cube size, cube mass and lift height could be systematically varied according to the participant's capability level. A maximum of 30 s was allowed for each trial. The number of successful hit and total time on the task were provided as performance feedback after each block.

Outcome Measures

A preset benchmark block of the pinch task (Table 1) and the behavioral assessments including the Fugl-Meyer Assessment (FMA) [11] and the Box and Block test (B&B) [12] were conducted pre-and post- training to evaluate the effectiveness of the virtual reality system. Total time on the task and the peak pinch force during the benchmark block were further calculated.

Testing Procedure

Both participants practiced the pinch task for 15 min over three weeks for 12 training sessions. An experienced physical or occupational therapist worked with the participants. The responsibility of the therapist was to choose appropriate

Table 1 Task configuration of the benchmark block

Trial #	Size	Mass	Lift height
1	30	100	50
2	40	100	20
3	20	50	80
4	20	150	80
5	30	150	50
6	40	50	20
7	30	100	20
8	20	50	50
9	40	150	50
10	30	100	80

practice blocks and task parameters for participants. The practice blocks were chosen based on the participants' capability with some challenge. Participants sat in front of the computer screen with their tested thumb and index finger attached to the PHANTOM devices (Fig. 1, left). Before the trials, the 7-cm distance between the thumb and index finger were positioned and calibrated to reconfigure the metrics in the virtual environment. Participants were asked to grasp and lift a virtual cube within 30 s for 10 trials per block. The cube size, cube mass, and lifting height were systematically varied in order to keep a moderate level of difficulty. Performance summary feedback including successful hit and total time was provided after each 10-trial block.

Results and Discussion

The demographic data is shown in Table 2.

Training Performance

The participant poststroke completed 11 training sessions with the virtual reality-based pinch task (Table 3). The participant poststroke attempted an average of 38 trials per session with a 60 % success rate and without complaint of fatigue or

Table 2 Demographic data of study participants

	Age (y)	Sex	Dominant hand	Tested hand	Time after stroke (m)
Stroke	59	Male	Right	Right	24
Control	60	Female	Right	Right	

Table 3 Overall training performance

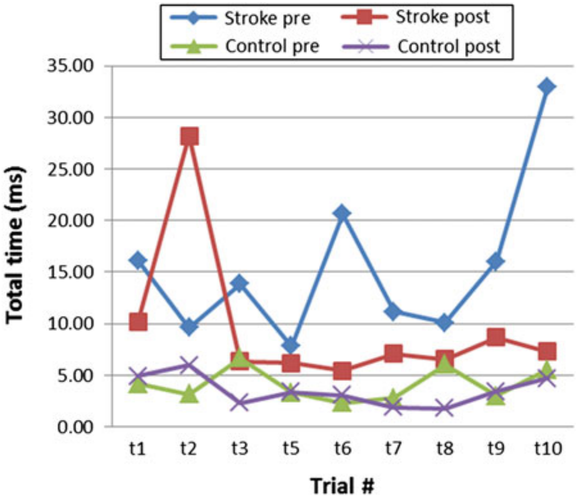
	Training days	Training time (hrs)	Blocks completed	Trials completed	Successful trials	Successful rate (%)
Stroke	11	2.28	48	415	250	60.24
Control	8	2.12	134	1,260	1,252	99.37

pain. Within the same block, time per trial and total time per block were longer for the participant poststroke compared to the control participant (Fig. 2).

Pre-Post Training Effect

After training, the participant poststroke decreased total time per block (Fig. 2) but peak pinch force did not change (Fig. 3). Peak pinch force was significantly greater for the participant poststroke across trials. The findings suggest that the participant poststroke could perform a virtual reality-based pinch task and improve speed of performance with practice without a change in peak force modulation. As for the behavior performance, the participant poststroke improved on the score of the Fugl-Meyer Assessment (FMA) and the total time on the benchmark block (Table 4). The results suggest the effectiveness of the virtual reality-based pinch task for stroke rehabilitation.

Fig. 2 Total time for each trial on the benchmark block pre- and post- training



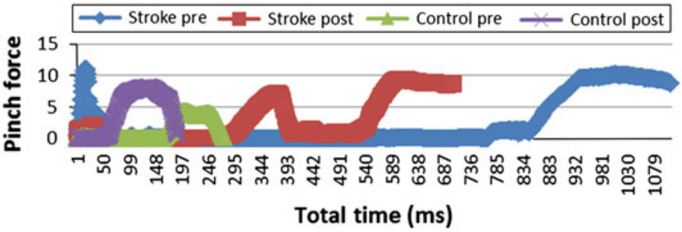


Fig. 3 An example (t3) of peak pinch force pre and post training for both participants

Table 4 Behavior and virtual reality performance data pre-post training

	FMA		B&B		Total time on BM (ms) (9 trials)*	
	Pre	Post	Pre	Post	Pre	Post
Stroke	48	55	17	17	138.37	85.88
Control					36.94	31.42

FMA Fugl-Meyer Assessment; *B&B* Box and Block test; *BM* Benchmark block
*Trial 4 was not included in calculating total time on BM

Table 5 Peak pinch force for each trial in the benchmark block pre- and post- training

		t1	t2	t3	t5	t6	t7	t8	t9	t10	mean
Stroke	Pre	6.27	10.40	8.75	9.97	12.43	11.07	9.76	11.68		9.64
	Post	9.62	8.94	10.81	10.48	12.00	9.48	10.28	12.41	11.23	10.58
Control	Pre	3.52	4.05	2.90	3.39	5.02	4.27	3.70	3.49	5.01	3.93
	Post	7.35	7.47	7.48	8.44	8.04	8.10	7.10	8.92	8.96	7.89

*Trial 4 was eliminated due to programming problems

Pinch Force pre- and post- Training

Peak pinch force in the benchmark block shows in Table 5 and Fig. 3. Pre-training peak pinch force was greater for the participant poststroke than the control participant. Both participants used a higher grip force after training. This may be associated with the concomitant increased lifting speed.

Conclusions

Post-stroke participant could perform a virtual reality-based pinch task and improve speed of performance with practice. A lack of explicit feedback regarding force during task performance may explain the similar peak force after training. Further investigation is warranted to better understand the effect of pinch force regulation during virtual reality-based training for patients poststroke.

Acknowledgments This study was supported by the Interdisciplinary Study of Neuroplasticity and Stroke Rehabilitation (ISNSR), an NIH Exploratory Center for Interdisciplinary Research (Grant # P20 RR20700-01) and the Integrated Media Systems Center, an NSF Engineering Research Center (Cooperative Agreement # EEC-9529152), both at the University of Southern California.

References

1. Rosamond W et al (2008) Heart disease and stroke statistics–2008 update: a report from the American heart association statistics committee and stroke statistics subcommittee. Circulation 117(4):e25–146
2. Andrews K, Stewart J (1979) Stroke recovery: he can but does he? Rheumatol Rehabil 18(1):43–48
3. Sterr A, Freivogel S, Schmalohr D (2002) Neurobehavioral aspects of recovery: assessment of the learned nonuse phenomenon in hemiparetic adolescents. Arch Phys Med Rehabil 83(12):1726–1731
4. Blennerhassett JM, Carey LM, Matyas TA (2006) Grip force regulation during pinch grip lifts under somatosensory guidance: comparison between people with stroke and healthy controls. Arch Phys Med Rehabil 87(3):418–429
5. Blennerhassett JM, Matyas TA, Carey LM (2007) Impaired discrimination of surface friction contributes to pinch grip deficit after stroke. Neurorehabil Neural Repair 21(3):263–272
6. Merians AS et al (2002) Virtual reality-augmented rehabilitation for patients following stroke. Phys Ther 82(9):898–915
7. Crosbie J et al (2005) Development of a virtual reality system for the rehabilitation of the upper limb after stroke. Stud Health Technol Inform 117:218–222
8. Stewart JC et al (2007) Intervention to enhance skilled arm and hand movements after stroke: a feasibility study using a new virtual reality system. J Neuroeng Rehabil 4:21
9. Wade E, Winstein CJ (2011) Virtual reality and robotics for stroke rehabilitation: where do we go from here? Top Stroke Rehabil 18(6):685–700
10. Stewart JC et al (2007) Training reach movements in individuals with hemiparesis: effect of a virtual environment. J Neurol Phys Ther 31(4):192
11. Fugl-Meyer AR et al (1975) The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand J Rehabil Med 7(1):13–31
12. Desrosiers J et al (1994) Validation of the Box and Block Test as a measure of dexterity of elderly people: reliability, validity, and norms studies. Arch Phys Med Rehabil 75(7):751–755

Developing Kinect Games Integrated with Virtual Reality on Activities of Daily Living for Children with Developmental Delay

I-Ching Chung, Chien-Yu Huang, Shyh-Ching Yeh, Wei-Chi Chiang and Mei-Hui Tseng

Abstract Children with developmental delay (DD) often have difficulty in executing activities of daily living (ADL). Although independence in ADL is one of the ultimate goals of rehabilitation for children with DD, ADL training is challenging in the contexts of hospitals and clinics due to a lack of natural settings, and the difficulty in transferring the skills learned in hospitals and clinics to home environment. Kinect games integrated with virtual reality (VR) simulating a home environment can provide a natural environment for effective ADL training on children with DD. Thus, the aim of the study is to develop game-based ADL training tasks using Kinect integrated with VR for children with DD. Kinect games for training purposes are developed. In addition, two pilot studies are conducted for typically developing children and children with DD aged 3–5.9 years respectively to test the applicability of Kinect games. Kinect games integrated with VR designed for ADL training provide opportunities for children with DD to practice ADL in simulated real-life situations, which reinforces the effectiveness of training at clinics and decreases the burn of caregivers in training the child. The efficiency and feasibility of ADL training could thus be improved.

Keywords Kinect games · Virtual realty · Children

I.-C. Chung · C.-Y. Huang · W.-C. Chiang · M.-H. Tseng (✉)
School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei, Taiwan
e-mail: mhtsengster@gmail.com

S.-C. Yeh
Department of Computer Science and Information Engineering, National Central University, Jhongli, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_124, © Springer Science+Business Media Dordrecht 2014

Introduction

The child's independence in activities of daily living (ADL), including basic ADL and instrumental ADL, is crucial for children to gain self-control and could decrease caregiver burden. Children with developmental delay (DD), however, often have difficulty executing ADL, such as toileting and shopping. ADL training is thus important for children with DD and is one of primary goals of rehabilitation. Nevertheless, ADL training is challenging in the contexts of hospitals and clinics. The ADL training setting in hospitals and clinics is different from home setting, resulting in weak generalization to the real world environment. In addition, ADL training at home often causes the tension between children and caregivers.

Kinect games allow players using their whole body to engage in games and interact with scenes and characters in the games. Virtual reality (VR) allows operators perceiving sensory inputs close to the real world so that operators could completely blend into the simulated environment. When ADL training designed in the form of Kinect games integrated with VR, children could use their whole body to practice activities in the real-life context simulated by the computer. The problem of weak generalizability could hence be solved. Furthermore, Kinect games integrated with VR provide interactive-interface, the animation stories with VR, and visual and auditory feedbacks. These elements increase the appealingness of ADL training to children, with resulting enhancement of children's motivation for participating in training activities, and consequently promote the training effectiveness.

Besides, as regards equipment, only a few items are needed for Kinect games, i.e., a computer and a sensor. Such simple equipment exempts caregivers from the burden of preparing ADL appliances in a real-life situation, thereby maintaining the fidelity of caregivers to implement the ADL training at home.

Kinect games integrated with VR have been widely applied to rehabilitation, including stroke rehabilitation [1–5], balance training [3, 6–8], and motor training [1, 9]. All the aforementioned studies reported that participants had significant improvements, indicating that Kinect games integrated with VR are potentially useful tools for ADL training and benefit children with DD. However, no Kinect games integrated with VR to date have been developed for ADL training in children with DD.

The aim of the study is to develop Kinect games integrated with VR for ADL training for children with DD in order to provide opportunities for children with DD to practice ADL in the simulated real-life situations. Since the Kinect games integrated with VR require only simple equipment, the feasibility and efficiency as well as caregivers' fidelity of implementing the training program at home are improved.

Proposed Method

This study has two phases: (1) the phase of developing Kinect games, and (2) the tryout phase of Kinect games.

The Phase of Developing Kinect Games

Equipment

Kinect

The present study utilizes Kinect, a type of interactive motion-sensing systems, as the main device to catch sensory inputs such as body motions and facial expressions and sounds. Kinect sensor detects the human body as the main controller without the needs of additional operational devices. During the operating process, Kinect sensor could record the patterns and the accuracy of the body movement which could be used to record the progress of the child with DD.

Kinect Games

Training purposed Kinect games integrated with VR for ADL training purposes are developed for children with DD. One example of the program is described as follows. For the purpose of shopping in the community, a 360-degree spherical image of a virtual convenience store is provided in the game. Children are asked to walk around the store, search for the assigned goods, pay appropriate amount of money to clerks or get accurate exchange back, finally take the goods and walk out from the store. Other ADL training activities are also developed in the present study, including answering the phone at home and expressing the need for toileting, etc.

Procedure

1. Applying Kinect Sensor and Unity
The program applies Kinect sensor to capture three information, including the colorful image, 3D depth alignment and auditory information. The main tracking information from players is the skeletal information, including the position and the direction. The unity is a powerful rendering game engine integrated with complete set of intuitive tools, including technique of light mapping, a type of rendering path. Users could create interactive 3D content and publish games for multiplatform easily.
2. Writing Scripts for Kinect Games and Developing Kinect Games
Researchers first collect common difficulties or problems encountered by children with DD when they executing activities of daily living (ADL) from clinicians, parents, and teachers. Then, researchers write up stories containing the ADL training tasks targeted at those problems. Besides, each activity is graded by levels of difficulty in order to fit children with DD functioning levels. Scripts are designed for Kinect games through 3D animation and computer programs. Kinect sensor could then be used to play Kinect games.

The Tryout Phase of Kinect Games

A Pilot Study of Kinect Games on Typically Developing Children

Participants

The present study recruit twenty children aged 3–5.9 years and their caregivers. In addition, caregivers should be able to read Mandarin Chinese in order to fill out the Survey.

Instrument

The Survey of Feedback for Kinect games (Appendix A.) is used to evaluate the applicability of the Kinect games. Four areas are accessed in the survey, including difficulties, interestingness, fluency, and comments for improvement. Caregivers complete the survey through observing the child's performance and asking the child's opinions. Researchers then revise the Kinect games based on the survey results. The survey takes about 10 min.

Procedures

1. Cover letters explaining the research project and consent letters are mailed to kindergartens in the greater Taipei area to recruit participants. After caregivers returned the signed consent letters, researchers either set up the computer along with the Kinect sensor in children's home or ask participants to come to the research lab.
2. Children are asked to play Kinect games, and caregivers fill out the Survey of Feedback for Kinect Games.
3. Researchers revise the scripts and program of Kinect games according to opinions from children and caregivers.
4. Repeat step 2 and step 3 until no problems are reported from caregivers or children.

A Pilot Study of Kinect Games: Children with DD

Participants

The study recruits 20 children with DD aged 3–6.9 years and their caregivers. In addition, caregivers should be able to read Mandarin Chinese in order to fill out the Survey.

Instruments

The Survey of Feedback for Kinect Games is also used in this study.

Procedures

1. Cover letters explaining the research project and consent letters are mailed to pediatrics clinics of Department of Physical Medicine and Rehabilitation at hospitals or medical centers, and child developmental centers in the greater Taipei area to recruit participants. After caregivers returned the signed consent letters, researchers either set up the computer along with the Kinect sensor in children's home or ask participants to come to the research lab.
2. The revised Kinect games are provided to children with DD. In addition, the provided Kinect games are just-right challenge activities according to children's abilities and their intervention goals.
3. The following steps are following the same with the procedures on typically developing children.

Discussion

Enhancing the Feasibility of ADL Training Programs and Promoting the Effects of Rehabilitation Therapy

It is easier for caregivers and clinicians to train real-life ADL skills using Kinect games that create simulated real-life situations. The Kinect games can be played both in clinics and at home so that children have sufficient practice opportunities to ensure learning new ADL skills and achieving the optimal rehabilitation effects. Moreover, caregivers are usually unable to fully understand the home program dictated by therapists right after treatment sessions. With Kinect games, it becomes easier for caregivers to carry out home programs for children. Furthermore, with Kinect games caregivers can obtain the progress of children with objective data.

Enhancing Children's Motivation in Participation

Kinect games have been gaining more attention and become more prevalent as a new tool in the field of rehabilitation for children with cerebral palsy due to its fascination. The interactive interface, the animation stories with VR, and the feedback of audio and video effects could enhance the children's motivation, attention, and vitality in rehabilitation therapy to further promote the therapy effects and achieve functional independence.

Recording and Monitoring Progress of ADL Training for Children with DD

The Kinect ADL training system allows therapists selecting appropriate training program according to evaluation results. In addition, children could practice home-based ADL training activities whenever they like to such that therapeutic effects can be enhanced. Besides, with online training program, children's performance on ADL tasks and their progress are automatically recorded in the system, which allows the rehabilitation professionals to obtain the profile of children's ADL skills thus, effective ADL training program will be established.

Conclusion

This is the first study using the Kinect games integrated with VR to design ADL training program for children with DD. There are three advantages of applying Kinect games on ADL training: (1) improve the feasibility of ADL training, (2) enhance the child's motivation on the training, and (3) record and Monitor progress of ADL training. Thus, effective ADL training program will be established by clinicians as soon as possible result in three-win situation of the medical personnel, the caregivers, and the children.

Appendix A. Examples of The Survey of Feedback for Kinect games:

A. Operation

1. Is it easy to find the icons of each Kinect games?
□ Yes, □ No, Suggestion: ______________________________
2. Is the plot interesting?
□ Yes, □ No, Suggestion: ______________________________

B. Satisfaction

Items	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1. I felt satisfied with the content of the training programs.	SA	A	N	D	SD
	The reason is:				
2. My child felt interesting about the Kinect games.	SA	A	N	D	SD
	The reason is:				

References

1. Gallo L (2011) Controller-free exploration of medical image data: experiencing the Kinect. In: Paper presented at the 24th international symposium on computer-based medical system (CBMS)
2. Mouawad MR, Doust CG, Max MD et al (2011) Wii-based movement therapy to promote improved upper extremity function post-stroke: a pilot study. J Rehabil Med 43:527–533
3. Padala KP, Padala PR, Burke WJ (2011) Wii-Fit as an adjunct for mild cognitive impairment: clinical perspectives. J Am Geriatr Soc 59:923–932
4. Saposnik G, Manmdani M, Bayley M et al (2010) Effectiveness of virtual reality exercises in stroke rehabilitation (EVREST): Rationale, design, and protocol of a pilot randomized clinical trial assessing the Wii gaming system. Int J Stroke 5:47–51
5. Yavuzer G, Senel A, Atay MB et al (2008) "Playstation Eyetoy Titles" improve upper extremity-related motor functioning in subacute stroke: a randomized controlled clinical trial. Eur J Phys Rehabil Med 44:237–244
6. Clark R, Kraemer T (2009) Clinical use of Nintendo Wii bowling simulation to decrease fall risk in an elderly resident of a nursing home: a case report. J Geriatr Phys Ther 32:174–180
7. Kliem A, Wiemeyer J (2010) Comparison of a traditional and a video game based balance training program. Int J Comput Sci Sport 9:80–91
8. Smith ST, Sherrington C, Studenski S et al (2009) A novel dance dance revolution (DDR) system for in-home training of stepping ability: basic parameters of system use by older adults. Br J Sports Med 45:441–445
9. Bryanton C, Bosse J, Brien M, McLean J, McCormick A, Sveistrup H (2006) Feasibility, motivation, and selective motor control: virtual reality compared to conventional home exercise in children with cerebral palsy. CyberPsychology Behavior 9(2):123–128
10. Kerrebrock N, Lewit EM (1999) Children in self-care. The future of children pp 151–160
11. Case-Smith J (1994) Defining the specialization of pediatric occupational therapy. Am J Occup Ther 48(9):791–802
12. Dougherty J, Kancel A, Ramer C et al (2011) The effects of a multiaxis balance board intervention program in an elderly population. Mo Med 108:128–132
13. Flynn S, Palma P, Bender A (2007) Feasibility of using the Sony playstation 2 gaming platform for an individual poststroke: a case report. J Neurol Phys Ther 31:180–189
14. Kerrebrock N, Lewit EM (1999) Children in self-care. The Future of Children, 151–160
15. Lange B, Flynn S, Rizzo A (2009) Initial usability assessment of off the-shelf video title consoles for clinical title-based motor rehabilitation. Phys Ther Rev 14:355–363
16. Rand D, Kizony R, Weiss P (2008) The Sony playstation II EyeToy: low-cost virtual reality for use in rehabilitation. J Neurol Phys Ther 32:155–163

Automate Virtual Reality Rehabilitation Evaluation for Chronic Imbalance and Vestibular Dysfunction Patients

Ming-Chun Huang, Shuya Chen, Pa-Chun Wang, Mu-Chun Su, Yen-Po Hung, Chia-Huang Chang and Shih-Ching Yeh

Abstract Dizziness is a major consequence of chronic imbalance and vestibular dysfunction, which prevents people performing their routine tasks and affects their quality of life. It may lead to severe injuries caused by unexpected falling. Medicine treatment can alleviate the syndrome of dizziness and past research shows that dizziness can be further reduced if appropriate vestibular function rehabilitation exercises are practiced regularly. Nevertheless, these exercises are usually time-consuming and tedious because of repetitive motions. Most of patients ceases practicing accordingly and reduce the effectiveness of rehabilitation. In order to encourage patients to be involved in the rehabilitation process, interactive rehabilitative gaming systems are introduced in the recent research. Virtual reality technology is used to enhance the gaming experiences and vision sensors are served as gaming inputs. This paper proposes a series of novel virtual reality games adapted by Cawthorne-Cooksey exercises, which are extensively used in clinical for chronic imbalance and vestibular dysfunction rehabilitation. 32 patients participate in the rehabilitation processes within two months period and their gaming parameters and quantified balance indices are analyzed by the

M.-C. Huang (✉)
Computer Science Department, University of California, Los Angeles, USA
e-mail: mingchuh@cs.ucla.edu

M.-C. Su · Y.-P. Hung · S.-C. Yeh
Computer Science and Information Engineering, National Central University, Chungli, Taiwan, Republic of China

S. Chen
Physical Therapy, College of Health Care, China Medical University, Taichung, Taiwan, Republic of China

P.-C. Wang
Otolaryngology, Cathay General Hospital, Fu-Jen Catholic University School of Medicine, Taipei, Taiwan, Republic of China

C.-H. Chang
Outcomes Research Unit, Cathay Medical Research Institute, Hsichih, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_125, © Springer Science+Business Media Dordrecht 2014

supported vector machine (SVM) classifier. It shows that $\sim 81\ \%$ patients improve their game parameters and balance indices after undertaking the dizziness rehabilitation training compared to the measurement they had. Our clinical observations also reveal that our patients have higher willing and motivation to regularly perform rehabilitation with the proposed system.

Introduction

Chronic imbalance and vestibular dysfunction usually lead to dizziness [1]. Common dizzy types are Meniere's disease, Benign Positional Vertigo, Vertebrobasilar insufficiently, and Vestibular neuritis, and sick headache. Dizzy may persist for minutes, hours, or longer, and thus interrupts people's daily routine [2–4].

Moreover, Dizziness may cause severely dangerous while people holds glass/knife in the kitchen or stands in a bathroom. Steady posture may be difficult to maintain, and the person may experience shaking, tilting, difficulty standing, and falls. Worst of all, the syndrome of dizziness will keep deteriorate if no treatment is applied. Common treatments are vestibular surgery, medicine treatment, and balance training exercises. Past research reveals that the combination of these treatments can effectively reduce the syndrome of dizziness [5]. Compared to surgery and medicine treatment, balance training is the more non-invasive and suitable for most of patients. Cawthorne-Cooksey is a balance training exercises designed to improve patients' ability of keeping balance by training patients' ability to control their eyeball movement, head movement, extremities stretch, and bilateral body balance. Through undertaking repeated training exercises, patients' ability of cervico-ocular reflex should be enhanced and provides better compensation for retaining balance situation during dizzy [5, 6]. Cawthorne-Cooksey vestibular rehabilitation exercises requires the subject move their eyeball without moving the orientation of the head, move the head without changing their body position, rotate and bend the full body with the force of waist, rotate head and should with the force of shoulder and throwing/catching objects in between both hands. These exercises are usually time-consuming and tedious because of repetitive motions. In addition, it is difficult to customize training exercises for individual patients because the differences among the patients are not quantifiable. To overcome the limitation of the traditional training methods, interactive game-based rehabilitation programs and sensor based recording system are introduced [2, 7–10]. Gaming and sensing systems not only motivate patients to exercise more, but also quantified the process of balance training into meaningful balance indices [11, 12]. Virtual reality technology makes gaming more playful and close to the daily life setup. Physical therapists can adjust the difficulty levels and game contents for meeting the need of individual patient. On the other hand, camera based recording system reduces the interference of the training process, but still record training processes in full. Optical tracking technology and skeleton detection technology provided by Microsoft Kinect further provides abundant information of human motion. A full rehabilitation system can provides immediate visual, auditory, and tactile feedback to make patients more involved in the training processes and gain better training effectiveness.

System Architecture

The proposed VR interactive rehabilitation system consists of game-based training system and center of pressure measuring system. 4 training tasks adopt from Cawthorn-Cooksey vestibular rehabilitation exercises: eyeball movement, head movement, extremities stretch, and bilateral body balance. These exercises are transformed into interactive 3-dimensional VR games. The difficulties and contents of these games can be adjusted by physical therapists based on the condition of the respective patient. Indeed, the prototype games strictly follow the guidance of our medical collaborators and are implemented as similar as the functionality of Cawthorn-Cooksey exercises as possible. Patients are expected to wear 3D glasses and receive real-time gaming feedback in terms of visual and audio output. System flowchart (Fig. 1) presents the work flow of the system from user interaction, data collection, feature selection, to SVM classification. Patients should follow the visual and audio instructions to complete the assigned tasks in the games. Their kinematic performance in each session is automatically recorded by the optical tracking system and Microsoft Kinect. This information are served as gaming inputs, performance indices for assisting therapists and programmers to design more appropriate gaming scenarios. Same for the pre/post-test of the center of

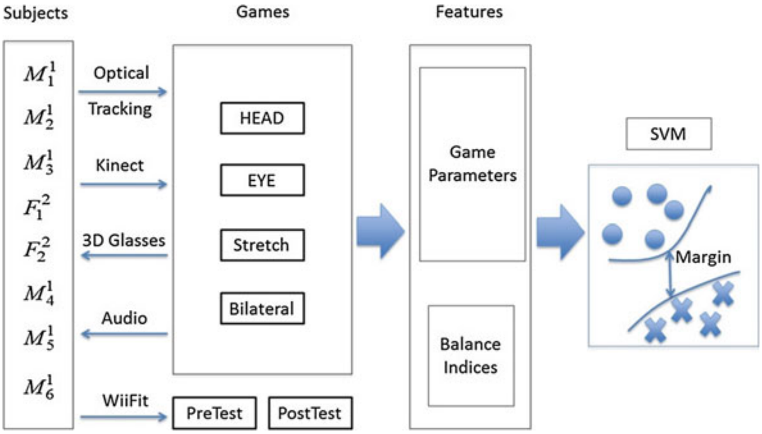


Fig. 1 System flowchart

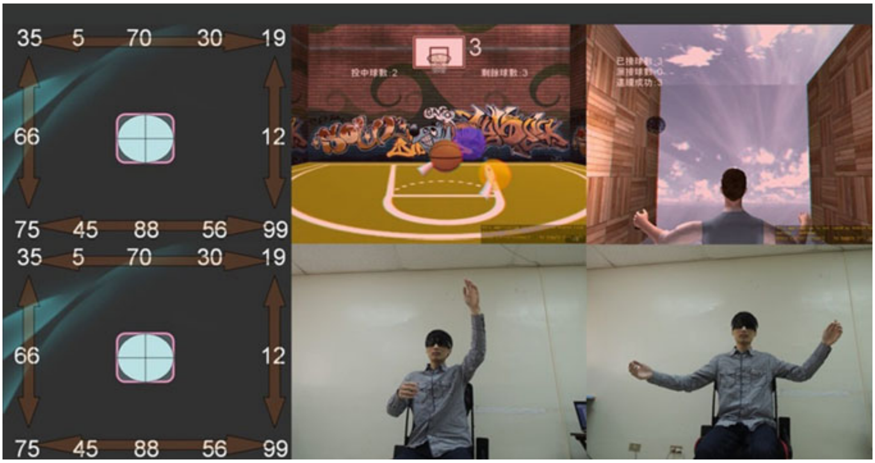


Fig. 2 Four Cawthorn-Cooksey based game snapshots are shown and the taped corresponding subjects' movements

pressure measuring system, center of pressure readings and time stamps are continuously recorded by WiiFit.

Cawthorn-Cooksey Based Games

Four rehabilitative games are designed to simulate the tasks of Cawthorn-Cooksey exercises (Fig. 2). The first game requires users to read numbers randomly generated on the screen. The numbers are shown on the margin of the screen. The goal of this exercise is to train the eyeball movement; therefore, users should keep their head in a static position and move their eyeball in order to read the number on the corner. The ratio of hit or miss is recorded as performance indices. The game for training head movement is similar to eyeball movement training game. Instead, users are asked to gaze on a red box in the screen and read the random number around the box by moving their heads. In order to encourage users to perform stretch exercise, a basketball shooting game is designed. Users should grasp a virtual basketball in the game, lift their arms to a randomly assigned location indicated by a red ball, and the ball will be threw to the virtual basket automatically. During the exercise, users should move their shoulder and stretch their arms to assigned levels. The ratio of hit or miss is recorded as performance indices as well. The last game is for bilateral coordination exercise. Users are asked to capture a ball threw from one hand to the other hand repeatedly. The throwing process will continue until the users miss a catch.

Pre/Post-Test

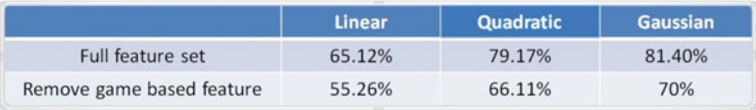
Pre/Post-test center of the pressure data are collected by WiiFit. The trajectory of the center of the pressure can derive statokinesigram (SKG) [13, 14], which is an important index of balance level. The smaller envelope of the trajectory is preferred in our case. Smaller envelope area means that users have less variance in terms of their center of the pressure while they are standing. Together with SKG, maximum X and Y directional placements (MaxML and MaxAP), X and Y directional standard deviations (SDML and SDAP), and average X and Y directional speeds (meanML and MaxAP) can be calculated before and after each game-based training. These indices and the gamed based parameters form a feature for each subject and become an input data for SVM classification.

Binary Supported Vector Machine

Binary linear SVM is a classifier which attempts to partition data set into two categories with the maximum barrier in between. The barrier, formally named as a margin, stands for the width increased by before hitting a data point for both categories. Intuitively, a good separation is achieved by finding the hyper-plane that has the largest distance to the nearest training data point of any class. However, some data set may contain data points which are not linear separable. To overcome this situation with linear SVM, a kernel trick is performed to project those data points into a higher feature dimension; presumably, those data points are projected into a space where they are linear separable. Commonly used kernels are linear, quadratic, and Gaussian kernels. Process data points with SVM classifier is simple if each data point is correctly labeled, such as marked as +1 and -1 for respective group. The normal and offset vector of the hyper-planes can be obtained via training processes. After training, a set of supported vectors can be obtained. For binary linear SVM case, testing data should be projected on to the obtained hyper-plane and multiply with its label to determine the group it belongs to. Accuracy can be computed by analyzing the differences between the results generated by projection from the testing data labels.

Evaluation

32 subjects, including both males and females, participate in a series of VR gaming experiments. Each subject participates in at least 6 trials during approximately two month period. Subjects are required to stand on Wiifit to record their seven indices of center of pressure before and after completing all games in the training exercises. Subjects experience 4 types of Cawthorn-Cooksey based games.



	Linear	Quadratic	Gaussian
Full feature set	65.12%	79.17%	81.40%
Remove game based feature	55.26%	66.11%	70%

Fig. 3 Experimental results of binary linear SVM (*first row* uses all features including game related parameters and center of pressure indices and *second row* uses only balance indices of the center of pressure)

All the kinematics information are recorded by the optical tracking system and Microsoft Kinect. In order to make the rehabilitation process smooth, a copy of the technical manual was provided to assist the therapists in operating the interactive rehabilitation system for each subject. All subjects after completing the sixth trial are invited to join our survey, targeting on analyzing the sufficiency of gaming instructions, game appearance, system usefulness/playfulness, motivation promotion, and the easiness in terms of daily use for all game types: eye movement, head movement, hand stretching, and bilateral balance. In addition, subjects' personal information and their type of dizziness are recorded. Benign Paroxysmal Positional Vertigo, BPPV, Meniere's disease, Vestibular neuritis, vertebrobasilar insufficiency, VBI, and headache induced dizziness.

Recorded game parameters and balance indices from Wiifit are classified by SVM classifier. Linear, quadratic, and Gaussian kernels are applied to understand the structure of the collected data. From the Fig. 3, we can identify that the data are not linear separable. This is because different patients' severity does not equal and their recover speeds are not the same. Some patients get much more improvement during gamed-based rehabilitation and others may get relative less improvement. With Gaussian kernels, we still can achieve ~81 % recognition accuracy, which indicates that most of improvement can be identified with SVM classifier and the differences among all patients are tolerable. We are also interested in verifying that gaming related parameters actually help in classification. Hence, in the second row of the Fig. 3, we remove the features from games and run the training and testing procedure again. We can see that the recognition rate decrease to ~70 % accordingly. It reveals that the gaming parameters provide valuable information as well to diagnose patients' recovery.

Conclusion

This paper proposes a series of novel and interactive virtual reality games adapted by Cawthorne-Cooksey exercises for promoting chronic imbalance and vestibular dysfunction rehabilitation. In order to evaluate the performance of system, collected quantified game parameters and balance index are analyzed by the SVM classifier. It shows that ~81 % patients receive better game parameter balance indices after undertaking the dizziness rehabilitation training compared to the game parameter and balance indices they had. The results of the follow-up questionnaire reveal that using new VR technology for rehabilitation is helpful to promote chronic imbalance and vestibular dysfunction rehabilitation.

References

1. Monsell EM (1995) New and revised reporting guidelines from the Committee on hearing and equilibrium. American Academy of Otolaryngology-Head and Neck Surgery Foundation, Inc. Otolaryngology Head Neck Surgery 1995, vol 113, pp 176–178
2. Gauchard GC, Jeandel C, Perrin PP (2001) Physical and sporting activities improve vestibular afferent usage and balance in elderly human subjects. Gerontology 47:263–270
3. Greenspan SL, Myers ER, Maitland LA, Resnick NM, Hayes WC (1994) Fall severity and bone mineral density as risk factors for hip fracture in ambulatory elderly. JAMA 271:128–133
4. Luukinen H, Herala M, Koski K, Honkanen R, Laippala P, Kivelä SL (2000) Fracture risk associated with a fall according to type of fall among the elderly. Osteoporos Int 11:631–634
5. Schubert MC, Whitney SL (2010) From Cawthorne-Cooksey to biotechnology: where we have been and where we are headed in vestibular rehabilitation? J Neurol Phys Ther 34:62–63
6. Corna S, Nardone A, Prestinari A, Galante M, Grasso M, Schieppati M (2003) Comparison of Cawthorne-Cooksey exercises and sinusoidal support surface translations to improve balance in patients with unilateral vestibular deficit. Arch Phys Med Rehabil 84:1173–1184
7. Adamovich SV, Fluet GG, Merians AS, Mathai A, Qiu Q (2009) Incorporating haptic effects into three-dimensional virtual environments to train the hemiparetic upper extremity. IEEE Trans Neural Syst Rehabil Eng 17:512–520
8. Holden MK (2005) Virtual environments for motor rehabilitation: review. Cyberpsychol Behav 8:187–211
9. Lahiri U, Warren Z, Sarkar N (2011) Design of a gaze-sensitive virtual social interactive system for children with autism. IEEE Trans Neural Syst Rehabil Eng 19:443–452
10. Weiss PL, Katz N (2004) The potential of virtual reality for rehabilitation. J Rehabil Res Dev 41:vii–x
11. Hunag MC, Chen E, Xu W, Sarrafzadeh M (2012) Gaming for upper extremities rehabilitation. Proceeding of conference on wireless health 2012, vol 27
12. Huang MC, Xu W, Su Y, Lange B, Chang CY, Sarrafzadhe M (2012) Smart Glove for upper extremities rehabilitation gaming assessment. Proceeding of conference on pervasive technologies related to assistive environment 2012, vol 20
13. Takagi A, Fujimura E, Suehiro S (1985) A new method of statokinesigram area measurement. Application of a statistically calculated ellipse. In: Igarashi M, Black FO (eds) Vestibular and visual control on posture and locomotor equilibrium, Karger, Basel, pp 74–79
14. Paillard T, Costes-Salon C, Lafont C, Dupui P (2002) Are there differences in postural regulation according to the level of competition in judoists? Br J Sports Med 36:304–305

Part XII
Recent Advances on Video Analysis and its Applications

Machine-to-Machine Interaction Based on Remote 3D Arm Pointing Using Single RGBD Camera

Huang-Chia Shih and En-Rui Liu

Abstract In this paper, we propose a prototype of machine-to-machine interaction system using a single RGBD camera. Based on the 3D arm pointing and tracking algorithm, system enables user to select multiple objects using their two arms, without infrared remote controller and wireless transmitter. A fast gesture recognition method is used to identify the instruction assignment for the target objects. The system consists of three modules, including image capturing module, 3D arm pointing and target tracking module, and interaction module. The image capture module use the RGBD camera installed behind of user to capture the arm parameters in real-world space. The pointing and tracking algorithm is referred to the environmental information to establish the master and slave objects. For interaction module, the command gesture is determined by the gesture recognition, interacting with target devices by the decoded instruction. Finally, system generates a control signal for triggering the master object to communicate with the slave object according to the fetched the instruction.

Keywords Machine-to-machine interaction · Tracking algorithm · Human-computer interaction · 3D pointing · Particle filtering · RGBD camera · Motion capturing

Introduction

With the development of human-computer interaction (HCI) technique is rapidly increasing, we can expect that the computer system allows amounts of interactions around our environment. The fundamental requirement of HCI is to interact with

H.-C. Shih (✉) · E.-R. Liu
Human-Computer Interaction Multimedia Lab, Department of Electrical Engineering,
Yuan Ze University, Taoyuan, Taiwan, Republic of China
e-mail: hcshih@saturn.yzu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_126, © Springer Science+Business Media Dordrecht 2014

an interface, this enables user to control the actions of machine intuitively. Nowadays, the remote controlling has been widely used as the major interface for delivering information and commanding the computerized equipment. The evolution of interactive media changes from the infrared ray (IR), wireless signal, and the natural body control. In recent years, the kinect sensor established by the Microsoft Company that has been attracted a lot of attention and used to real-time capture the human motion parameters.

The conventional control mechanism is based on a remote sensing controller via wired or wireless approach to communicate with devices located at a distance. For example, user tends to control the television, camera, or other electronic devices. In addition, when users send the control signal via remote controller, a target device will execute the corresponding action based on the received control signal such as the commands of turn on/off the power, turn up/down the voice. According to the technology evolutions, the mechanism of HCI changes from direct touching, with a small handset remote controller, and the speech-based and gesture-based manner. Basically, a camera and microphone is used to capture user's voice and gesture.

This study developed a prototype of machine-to-machine interaction based on the arm pointing and gesture recognition captured by a single RGBD camera (i.e., kinect sensor). This provides user to assign machines intuitively and command the interaction between the machines. Figure 1 shows the examples of arm pointing with a single RGBD camera.

System Design

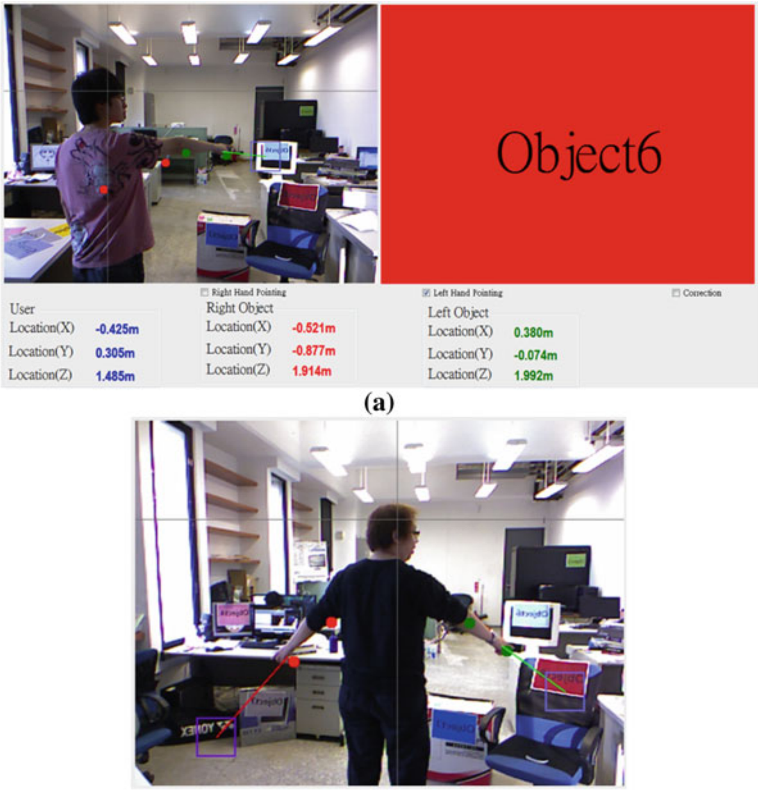
This interaction system consists of three processes, including (1) pointing process, (2) tracking process, and (3) recognition process.

Pointing and Tracking Algorithm

In this paper, the extension line from elbow to hand is used to describe the pointing direction. The depth information provided by kinect sensor applies for initializing object location. The particle filtering [1] algorithm is used to track the target object, it avoids the imperfect depth information resulted from the unstable lighting condition. This instability is complemented using a probabilistic model.

A. *Arm pointing by extension line*

In real world coordination, a pointing unit vector $\boldsymbol{u} = (\mathrm{u_x}, \mathrm{u_y}, \mathrm{u_z})$ formed by the user's skeleton with the elbow point $\boldsymbol{E}$ and hand point $\boldsymbol{H}$. Theoretically, the target object location $\boldsymbol{T}$ satisfies a 3-dimensional line equation along with the unit



□ Right Hand Pointing ☑ Left Hand Pointing □ Correction

User
Location(X) -0.425m
Location(Y) 0.305m
Location(Z) 1.485m

Right Object
Location(X) -0.521m
Location(Y) -0.877m
Location(Z) 1.914m

Left Object
Location(X) 0.380m
Location(Y) -0.074m
Location(Z) 1.992m

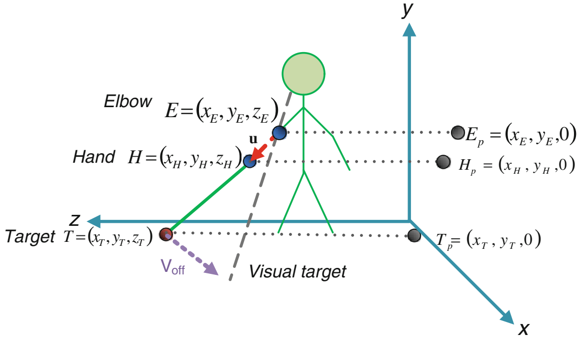
(a)

(b)

Fig. 1 The examples of pointing action with a single kinect sensor behind of user; **a** single arm pointing **b** two arms pointing

vector ***u***. However, due to the stereovision system, it exists a visual error between optical line and pointing line. As shows in Fig. 2, the eyes specify target ***T*** which pointing from the direction of ***u*** is a constant offset higher than the real one.

Fig. 2 The conceptual illustration of 3D arm pointing based on the extension line



Therefore, we need to modify the pointing direction using an offset vector $\boldsymbol{v_{off}}$. To project onto the x-y image plane, we obtain three points include $\boldsymbol{E}_p$, $\boldsymbol{H}_p$, and $\boldsymbol{T}_p$ which denote the corresponding feature points projected from 3-dimensional space. Therefore, we can easily to compute the offset vector from these three points. Based on the depth information, it enables the target object location more consistent with the real pointing direction.

B. *Target tracking by particle filtering*

Particle filtering (PF) [1] is based on the Bayesian theory and the Mote Carlo Sequential sampling method to handle the nonlinear and non-Gaussian problems. The probability propagation model is selected based on prior and likelihood model of every candidate.

The PF approach requires two probabilistic models: the state model and the observation model. The state model describes the evolution of the system from its past state whereas the observation model relates the current state observations to the current state of the system. Based on the observation model, the measurement of the weight for the each particle is an important process for updating the priori. Most of the PF-based visual tracking approaches which applied the color-based histogram similarity manner to achieve the measurement. Based on Bayesian sequential estimation, the prior density function can be computed by two recursive stages: *prediction stage* and *updating stage*.

Gesture Recognition by Ring Projection

We employ the human skeleton toolkit provided by Microsoft [2] to obtain the center position of hand. However it requires user with upstanding pose. In practical applications, when user shows the gesture, they are not always standup stead of siting or lying. To cope this problem, the OpenNI library [3] is used to detect and track the hand.

The goal of the gesture recognition is to recognize the form of hand gestures, supporting for inter-machine communications. Here, the centroid of tracked object is utilized for aligning the hand of commanding. A maximal bounding circle centered by the centroid of circle is obtained. Then the ring projection transformation (RPT) [3] is used to project the commanding gesture into one-dimensional histogram feature map. The cross-correlation applies for identifying the type of gesture.

The RPT is constructed along circular rings of the radius, because the one-dimensional ring-projection model is invariant to the rotation of its corresponding image plane. Figure 3b shows a template image in two distinct orientations and the plots of RPT values as functions of radius r. This template applies for the centroid of hand as shown in Fig. 3a, and computes two ring-projection plots, which are approximately identical, regardless of orientation changes.

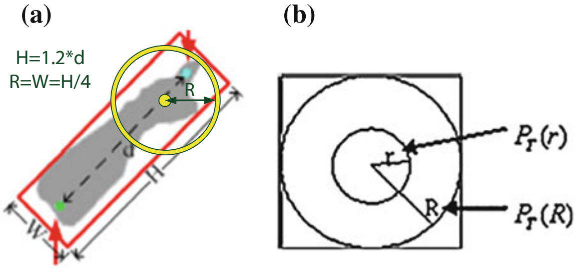


Fig. 3 Gesture recognition using the ring projection

Table 1 The performance of arm pointing

Pose/arm(s)		4 Targets			8 Targets		
		Times	Hits	Accuracy (%)	Times	Hits	Accuracy (%)
Stand up	Single arm (left)	20	18	90	20	17	85
	Single arm (right)	20	19	95	20	18	90
	Two arms	20	16	80	20	15	75
Sit	Single arm (left)	20	17	85	20	17	85
	Single arm (right)	20	16	80	20	17	85
	Two arms	20	15	75	20	15	75

Experimental Results

To evaluate our system, five untrained subjects test the pointing system without any marker and remote controller. Subjects stand and sit to point 4 and 8 targets with their two arms. The results of arm pointing is shown in Table 1. Normally, this system performs the pointing results with single arm, which obtains higher accuracy than it with two arms. Based on the RGBD camera, it allows system to deal with partial occlusion problem using the depth information. In addition, the interference of lighting condition can be suppressed.

Conclusions

This paper proposed a prototype of the machine-to-machine interaction based on the visual arm pointing captured by a single RGBD camera without any remote infrared or wireless transmitter. A fast gesture recognition method used to decode the instruction assignment between the devices. The pointing and tracking algorithm is referred to the space information to establish the master and slave objects.

References

1. Isard M, Blake A (1998) Condensation—conditional density propagation for visual tracking. Int J Comput Vision 29(1):5–28
2. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: Proceeding of the computer vision and pattern recognition (CVPR), pp 1297–1304
3. [Online] OpenNI. http://www.openni.org
4. Tsai DM, Tsai YH (2002) Rotation-invariant pattern matching with color ring-projection. Pattern Recogn 35:131–141

Automatic Peak Recognition for Mountain Images

Wei-Han Liu and Chih-Wen Su

Abstract In this paper, we propose a novel method for automatically recognize the peaks of mountains based on the shape of skyline. Since the appearances of a mountain are variable due to the changes of weather, season or region, the skyline of mountains is extracted for the matching of different mountain images. We use support vector machine (SVM) to predict the possible skyline segments and the linking of incomplete fragments of skyline is formulated as a shortest path problem and solved by dynamic programming strategy. In order to resist the geometric distortion caused by view change, we perform 2D curve matching on the extracted skylines for the peak recognition task. Our experimental results demonstrate that the proposed method is our method is effective for the mountain recognition under complicated and variable circumstances.

Keywords Skyline localization · Mountain annotation

Introduction and Previous Work

In last decade, many state-of-the-art image and visual subject retrieval approaches for place instance recognition [1, 2] and scene category recognition [3, 4] have been proposed. However, the automatic image-based location recognition for

This work was supported by National Science Council, Taiwan, under the Grants NSC101-2221-E-033-062.

W.-H. Liu · C.-W. Su (✉)
Department of Information and Computer Engineering, Chung Yuan Christian University, Chungli, Taiwan, Republic of China
e-mail: lucas@cycu.edu.tw

W.-H. Liu
e-mail: g10077030@cycu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_127, © Springer Science+Business Media Dordrecht 2014

natural scene is still a challenging task. For example, the color and texture of vegetation may be very different due to the changes of seasons, weather and sunlight. It makes appearance-based techniques [1, 2] difficult to find the robust feature descriptors of a natural scene under different circumstances.

In this paper, we consider the problem of recognizing the mountains in the image without geotagging information. Even for the geotagged photograph, only the position of the photographer is associated with the image in the most cases. To precisely annotate the names and positions of peaks in image, we extract skyline to locate the positions of mountains and then perform partial matching process on skylines to find the names of peaks.

Currently, there is a limited number of studies focus on the image-based mountain recognition in recent years. In [5], Babound et al. detected mountains silhouette by compass edge detector and some post-processing steps. They compared the silhouette with a 3D terrain model of the mountains to extract silhouette accurately. The location of each mountain peak can be assigned accurately on images. In [6], Baatz et al. proposed a system for large scale location recognition based on digital elevation models. They extracted the sky and represent the visible horizon by a set of contour words. Then, a bag-of-words like approach with integrated geometric verification was performed for looking the panorama that has a similar frequency of contour words with a consistent direction. Although the above mentioned methods are accurate and reliable, a precise 3D terrain model of the mountains on image is indispensable for matching the silhouette of mountains in their method.

The remainder of the paper is organized as follows. The next section contains the process of skyline extraction. In 2D Curve Matching for Skylines, we introduce the proposed 2D curve matching algorithm for skylines. In Experimental Results, we discuss the experiment results. Conclusion contains some concluding remarks.

Skyline Extraction

In order to extract the skyline automatically, the algorithm we proposed in [7] is adopted. First, we use Canny edge detector [8] to extract edges from an image. The neighboring pixels around the edges are then be collected as 8 × 8 blocks and trained by SVM for finding the possible edge segments of skyline. We extract a 210 dimensional feature descriptor for an 8 × 8 block. Each one contains 192 dimensions for the RGB color 8 × 8 neighboring pixels, 2 dimensions for the variances of illumination aside the edge and 16 dimensions for the location information of edge pixels. All of the features mentioned above are normalized within the interval of [0, 1].

Since edge extraction is always an ill-posed problem, a detected skyline edge segment may be a part of true skyline or just an irrelevant edge. We proposed an edge-linking algorithm to find the bridges for linking the skyline-like edge pixels.

We calculate an energy map by weighting a gradient image and skyline-like edge pixels, which is based on the following equation:

$$e_p = \begin{cases} 0, & \textit{if } p \in S \\ \alpha_p & (1 - G_p + \beta), \quad \textit{otherwise} \end{cases}, \tag{1}$$

where e_p is the energy for the pixel p. S denotes the set of skyline pixels and G_p denotes the normalized gradient value of p. α_p and β are the weights of the energy function of non-skyline pixels.

In order to determine a shortest path with the lowest cumulative energy, we define the following recursive functions for solving the problem caused by steep mountain ridge:

$$E^0_{p(x,y)} = \begin{cases} e_{p(x,y)}, \textit{ if } x = 1. \\ \min(E^f_{p(x-1,y-1)}, E^f_{p(x-1,y)}, E^f_{p(x-1,y+1)}) + e_{p(x,y)}, \quad \textit{otherwise}. \end{cases} \tag{2}$$

$$E^i_{p(x,y)} = \min(E^{i-1}_{p(x,y-1)} + e_{p(x,y)}, E^{i-1}_{p(x,y)}, E^{i-1}_{p(x,y+1)} + e_{p(x,y)}). \tag{3}$$

$$E^f_{p(x,y)} = E^j_{p(x,y)}, \textit{ if } E^j_{p(x,y)} = E^{j-1}_{p(x,y)} \textit{ for all } x. \tag{4}$$

The idea of above recursive algorithm is similar to Seam Carving algorithm [9]. The major difference is the consideration of vertical path in Eq. (4). We incorporate vertical trace into the original Seam Carving algorithm to deal with the problem of steep mountain ridge. For the pixels at the leftmost column on image, $E^0_p = e_p$. Otherwise, E^0_p is the cumulative energy of a shortest path from the leftmost column of pixels to the pixel p. This path does not pass through any pixel at the same column of p. In contrast to E^0_p, the shortest path through the other pixels at the same column of p will be considered in E^i_p, where i denotes the ith iteration. Once the values of E^j_p becomes steady for all the pixels in the same column in jth iteration, E^j_p will be set as E^f_p which is the final lowest cumulative energy from the leftmost column of pixels to current pixel p. Some detection results in [7] are shown in Fig. 1. Our previous experimental results show that approximately 80 % of mountain images' average distance errors are lower than 3 pixels.

2D Curve Matching for Skylines

2D Curve matching is an important research issue for a long time. Mokhtarian and Mackworh proposed Curvature Scale Space (CSS) [10] to represent reliable features of a planer curve on CSS image. Afterwards, numerous CSS-based matching algorithms [11–14] have been proposed for solving the curve matching problem. In this paper, we also use curvature as the feature for the matching scheme, to resist the distortion of view change and skyline localization.

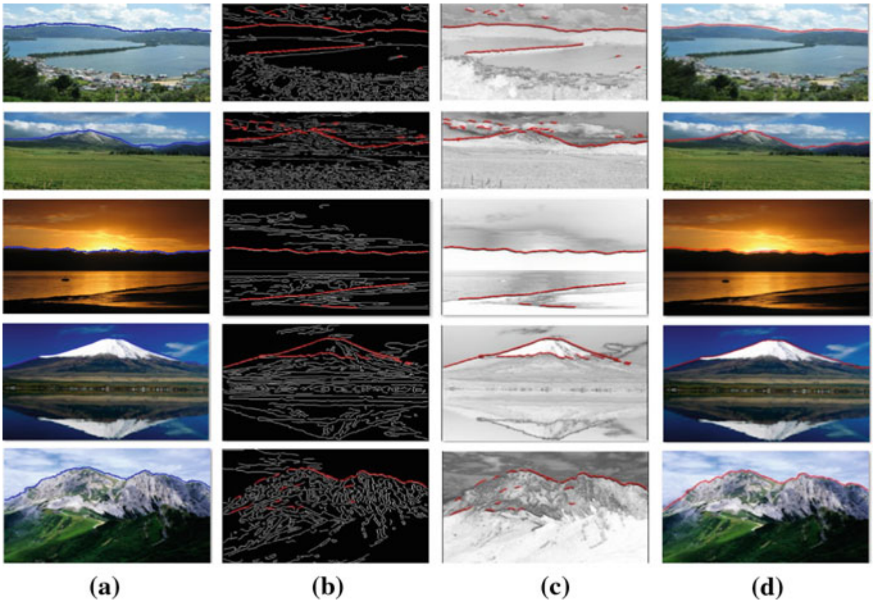


Fig. 1 **a** Original images with ground truth (*blue lines*). **b** The skyline-like edges classified by SVM on edge maps. **c** Energy maps. **d** The final detected skylines (*red lines*). [7]

Before we start the curve matching process, each mountain image is resized respect to the width of image and only a skyline pixel will be sampled for each image column. In other words, the number of skyline pixels is the same as the width of resized image. After the length of skyline has been normalized to N, we calculate the curvature values of the skyline pixels as follows:

$$\kappa(u,\sigma) = \frac{X_u(u,\sigma)Y_{uu}(u,\sigma) - X_{uu}(u,\sigma)Y_u(u,\sigma)}{\left(X_u(u,\sigma)^2 + Y_u(u,\sigma)^2\right)^{3/2}}, \tag{5}$$

$$\begin{aligned} X_u(u,\sigma) &= x(u) \otimes g(u,\sigma) \\ Y_u(u,\sigma) &= y(u) \otimes g(u,\sigma), \end{aligned} \tag{6}$$

where $x(u)$, $y(u)$ denote the position of a skyline pixel on the image and $g(u, \sigma)$ denotes a Gaussian filter of width σ.

Since the scale of a mountain is variable on different images, we perform brute force partial matching to find the best partial match between two skylines. Given a start position P_s and an end position P_e of skyline, we use P_s and P_e to align segments of skylines and each segment is always normalized to N pixels before matching process. On the other hand, since it is not really necessary to consider the rotation problems for mountain images, we discard the alignments with significant rotation over 15°. Then, the similarity between two skylines can be calculated based on the root-mean-square error (RMSE) of curvature value as follows.

$$\min_{i,j}\left(RMSE(A_i, B_j) + RMSE(\tilde{A}_i, \tilde{B}_j)\right), \tag{7}$$

where A_i and B_j are the aligned segments of skyline A and skyline B, respectively. $\tilde{A}_i$ and $\tilde{B}_j$ are the complement sets of A_i and B_j after alignment, respectively. Here, only the pixel that can find its corresponding pixel on the other skyline will be considered in the computation of RMSE ($\tilde{A}_i$, $\tilde{B}_j$). In order to speed up the matching process, we use Fast Library for Approximate Nearest Neighbors (FLANN) [15] to perform fast approximate nearest neighbor searches in high dimensional spaces.

Experimental Results

We have collected 130 mountain images from Internet for the evaluation of our proposed method. For each query image, there are about 10 images in the dataset. Figure 2 shows two examples of the query images and top 3 retrieval results. Although the mountain images are with different fields of view, especially for the last case in Fig. 2d, our proposed method can automatically extract the complete skylines and find best alignment between two skylines. The precision versus recall curve of is shown in Fig. 3. Our method achieved 80 % average precision at 65 % recall.

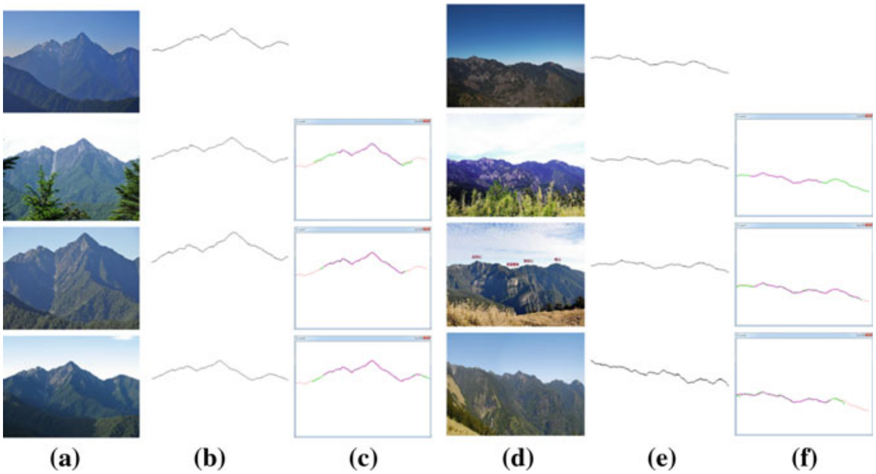


Fig. 2 Query mountain image (*first row*) and retrieved top ranked images. **a, d** Original images. **b, e** Detected skylines. **c, f** The best partial matching segments (*pink lines*)

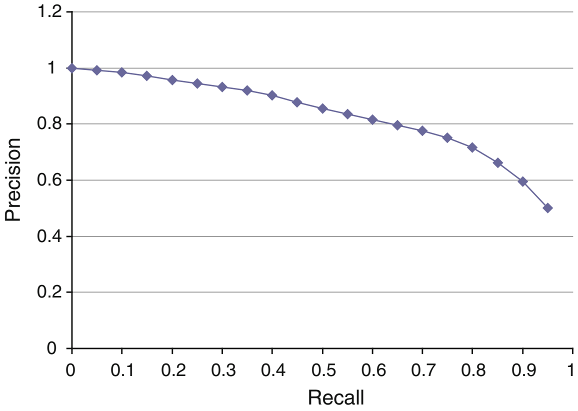


Fig. 3 Precision versus recall *curve* (average over multiple queries)

Conclusion

In this paper, we proposed a systematic method to automatically extract skyline and retrieve mountains images by curve matching. To resist the geometric distortion caused by view change, we used FLANN to find best partial alignment between two skylines effectively. The experimental results showed that our proposed method is accurate and reliable. We would like to explore the further applications of our method on the navigation and annotation of outdoor images in our future work.

References

1. Chum O, Philbin J, Sivic J, Isard M, Zisserman A (2007) Total recall: automatic query expansion with a generative feature model for object retrieval. In ICCV
2. Schindler G, Brown M, Szeliski R (2007) City-scale location recognition. In CVPR
3. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In CVPR
4. Hays J, Efros AA (2008) IM 2 GPS: estimating geographic information from a single image. In CVPR
5. Babound L, Cadik M, Eisemann E, Seidel HP (2011) Automatic photo-to-terrain alignment for the annotation of mountain pictures. In CVPR
6. Baatz G, Saurer O, Köser K, Pollefeys M (2012) Large scale visual geo-localization of images in mountainous terrain. In ECCV
7. Hung Y-L, Su C-W, Chang Y-H, Chang J-C, Tyan H-R (2013) Skyline localization for mountain images. In ICME
8. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8:679–698
9. Avidan S, Shamir A (2007) Seam carving for content-aware image resizing. ACM Trans Graph 26(3):10
10. Mokhtarian F, Mackworth AK (1992) A theory of multiscale curvature-based shape representation for planar curves. IEEE Trans Pattern Anal Mach Intell 14(8):789–805

11. Abbasi S, Mokhtarian F, Kittler J (1999) Curvature scale space image in shape similarity retrieval. Multimedia Syst 7:467–476
12. Mokhtarian F, Abbasi S (2002) Shape similarity retrieval under affine transforms. Pattern Recogn 35(1):31–41
13. Pinheiro AM (2005) Identification of similar shape contours based on the curvature extremes description. In ICIP
14. Mai F, Chang CQ, Hung YS (2010) Affine-invariant shape matching and recognition under partial occlusion. In ICIP
15. Muja M, Lowe DG (2009) Fast approximate nearest neighbors with automatic algorithm configuration. In VISAPP

Gesture Recognition Based on Kinect

Chi-Hung Chuang, Ying-Nong Chen, Ming-Sang Deng and Kuo-Chin Fan

Abstract In recent years, depth cameras have become a widely available sensor type that captures depth images at real-time frame rates. For example, Microsoft KINECT is a powerful but cheap device to get depth images. Even though recent approaches have shown that 3D pose estimation and recognition from monocular 2.5D depth images has become feasible, there are still some challenge problems like gesture detection and recognition. In this paper, we propose a gesture recognition method and use that to make a puzzle game with kinesthetic system. Gesture is very important for our system because instead of using some devices like keyboard or mouse, users will play the puzzle game by their own hand. We will focus on using ROI and some algorithms that we proposed to do gesture detection and recognition.

Keywords KINECT · Depth cameras · Pose estimation · Gesture detection

Introduction

Kinesthetic system is the interactive media system that has been known recently. Because of the immediateness the users gain more feedback during the operation process. By utilizing this characteristic, we integrate this system with teaching resources and make it a digital interactive platform which boost the concentration and interest of students. Hence, it becomes an excellent teaching medium. This research uses Microsoft KINECT cameras and its developed software as the foundation of kinesthetic system. KINECT cameras derive three kinds of data, which are colorful images, 3D image depth information, and audio sources. There

C.-H. Chuang (✉) · Y.-N. Chen · M.-S. Deng · K.-C. Fan
Department of Computer Engineering and EntertainmentTechnology, Tajen University, Pingtung, Taiwan, Republic of China
e-mail: chchuang@mail.fgu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_128, © Springer Science+Business Media Dordrecht 2014

are three cameras in KINECT: the middle one is common RGB color camera and the others are infrared launcher and 3D depth sensor imaged by infrared CMOS cameras. The data sources derived by this system detect the users' motions mainly by 3D depth sensor.

KINECT utilizes the technology of Light Coding to reach the image detecting and tracking. Light Coding [1] is a kind of technology adopted as a way to process depth information of image. This theory used coding of the measured space by Continuous light (close to infrared ray) and decodes by chips to result an image with depth. After understanding how KINECT gets images, the next step is how to process the job of recognition. The data derived from Light Coding technology are basic image information. The key point is to recognize images and transfer into order of action. KINECT can transfer 3D depth image into system of Skeleton tracking. Shotton et al. [2] is the way to integrate the colorful images with depth image to find out body node and the skeleton. Girshick et al. [3] is the procedure of Regression to increase the accuracy of the correspondence between the body pose and skeleton. This system can detect up to six people simultaneously, including recognize simultaneously the motions of two people simultaneously. It can record everyone's twenty sets of details, including body, limbs and finger. Having the approach of skeleton tracing, we can make further application, such as pose estimation [4], action recognition [5], image segmentation [6], body pose reconstruction [7, 8] Building rich 3D maps of environments etc. KINECT developed system indentified many body node from human body, but it didn't identify the detail part of hands. The system can indeed detect the position of palms. However, there is no way to effectively detect or recognize whether the palms are open or close, even the rotation of palms. In order to solve this problem, this system adopts the skeleton tracing system which can gain ROI information of image from palm node. ROI (Region Of Interest) is the area where program designers' interests are, and it is also the core of further image process and recognition. During this phrase, we use OPENNI as the foundation to detect the palm node, and take the sounding area as key palm image to process further recognition and analysis. There are other researches about KINECT palm detection, [9, 10] KINECT is used to detect palm and recognition of palm, palm and fingertip tracking. The hands were segmented using the depth vector and center of palms were detected using distance transformation on inverse image.

This research focuses on this part. We take further palm detection and discuss gesture recognition. Moreover, we can do jigsaw puzzle by corresponding this gesture recognition system with kinesthetic system.

Selection System

The kinesthetic system this research adopted structures its core program on the skeleton monitoring system as Fig. 1. The skeleton monitoring system is through the 3 M depth pictures which are detected by Light Coding method with color

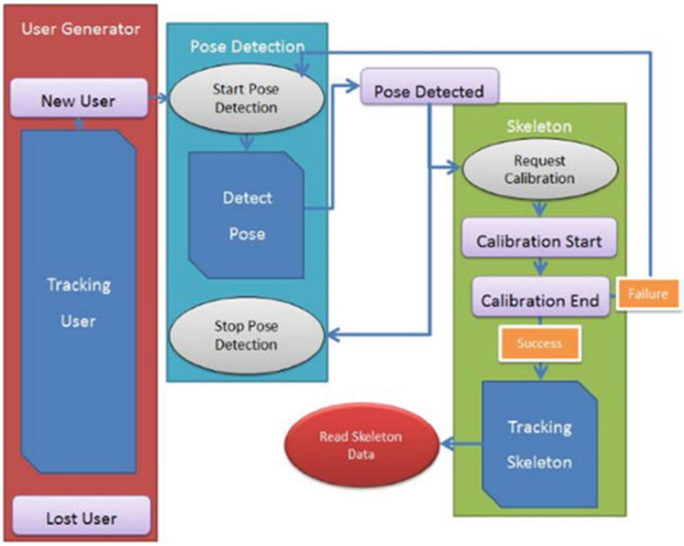


Fig. 1 Skeleton monitoring system

cameras. Once a man are detected, it will process pose detection and recognition on this man. After recognition, it will work on pose correction and track this man until he leaves the scope or the program stops. In the flow chart, the program first starts the man monitoring system. If there is a man detected in the shooting scope, it will start "New User" function and then start pose detection. "New User" and "Lost User" on the last part represent the callback function of two events: "new user detected in the scope" and "user leaving the scope for a while". When these two evens happen, these two functions will be called separately. When a new user is detected and "New User" is called, the program "New User" will call the pose detection unit "Start Pose Detection" to detect a man's pose. When the unit "Pose Detection" detects a user, it will call the callback function "Pose Detected" to process the further move. In this case, "Pose Detected" works on two jobs. One is to check if the numbers of users who is being traced reach the limit; the other is to call the function "Request Calibration", which is a unit processing the skeleton to correct and analyze the body skeleton. Once "Request Calibration" is called, the skeleton process unit will correct and analyze the skeleton. When this unit is working, the skeleton process unit will call the callback function "Calibration Start" to aware the program developer that it starts to process the skeleton correction and analysis. After the skeleton correction is done, it will call the callback function "Calibration End". However, when "Calibration End" is called, it only means the end of the skeleton correction and recognition. It is not sure the skeleton recognition succeeds or fails. If it succeeds, it will go into the next step, which is the function "Start Tracking" to make system to track the skeleton data. If it fails, it will make the pose detection unit to restart to detect the user's pose. After the

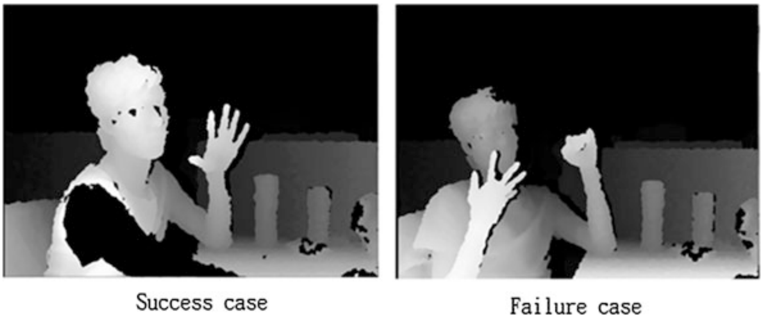


Fig. 2 Palm detection

success of the skeleton correction and start to track skeleton, the user only needs to call the function to read joints data, and gain the latest joint-related information to build a whole body skeleton data. Kinect development system identify many body joints, but no detail identification on palm joints. That is, when the user's palm in the shooting scope, the system can actually detect the position of palms. However, it can't detect or recognize whether the palm is open or close, or even rotates. From Fig. 2, the green dots represent the result of the palm detection. If the palms are under normal circumstances, the result will be complete and the tracking will be normal. But, if the user clenches or rotates his palms, the system will lose the tracking ability. Therefore, the probability of detection will be lower as well. This research mainly aims to make the user be able to do the jigsaw puzzle instinctively. It has to reach the stable detection and recognition on the clenched palms to allow the user to grab virtual jigsaw and move it.

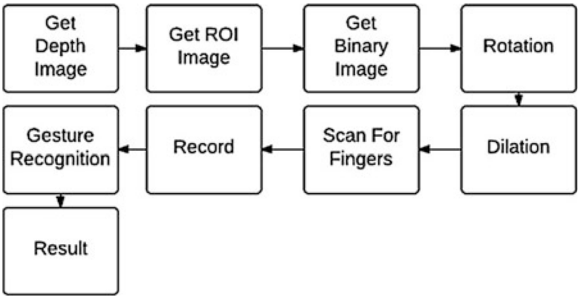
Gesture Recognition

This system enlarges OPENNI development kit and apply it into the recognition of fingers and gesture. The user can play the jigsaw puzzle which this research made by hand gestures through this system. There are more applications about this hand gesture recognition system, such as using Microsoft PowerPoint. The speaker can change and operate the slides only by hand gesture. Besides Microsoft PowerPoint, it can also be integrated with other software. The more researches, the more functions.

The Concept and Procedure

Since Kinect itself is able to track the palm position, we use it to frame the scope of the palms and work on the calculation of the fingertips positions. After getting the data, we use the numbers of fingertips and the angles between them to

Fig. 3 The flowchart of palm recognition



determine the movement of hand gesture. The computer will follow according to different hand gestures. Our calculation method use Kinect's original data to lower the program calculation and space cost. We rotate our palm to proper position in order to get the fingertips position and rotating angels more conveniently and we can save the direction of the palm. The detail calculation is as following Fig. 3.

In the part of image process, we describe separately each phrase according to the flow chart as following:

(a) **Getting ROI image data**: ROI (Region Of Interest) is the area the user interests. It is the core of further image process and recognition. In the phrase, using OPENNI Development environment to find palm as base, we process further recognition and analysis by getting the surrounding area as the key palm image area. However, because the position of palm and waist received by Kinect is quietly unstable, we will not adopt it. (b) **Binary image**: Each point on palm must be close to the palm Depth due to the same hand. Therefore, we frame Palm and compare each point with the palm Depth in the same time. The result like Fig. 4. (c) **Rotation**: The method we mark the finger position needs to rotate the palm to Correct position (palm's up). (d) **Dilation**: This method is to remove flaws due to some image process, such as white spots. (e) **Finding out the fingertips**: the steps

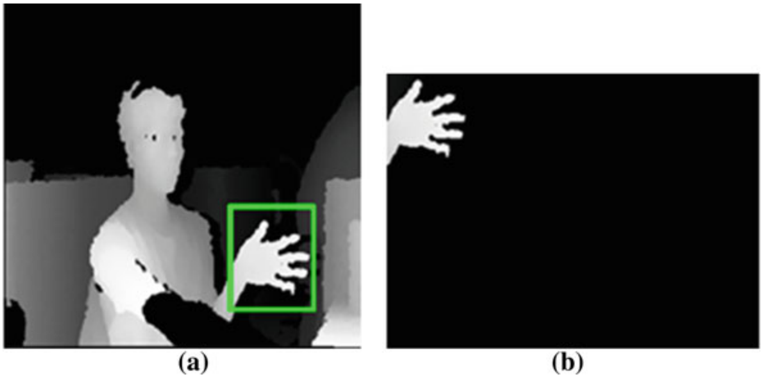


Fig. 4 Getting palm data (a palm tracking (b)palm image

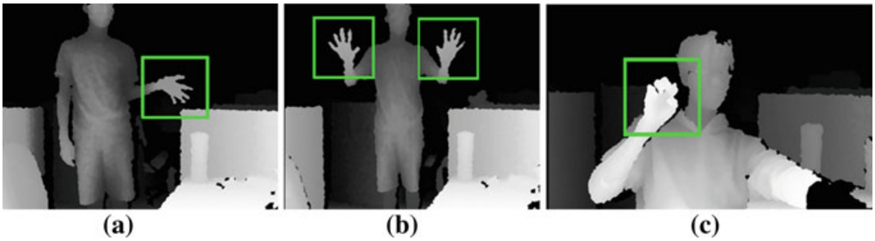


Fig. 5 Results of palm detection (a) one palm is detected (b) Two palms are detected (c) Clasped palm is detected

of finding out the fingertips as follows as: Step 1: We find the number “1”, record it as fingertip (saved in the matrix) and put its position into Stack. Step 2: We take out the top position in the Stack matrix, and search five direction around Site (right, right down, down, left down, left), if the value equal ⌈1⌋, the value set ⌈0⌋, and put it into Stack. Step 3: Into recursive, we repeat Step 2 until clear Stack. Step 4: We repeat Step 1 until the end of the matrix. (f) **Hand gesture determination**: Through the palm and waist points given by OPPENNI and the fingertips found by the methods above, we determine the hand gesture by the position data of points, distance and angle. This research uses it to recognize the movement of user’s palm in the jigsaw puzzle to determine status. When the user grab the virtual jigsaw, the system will recognize the gesture and connect the relative positions of the jigsaw with the coordinates of the actual palm. That will make the user feel like grabbing virtual jigsaw. When the user open his palm, the system will stop connecting jigsaw with palm, which has the effect of unclenching the virtual jigsaw.

Experimental Results

In this paper, using our proposed method of gesture recognition algorithm, With KINECT human skeleton tracking, the system can track and identify parts of the palm to achieve better results. From Fig. 5, the palm is detected very accurate.

Conclusion

The kinesthetic system is the interactive system which has risen recently because of depth image can be captured directly. The user gets more feedback during the operation. The hand gesture detection and recognition plays an important in the kinesthetic system. How to track and then recognize palms in various operation environments has been a critical issue. In this paper, the hand recognition system based on Kinect becomes practical. This method proves a successful calculation method even under various environments. That is the reason why it functions so well in the jigsaw puzzle. This kind of kinesthetic system has not only provided the user with different ways of interacting with a computer, but also drawn the attention of the user. Therefore, the user is able to operate it more fluently and more interactively.

References

1. Albitar C, Graebling P (2007) Christophe DOIGNON robust structured light coding for 3D reconstruction computer vision. International conference on computer vision (ICCV)
2. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: Proceeding of the IEEE conference on computer vision and pattern recognition (CVPR)
3. Girshick R, Shotton J, Kohli P, Criminisi A, Fitzgibbon A (2011) Efficient regression of general-activity human poses from depth images. In: Proceeding of international conference on computer vision (ICCV)
4. Grest D, Woetzel J, Koch R (2005) Nonlinear body pose estimation from depth images. In: Proceeding DAGM
5. Ye G, Liu Y, Hasler N, Ji X, Dai Q, Theobalt C (2012) Performance capture of interacting characters with handheld kinects. In: Proceeding of European conference on computer vision (ECCV)
6. Abramov A, Pauwels K, Papon J, Worgotter F, Dellen B (2012) Depth-supported real-time video segmentation with the Kinect. In: IEEE workshop on applications of computer vision (WACV)
7. Baak A, Muller M, Bharaj G, Seidel HP, Theobalt C (2011) A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Proceeding of international conference on computer vision (ICCV)
8. Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, Kohli P, Shotton J, Hodges S, Fitzgibbon A (2011) KinectFusion: real-time dense surface mapping and tracking. In: Proceeding of IEEE international symposium on mixed and augmented reality (ISMAR), pp 127–136
9. Frati V, Prattichizzo D (2011) Using Kinect for hand tracking and rendering in wearable haptics. World haptics conference (WHC). Dipt. di Ing. dell''Inf., Univ. di Siena, Siena, Italy
10. Raheja JL (2011) Tracking of fingertips and centers of palm using KINECT. Computational intelligence, modelling and simulation (CIMSiM)

Fall Detection in Dusky Environment

Ying-Nong Chen, Chi-Hung Chuang, Chih-Chang Yu and Kuo-Chin Fan

Abstract Accidental fall is the most prominent factor that causes the accidental death of elder people due to their slow body reaction. Automatic fall detection is then an emerging technology which can assist traditional human monitoring and avoid the drawbacks suffering in health care systems especially in dusky environments. In this paper, a novel fall detection system based on coarse-to-fine strategy is proposed focusing mainly on dusky environments. Since the silhouette images of human bodies extracted from conventional CCD cameras in dusky environments are usually imperfect due to the abrupt change of illumination, our work adopts thermal imager instead to detect human bodies. In our approach, the downward optical flow features are firstly extracted from the thermal images to identify fall-like actions in the coarse stage. The horizontal projected motion history images (MHI) features of fall-like actions are then designed to verify the fall by the proposed nearest neighbor feature line embedding (NNFLE) in the fine stage. Experimental results demonstrate that the proposed method can distinguish the fall incidents with high accuracy even in dusky environments and overlapping situations.

Keywords Fall detection · Optical flow · Motion history image · Nearest neighbor feature line

Introduction

Accidental fall is the most prominent factor that causes the accidental death of elder people due to their slow body reaction. Fall accidents usually occur at night with nobody except the elder people if they live alone. It is usually too late to

Y.-N. Chen (✉) · C.-H. Chuang · C.-C. Yu · K.-C. Fan
National Central University, Jhongli, Taiwan, Republic of China
e-mail: 93542021@cc.ncu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_129, © Springer Science+Business Media Dordrecht 2014

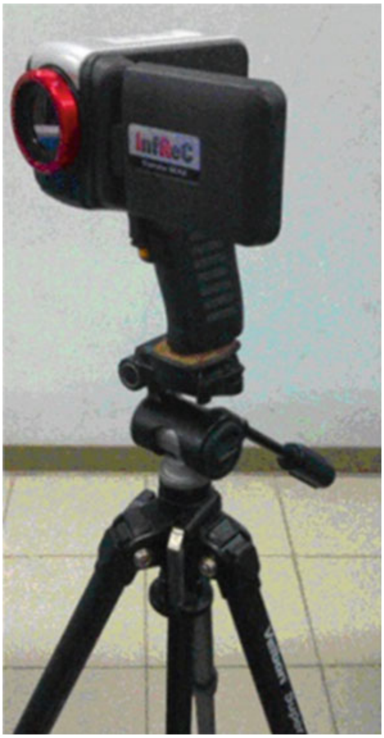


Fig. 1 The thermal imager

remedy the tragedy when the body is discovered hours or days after with the occurrence of accidental fall. In the occurrence of fall incident, humans usually lie flat on the ground. However, we cannot merely use the images to perceive whether this person is lying on the ground. Hence, we have to detect and avoid the risk caused by fall action before the existence of really lying on the ground. According to the survey, a sudden fainting or body imbalance is the main reason to cause a fall. No matter what kind of reasons, fall is a warning that the subject may be in danger. Moreover, if the incidents occur in a dusky and unattended environment, people usually miss the prime time for rescue because the silhouette images of human bodies extracted from conventional CCD cameras in dusky environments are usually invisible or imperfect due to the illumination constraint. To remedy this problem, a fall detection system using a thermal imager (see Fig. 1) to capture the images of human bodies is proposed in this paper. By using the thermal imager, the human bodies can be accurately located even in a dusky environment. For comparison, Fig. 2a shows the images obtained by a CCD camera in a dusky environment, whereas Fig. 2b shows the images obtained by a thermal imager in the same environment. It is obvious that the thermal imagers can extract more clear and intact human bodies in the dusky environments than CCD cameras.

Moylan [1] illustrated the gravity of falls as a health risk with abundant statistics. Larson [2] described the importance of falls in elderly. National Center for

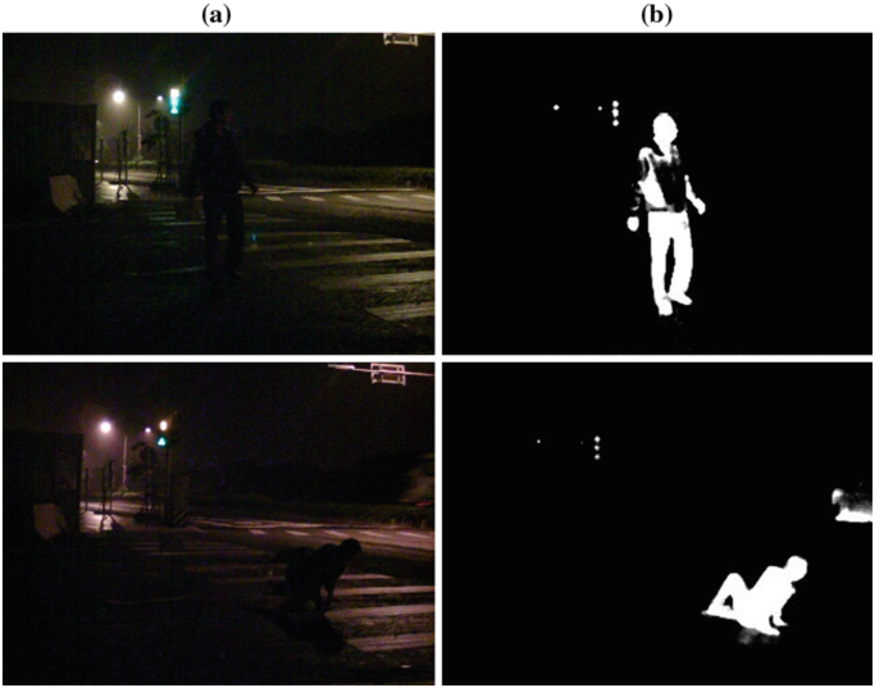


Fig. 2 Image extraction results captured by (**a**) CCD camera, (**b**) thermal imager

Health Statistics showed that more than one-third of ages 65 or older fall each year. Moreover, 60 % of lethal falls occur at home, 30 % occur in public region, and 10 % happen in health care institutions for ages 65 or older [3]. In the literatures of fall detection, Tao et al. [4] applied the aspect ratio of the foreground object to detect fall incidents. Their system firstly tracks the foreground objects and then analyzes the sequences of features for fall incidents detection. Anderson et al. [5] also applied the aspect ratio of the silhouette to detect fall incidents. The rationale based mainly on the fact that the aspect ratio of the silhouette is usually very large when the fall incidents occur. On the contrary, the aspect ratio is much smaller when the fall incidents do not occur. Juang [6] proposed a neural fuzzy network method to classify the human body postures, such as standing, bending, sitting and lying down. In [7], Foroughi et al. proposed a fall detection method using an approximated eclipse of human body silhouette and head pose as features for multi-class support vector machine (SVM). Rougier et al. [8] applied the motion history image (MHI) and variations of human body shape to detect falls. In [9], Foroughi et al. proposed a modified MHI integrating the time motion image (ITMI) as the motion feature. Then, the eigenspace technique was used for motion feature reduction and fed into individual neural network for each activity. Liu et al. [10] proposed a nearest neighbor classification method to classify the ratio of human body silhouette of fall incidents. In order to differentiate between the fall and lying, the time difference between fall and lying was used as a key feature. Liao et al. [11] proposed a slip and fall detection system based on Bayesian Belief Network (BBN). They used the integrated spatiotemporal energy (ISTE) map to obtain the motion measure. Then, the BBN model of the causality of the slip and fall was constructed for fall prevention. Olivieri et al. [12] proposed a spatio-temporal motion feature to represent activities termed motion vector flow instance (MVFI) templates. Then, a canonical eigenspace technique was used for MVFI template reduction and template matching.

In this paper, a novel fall detection mechanism based on coarse-to-fine strategy which is workable in both day and dusky environments is proposed. In the coarse stage, the downward optical flow features are extracted from the thermal images to identify fall-like actions. Then, the horizontal projected motion history images (MHI) features of fall-like actions are used in the fine stage to verify the fall by the nearest neighbor feature line embedding.

The Proposed Fall Detection Mechanism

The proposed fall detection mechanism consists of two modules including human body extraction and fall detection. In human body extraction module, temperature frames obtained from the thermal imager are processed with image processing techniques to obtain intact human body contours and silhouettes. In fall detection module, a coarse-to-fine strategy is devised to verify fall incidents. In the coarse stage, the downward optical flow features are extracted from the temperature images to identify possible fall down actions. Then, the 50-dimensional temporal based motion history images (MHI) feature vector is projected into the nearest neighbor feature line space to verify the fall down incident in the fine stage. To improve fall detection accuracy, complete silhouettes of human body must be extracted to obtain accurate bounding box of human body. To this end, the temperature images captured from a thermal imager are binarized by Otsu's method firstly. Then, the morphological closing operation is employed to obtain a complete human silhouette. Finally, a labeling process is performed to locate each human body in the image and filter out background noises. The process of human body extraction is depicted in Fig. 3. Figure 3a shows the temperature images captured from the thermal imager, Fig. 3b shows the Otsu's binarization results, and Fig. 3c shows the results of morphological closing operation. The bounding box of the human silhouette can be successfully generated after the morphological closing operations.

After the bounding box of human body has been determined, a coarse-to-fine strategy is utilized to verify fall incidents. The purpose of the coarse stage is to identify possible fall actions. Wu [13] had shown that a fall could be described by the increase in horizontal and vertical velocities. Moreover, this work observes that the histogram of vertical optical flows has also demonstrated the significant difference between walking and falling. In our work, a multi-frame optical flow

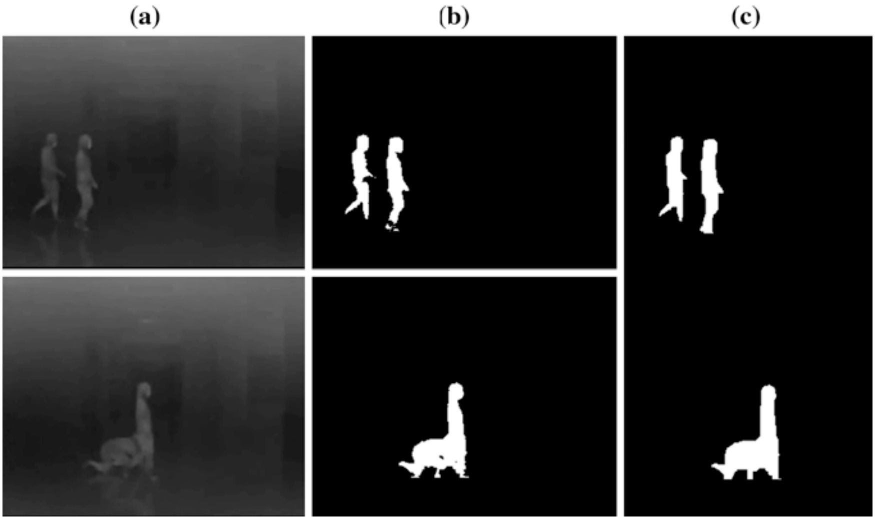


Fig. 3 Human body extraction. **a** Temperature gray level images, **b** binarization results, and **c** morphological closing operation results

method proposed by Wang [14] is adopted to extract the downward optical flow features inside the extracted bounding box in this stage. A possible fall action can be identified by two heuristic rules:

(1) Rule 1: Given 20 consecutive frames, the average vertical optical flows exhibit downward more than 75 % of frames.
(2) Rule 2: The sum of the average vertical optical flows in 20 consecutive frames is larger than a threshold, say 10 in this study.

In the coarse stage, most non-fall actions can be filtered out via the downward optical flow features. However, some fall-like actions are identified as fall incidents due to the swing of arms. To solve this problem, we devise a feature vector which is formed by projecting the MHI horizontally to verify fall incidents in the fine stage. MHI proposed by Bobick [15] is a template which condenses a determined number of silhouette sequences into a gray scale image which is capable of preserving dominant motion information. Since the main difference between fall and other actions is the vertical component changes, our work projects the MHI horizontally to obtain a 50 dimensional feature vector.

In our previous work [16], NFLE has been proven its effectiveness in pattern recognition. However, three problems of the NFLE have also been indicated in [16]. Based on the motivation of mitigating the three problems of NFLE and verify a fall-like action after the coarse stage, a modified NFLE termed Nearest Neighbor Feature Line Embedding (NNFLE) is proposed as a fall verifier in the fine stage. In this paper, we define a fine stage feature vector from a MHI as a sample, In order to overcome the extrapolation and interpolation inaccuracy, feature lines for a query point are generated from the k nearest neighborhood prototypes. The selection strategy for discriminant vectors in NNFLE is designed as follows:

(1) The within-class scatter $\mathbf{S}_W$: The NNFLs are generated from the k_1 nearest neighbor samples within the same class for the computation of the within-class scatter matrix, i.e. a set $F_{k_1}^{+}(x_i)$.

(2) The between-class scatter $\mathbf{S}_B$: Select k_2 nearest neighbor samples in different classes from a specified point x_i to generate the NNFLs and calculate the between-class scatter matrix, i.e. a set $F_{k_2}^{-}(x_i)$.

Experimental Results

In this section, experimental results conducted on fall incident detection are illustrated to demonstrate the effectiveness of the proposed method. In this paper, we compare the proposed method with two state-of-the-art methods. The results are evaluated by using the simulated video data set captured from outdoor scenes. The total number of the video data set is 320. In each video, the environment is in the dusky environments as shown in Fig. 2. The data sets used in this subsection contains only one subject in each video sequence. Two state-of-the-art methods, BBN [11] and CPL [12], are implemented for comparison. The CPL takes a sequence as a sample, whereas the BBN and our proposed method take a frame as a sample. Therefore, the performance comparison of these three methods is based on each video sequence. In the experiments, 60 video sequences of one person are used as training sets and 100 video sequences of one person are used for testing. The performance comparisons of these three methods are tabulated in Table 1. From Table 1, we can notice that the proposed coarse-to-fine strategy of fall detection outperforms the other two methods. It implies that the proposed method is much more effective than the other two methods.

Table 1 The fall detection performance on the data set (%)

Method	Classification action	Reference action (videos)	
		Fall	Walk
CPL	Fall	92.00 (46/50)	8.00 (4/50)
	Walk	10.00 (5/50)	90.00 (45/50)
BBN	Fall	80.00 (40/50)	20.00 (10/50)
	Walk	12.00 (6/50)	88.00 (44/50)
Ours	Fall	**98.00 (49/50)**	**2.00 (1/50)**
	Walk	**0.00 (0/50)**	**100.00 (50/50)**

Conclusion

In this paper, a novel fall detection mechanism based on coarse-to-fine strategy in dusky environment is proposed. The human body in dusky environment can be successfully extracted using the thermal imager and the fragments inside human body silhouette can also be significantly reduced as well. In the coarse stage, the optical flow algorithm is firstly applied on thermal images. Most of walk actions are filtered out by analyzing the downward flow features. In the fine stage, the projected MHI is used as the features followed by a nearest neighbor selection strategy adopted in the NNFLE method to verify fall incidents.

References

1. Moylan KC, Binder EF (2007) Falls in older adults: risk assessment, management and prevention. Am J Med 120(6):493–497
2. Larson L, Bergmann TF (2008) Taking on the fall: the etiology and prevention of falls in the elderly. Clin Chiropractic 11(3):148–154
3. Gs S (1988) Falls among the elderly: epidemiology and prevention. Am J Prev Med 4(5):282–288
4. Tao J, Turjo M, Wong M-F, Wang M, Tan Y-P (2005) Fall incidents detection for intelligent video surveillance. In: Proceedings of the 15th international conference on communications and signal processing 2005, p 1590–1594
5. Anderson D, Keller JM, Skubic M, Chen X, He Z (2006) Recognizing falls from silhouettes. In: Proceedings of the 28th IEEE EMBS annual international conference 2006
6. Juang CF, Chang CM (2007) Human body posture classification by neural fuzzy network and home care system applications. IEEE Trans SMC Part A 37(6):984–994
7. Foroughi H, Aabed N, Saberi A, Yazdi HS (2008) An eigenspace-based approach for human fall detection using integrated time motion image and neural networks. In: Proceedings of the IEEE international conference on signal processing (ICSP) 2008
8. Rougier C, Meunier J, ST Arnaud A, Rousseau J (2007) Fall detection from human shape and motion history using video surveillance. In: Proceedings of the 21st international conference on advanced information networking and applications workshops, vol 2, pp 875–880
9. Foroughi H, Rezvanian A, Paziraee A (2008) Robust fall detection using human shape and multi-class support vector machine. In: Proceedings of the sixth Indian conference on CVGIP 2008
10. Liu CL, Lee CH, Lin P (2010) A fall detection system using k-nearest neighbor classifier. Expert Syst Appl 37(10):7174–7181
11. Liao YT, Huang CL, Hsu SC (2012) Slip and fall event detection using Bayesian belief network. Pattern Recogn 45:24–32
12. Olivieri DN, Conde IG, Sobrino XAV (2012) Eigenspace-based fall detection and activity recognition from motion templates and machine learning. Expert Syst Appl 39(5):5935–5945
13. Wu G (2000) Distinguishing fall activities from normal activities by velocity characteristics. J Biomech 33(11):1497–1500
14. Wang CM, Fan KC, Wang CT (2008) Estimating optical flow by integrating multi-frame information. J Inf Sci Eng 24(6):1719–1731
15. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267

16. Chen YN, Han CC, Wang CT, Fan KC (2012) Face recognition using nearest feature space embedding. IEEE Trans Pattern Anal Mach Intell 33(6):1073–1086
17. Li SZ, Lu J (1999) Face recognition using the nearest feature line method. IEEE Trans Neural Netw 10(2):439–433
18. Yan S, Xu D, Zhang B, Zhang HJ, Lin S (2007) Graph embedding and extensions: general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51

Mandarin Phonetic Symbol Combination Recogniztion in Visioned Based Input Systems

Chih-Chang Yu and Hsu-Yung Cheng

Abstract This paper proposes a Mandarin Phonetic Symbols combination recognition system. This work provides a prototype of a user-friendly and efficient human-centered interface. The system analyzes the fingertip trajectories to obtain the strokes of Mandarin Phonetic Symbols combinations. The trajectories are segmented into real strokes and virtual strokes. The system computes the prior probabilities of different classes of combinations and via Bayes Classifier. The likelihood that each segment is generated by a certain Mandarin Phonetic Symbols combination model is computed via Hidden Markov Models. The experiments have validated the effectiveness of the proposed system.

Keywords Mandarin phonetic symbols combination · Bayes classifier · Hidden Markov models · Human computer interface · Fingertip input

Introduction

In Chinese input systems, Mandarin Phonetic Symbol (MPS) plays a very important role. Mandarin Phonetic Symbols are extracted from Chinese characters. The importance of Mandarin Phonetic Symbols (MPS) and algorithms to recognize single MPS are discussed in [1]. Also, algorithms of locating the fingertip from the segmented foreground mask and tracking the fingertip are elaborated in [1, 2]. To develop a more user-friendly and efficient human-centered fingertip Mandarin input system, recognizing MPS combinations instead of single symbol is desired. There are total 403 different meaningful combinations of MPS. Each combination may consist of single, double, or triple symbols. For convenience, we

C.-C. Yu · H.-Y. Cheng (✉)
Department of Computer Science and Information Engineering, National Central University,
Jungli, Taiwan, Republic of China
e-mail: chengsy@csie.ncu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_130, © Springer Science+Business Media Dordrecht 2014

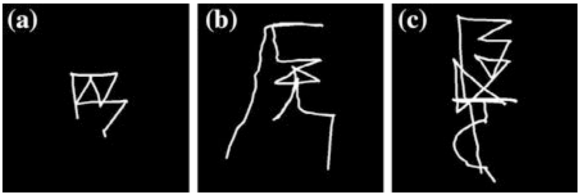


Fig. 1 Three different classes of MPS combinations: **a** single **b** double **c** triple

notate the single, double, and triple combinations as ω_1, ω_2 and ω_3. Figure 1a, b, and c shows an example of ω_1, ω_2 and ω_3 combination, respectively.

Mandarin Phonetic Symbol Combination Recognition

For effective MPS combination recognition, strokes must be analyzed. The strokes acquired by the trajectories of the fingertip include real strokes and virtual strokes. The input symbol combinations exhibit four different types of virtual strokes: entering stroke, leaving stroke, linking stroke, and other virtual strokes within a symbol. The entering, leaving and virtual strokes within a symbol are defined in Fig. 2. Note that in Fig. 2, starting strokes and ending strokes are real strokes. For MPS combinations, a new type of virtual stroke emerge, which is the linking virtual stroke between symbols, as illustrated in Fig. 3. Figure 3a and b show that the entering, leaving, and linking strokes may be very different for the same combination written by different users. In order to recognize the MPS combinations accurately, the entering and leaving strokes need to be eliminated, and the linking strokes need to be correctly located.

The MPS combination recognition procedure is illustrated in Fig. 4. The trajectories are encoded using 8-chain code [3] (Fig. 4a). The turning points are

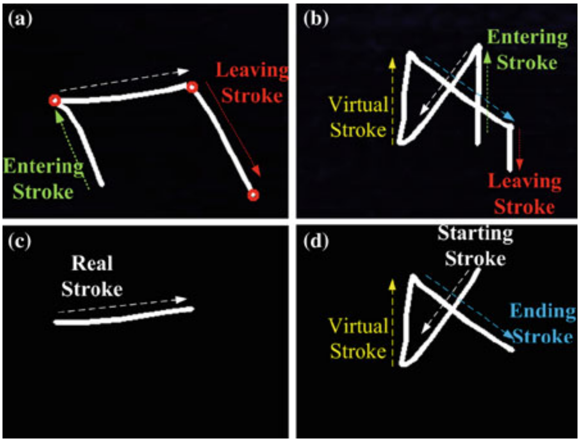


Fig. 2 Entering stroke, leaving stroke, starting stroke, and ending stroke

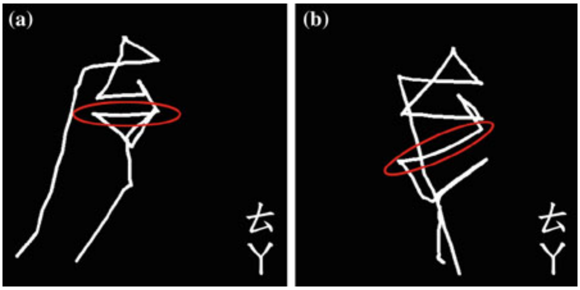


Fig. 3 Linking strokes between symbols

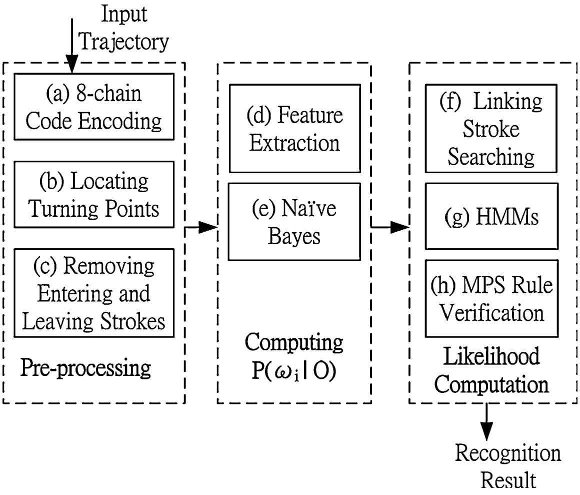

Input Trajectory
(a) 8-chain Code Encoding
(b) Locating Turning Points
(c) Removing Entering and Leaving Strokes
Pre-processing
(d) Feature Extraction
(e) Naïve Bayes
Computing P(ωi | O)
(f) Linking Stroke Searching
(g) HMMs
(h) MPS Rule Verification
Likelihood Computation
Recognition Result

Fig. 4 MPS combination recognition procedure

detected when the difference of two neighboring codes is larger than 2 (Fig. 4b). In other words, the neighboring strokes form an angle which is larger than 90° at a turning point. The trajectory between two turning points is regarded as a stroke. We check all the strokes and eliminate those strokes which are too short. The short strokes are usually noises and provide little contribution for recognition. Hence, we discard these short strokes for better recognition results. After obtaining the turning points, we can remove the entering and leaving strokes, which correspond to the trajectories before the first turning point and the last turning point (Fig. 4c). Some examples of encoded symbol combinations after removing entering and leaving strokes are displayed in Fig. 5. In Fig. 5, the green dots are the detected turning points between strokes.

To be able to recognize the combinations, we need to decide the number of symbols in each combination. Instead of distinguishing them into ω_1, ω_2, and $\omega_3$by making hard decisions, we compute the probabilities that the combination belongs to ω_i given the extracted feature vector $\mathbf{x}$, $P(\omega_i|\mathbf{x})$, $i = 1, 2, 3$. The posterior probability $P(\omega_i|\mathbf{x})$ computed by Bayes classifier (Fig. 4e) is decomposed into three terms as listed in Eq. (13). The prior probability $P(\omega_i)$ is set

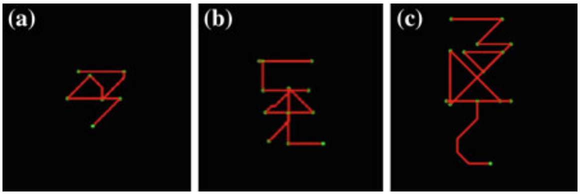


Fig. 5 After removing entering and leaving strokes and encoding the trajectories using 8-chain codes

according to the appearing frequency of samples from class i. The class conditional density functions $P(\mathbf{x}|\omega_i)$ are modeled by Gaussian functions whose parameters are estimated by training samples via maximum likelihood estimation. The term $P(\mathbf{x})$ is independent of class labels and therefore does not affect the decision. We use a Naïve Bayes classifier by making the assumption that all the features are independent. A bounding box is generated for each trajectory after preprocessing. The feature vector $\mathbf{x}$ extracted from the observed trajectory includes the height of the bounding box, the width of the bounding box, the aspect ratio of the bounding box, total number of codes after encoding, number of turning points, and the percentage of each of the eight chain codes in the observed trajectory.

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})} \tag{1}$$

After computing $P(\omega_i|\mathbf{x})$, we eliminate the class with the lowest probability and perform symbol segmentation based on the two remaining classes. For ω_1, there is no need to perform symbol segmentation since there is only a single symbol. For ω_2 and ω_3, linking stroke searching is necessary (Fig. 4f). Under the assumption of ω_2, we hypothesize that the *kth* stroke is the linking stroke. Under the assumption ofω_3, we hypothesize that the k_1*th* stroke and the k_2*th* stroke are the linking strokes. Note that $2 \le k \le M - 1$, and $2 \le k_1 < k_2 \le M - 1$, where M denotes the length of codes after encoding. With these hypotheses, the entire trajectory is segmented into one, two, or three segments. Each segment serves as the observation to compute likelihoods $P(S|\lambda_k)$ using pre-trained single symbol Hidden Markov Models via forward–backward algorithm [4, 5] (Fig. 4g). The single symbol Hidden Markov Model λ_k is trained using the chain code of each symbol via Baum-Welch algorithm [4, 5]. Finally, MPS rules are checked to verify the recognition results (Fig. 4h).

Table 1 Three different sets of symbols

Set	Notation	Symbols
Initial	α	ㄅㄆㄇㄈㄉㄊㄋㄌㄍㄎㄏㄐㄑ ㄒㄓㄔㄕㄖㄗㄘㄙ
Medial	β	ㄧㄨㄩ
Final	γ	ㄚㄛㄜㄝㄞㄟㄠㄡㄢㄣㄤㄥㄦ

For each segment, we compute the likelihoods $P(S|\lambda_k)$ that the segment is generated by the model λ_k of symbol k using forward–backward algorithm. The MPS symbols are divided into three sets as shown in Table 1. According to the rules of MPS, ω_1 can be from the last seven symbols in α, any symbol in β, or any symbol in γ. For ω_2, the combinations of the two symbols can be $\alpha + \beta$, $\beta + \gamma$, or $\alpha + \gamma$. For ω_3, the legal combination is $\alpha + \beta + \gamma$. These rules are helpful for recognition since we can set the likelihoods of illegal combinations as zero directly.

Next, we compute the likelihood that the entire observed trajectory T is generated by a certain model combination. For ω_1, the maximum likelihood of single symbol is selected as the likelihood of the observed trajectory $P(T|\lambda_k)$, as shown in Eq. (2). For ω_2, the likelihood of the observed trajectory $P(T|\lambda_{k1},\ \lambda_{k2})$ is computed using Eq. (3), where λ_i and λ_j are models from legal combinations as described above. For ω_3, the likelihood of the observed trajectory $P(T|\lambda_{k1},\ \lambda_{k2},\ \lambda_{k3})$ is computed using Eq. (4), where $\lambda_i \in \alpha$, $\lambda_j \in \beta$, and $\lambda_l \in \gamma$. In implementation, only top-ranked λ_i, λ_j, λ_j models need to be considered when computing Eqs. (3) and (4). The final decision is made by selecting the combination that contributes the highest weighted likelihood. The weights are the probabilities $P(\omega_i|\mathbf{x})$ computed by the Bayes classifier.

$$P(T|\lambda_k) = \max_i P(S|\lambda_i) \tag{2}$$

$$P(T|\lambda_{k1},\ \lambda_{k2}) = \max_{i,j} P(S|\lambda_i)P(S|\lambda_j) \tag{3}$$

$$P(T|\lambda_{k1},\ \lambda_{k2}, \lambda_{k3}) = \max_{i,j,l} P(S|\lambda_i)P(S|\lambda_j)P(S|\lambda_l) \tag{4}$$

Experimental Results

The experimental dataset for MPS combination recognition contains fingertip input trajectories contributed from 20 t esters. The features for the Naïve Bayes Classifier that computes $P(\omega_i|\mathbf{x})$ are collected from 111 ω_1 combinations, 219 ω_2 combinations, and 164 ω_3 combinations. The single symbol Hidden Markov Models are trained using 20 sets of the 37 MPS symbols. The HMMs are strictly left to right with four hidden states. Figure 6 illustrates examples of linking stroke searching for an ω_2 and an ω_3 combination. The yellow strokes in the figure are hypothesized linking strokes. The wrong hypothesis in Fig. 6a would result in a smaller likelihood than that obtained by the correct hypothesis in Fig. 6b. Similarly, Fig. 6c shows a wrong hypothesis for an ω_3 combination, and Fig. 6d shows the correct hypothesis for the combination. The MPS combination recognition rates from rank 1 to rank 10 are plotted in Fig. 7. An averaged recognition rate of more than 90 % for all possible combinations can be reached for rank 3.

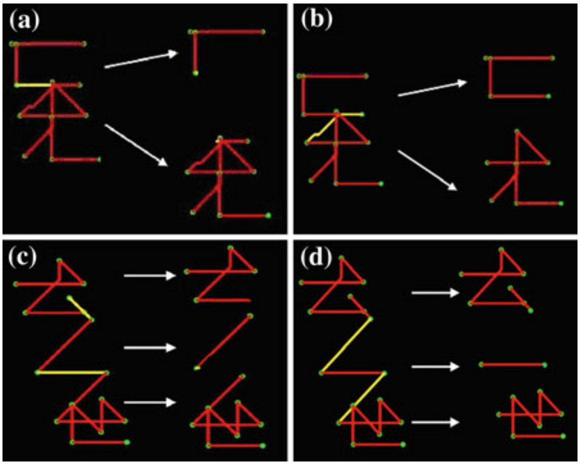


Fig. 6 Examples of MPS combination recognition

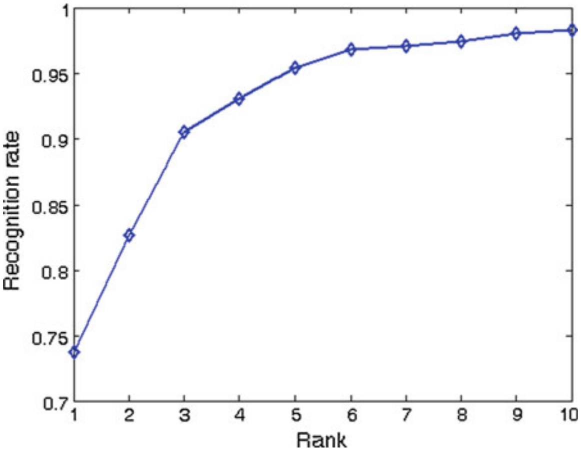


Fig. 7 MPS combination recognition accuracy

Conclusion

In this work, users employ their fingers as the input device instead of using pens or keyboards. Unlike other languages based on alphabets, Chinese character input systems are more challenging because of their logographic nature. The system segments the fingertip trajectories to obtain the strokes. The strokes are further analyzed to search for possible linking strokes between two different symbols in the MPS combination. The system computes the prior probabilities of different classes of combinations and via Bayes Classifier. The likelihood that each segment is generated by a certain Mandarin Phonetic Symbols combination model is computed via Hidden Markov Models. The experiments have validated the effectiveness of the proposed system. The ability to recognize MPS combinations rather than single symbol makes the human computer interface system much more user-friendly.

References

1. Taele P, Hammond T (2008) Geometric-based sketch recognition approach for handwritten mandarin phonetic symbols I. International Workshop on Visual Languages and Computing, pp 270–275
2. National Taiwan Normal University (2004) Practical audio-visual Chinese 1 textbook, vol 1. Cheng Chung Book Company, Ltd, Taipei
3. Yu CC, Cheng HY, Jeng BS, Lee CC, Hong WT (2011) Human-centered fingertip Mandarin input system using single camera. In: The 17th international conference on multimedia modeling
4. Cheng HY, Hwang JN (2011) Integrated video object tracking with applications in trajectory-based event detection. J Vis Commun Image Represent 22:673–685
5. Gonzalez RC, Woods RE (2002) Digital image processing, 2nd edn. Prentice Hall, Englewood Cliffs
6. Rabiner LR, Juang BH (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3:4–16
7. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286

Part XIII
All-IP Platforms, Services and Internet of Things in Future

Supporting Health Informatics with Platform-as-a-Service Cloud Computing

Garrett Hayes, Khalil El-Khatib and Carolyn McGregor

Abstract Recent progression in health informatics data analysis has been impeded due to lack of hospital resources and computation power. To remedy this, some researchers have proposed a cloud-based web service patient monitoring system capable of providing offsite collection, analysis, and dissemination of remote patient physiological data. Unfortunately, some of these cloud services are not effective without utilizing next-generation hardware management techniques. In order to make cloud based patient monitoring a reality, this paper shows how leveraging an underlying platform-as-a-service (PaaS) cloud model can provide integration with web service patient monitoring systems while providing high availability, scalability, and security.

Keywords Artemis cloud · Artemis framework · Cloud computing · Health informatics · Platform as a service · Remote patient monitoring

Introduction

As technology continues to develop in our increasingly connected world, previous non-network enabled medical devices have begun to join hospital networks around the world. Thanks to recent medical device improvements, hospitals are now able to remotely collect patient data from bedside monitoring devices. The collection

G. Hayes (✉) · K. El-Khatib · C. McGregor
University of Ontario Institute of Technology, Oshawa, ON, Canada
e-mail: Garrett.Hayes@uoit.ca

K. El-Khatib
e-mail: Khalil.El-Khatib@uoit.ca

C. McGregor
e-mail: Carolyn.McGregor@uoit.ca

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_131, © Springer Science+Business Media Dordrecht 2014

and analysis of this data has provided many opportunities for clinicians and researchers to discover new condition onset behaviours that are evident in a collection of physiological data streams before the current methods of detection for a range of conditions in critical care. Data collected from patients worldwide gives us the ability to extrapolate observations based on populous and historical data.

In order to support researchers using data collected from bedside monitoring devices, a number of researchers have proposed a cloud-based and offsite web service framework capable of collecting and analyzing patient data worldwide [1–4]. Unfortunately, underlying traditional technical infrastructure is incapable of providing a robust underlying platform that is highly available, expandable, and secure.

In this paper we will provide an overview of cloud computing, touch on a specific cloud framework for health informatics, and show how leveraging platform-as-a-service can provide a robust underlying architecture for such a deployment.

Cloud Computing Overview

The concept of cloud computing generally refers to the delivery of one or more infrastructure components—computational, storage, or otherwise—as a scalable and reliable location-independent service. These services, often located offsite, are exposed to a consumer via pre-set client interfaces [5]. Such interfaces may include web browsers, standardized APIs, and other types of Internet-ready protocols. The location of a client's provisioned cloud resources and data, unbeknownst to them, may span multiple servers, datacenters, or even countries.

Encompassing many service models, cloud computing allows the quick provisioning of resources regardless of underlying hardware or software components. Thanks to this client-centric model, enterprises are able to deploy or request additional infrastructure resources on demand, further streamlining their already complex IT operations.

The underlying amalgamation of various servers, storage area networks, and network hardware is conformed into a single standardized service interface referred to as a cloud. The resulting cloud provides the ability to elastically provision, span, and optimize resource and storage over multiple pieces or hardware while ensuring high availability and robust security.

Although the cloud concept seems esoteric by nature, the cloud model is being used to support many services on the web. Some of these services include popular products like Gmail, Facebook, Google Drive, Dropbox, Amazon AWS, Microsoft Office Web Apps, and many more.

While the majority of well-known cloud service deployments are Internet based (i.e. public clouds), in some instances enterprises may leverage the cloud architecture to consolidate resources and improve overall datacenter efficiency. In addition, when local resources are not cost efficient or sufficient enough for large projects, hybrid cloud models may be used to provision additional computation power on the fly. The four existing cloud models are as follows [5]:

1. Private Cloud: operated by a private company.
2. Community Cloud: operated by more than one company, often working together under a common agreement or area of interest and sharing a single cloud computing instance.
3. Public Cloud: hosted by an external cloud provider whom offers cloud provisioning services to all types of customers.
4. Hybrid Cloud: when two or more clouds are merged together, often through a standardized API interface, to provide cross-datacenter provisioning services on the fly.

Regardless of the cloud model used in a cloud deployment, various types of service models can be presented to administrators. Each service model aims to provide a specific set of resources pertinent to the requirements of the cloud consumer. These cloud service models are as follows: (1) Infrastructure-as-a-Service, (2) Platform-as-a-Service, (3) Software-as-a-Service, and (4) Data-as-a-Service.

Cloud Computing Advantages in Health Informatics

Cloud computing affords many advantages over traditional data center and hardware deployments. Benefits vary from increasing service availability to even reducing data center power consumption. The various benefits of cloud computing, particularly in health informatics, are as follows:

1. Accessibility
2. Availability
3. Resource Consolidation and Multitenancy
4. Reliability
5. Scalability
6. Security.

Cloud Computing in Healthcare

The applications of cloud computing are not just limited to providing virtual machine instances for servers and other Internet facing services. Rather, the cloud model can be applied to any type of datacenter, regardless of its operational goals. One interesting application of cloud computing is in health informatics.

Current cutting-edge health informatics research projects aim to discover new condition onset behaviours that are evident in physiological data streams earlier than traditional detection of conditions in critical care data [1]. To do this, some hospitals may participate in pilot programs that aim to collect real-time patient data from network enabled monitoring devices. This collected data is then analyzed to extract relevant temporal behaviours and usually stored for future data mining and analysis operations. Naturally, not all hospitals may have the capacity—in terms of networking, computational resources, and information technology support staff—to fully support such pilot projects. In this section we will look at a case study implementing a neonatal pilot research project and show how the use of off-site platform-as-a-service cloud computing can help maximize project efficiency both inside the participating hospitals and out.

Current Cloud Applications in Health Informatics

Although cloud computing is still relatively new in the health informatics field, recently there have been a few cloud technology deployments in North American healthcare facilities. In addition, the advent of cloud-based services has created a fundamental shift in the way practitioners meet with patients and manage their electronic health record data. Below we will discuss two significant cloud deployments in the ever-growing health informatics field.

One of the most interesting recent cloud deployments is the Minnesota Tele-Health system. The Tele-Health healthcare network in Minnesota, USA leverages cloud-based medical consultation communications in their hospitals to provide secure remote video conferencing between patients and doctors [6]. This system allows patients to meet with doctors on a flexible, on-demand schedule regardless of physical location. This cloud-based Tele-Health system is capable of prioritizing calls, integrating with human translators when needed, and providing routine medical information by amalgamating existing hospital services and employees. In addition, Tele-Health conforms to the HIPAA standard thanks to the implementation of robust cryptographic controls [6].

Focusing more on health practice management, CareCloud provides a cloud-based framework for managing, analyzing, and charting patient data [7]. Their secure HIPAA compliant cloud service provides a scalable electronic healthcare record solution for healthcare practitioners, allowing them to focus on their patients instead of managing health data [7]. This system allows practitioners to add, edit, and share patient data using a web-based cloud service that provides secure, redundant, and offsite health record storage [7]. Furthermore, CareCloud provides a community connection service that allows practitioners to meet virtually with their patients while sharing selective E-Health Record (EHR) data [7].

As we can see, both Tele-Health and CareCloud have provided innovative cloud-based services capable of streamlining existing healthcare practices while still conforming to HIPAA standards. However, those research projects do not support the transmission of high frequency physiological data streams in real-time for the provision of advanced clinical decision support remotely to support the Service of Critical Care.

Artemis Framework

The Artemis framework provides multidimensional real-time analysis of high-speed physiological data and also support advanced clinical research [1]. McGregor's platform showed the plausibility of supporting online real-time monitoring and analysis of clinical data to detect onset physiological conditions [1]. Collected data may be used to provide extensive insight into novel discoveries regarding physiological data for earlier onset detection of a range of developing physiological conditions.

The Artemis framework, seen in Fig. 1, consists of five phases which, when acting together, are capable of providing in-depth online analysis of developing physiological conditions in patients at participating hospital [1].

The first phase facilitates client data acquisition. A medical data hub, located within each hospital, is responsible for collecting real-time clinical data from network-enabled medical devices onsite. In addition, medical data hubs pair real time patient data with existing e-health data from the hospital's Clinical Information System [1].

Once all relevant data has been collected—in real time- the data is pushed to the online analysis portion of McGregor's proposed Artemis framework. Patient data is streamed in real time to the Artemis software-as-a-service web interface for analysis. Collected data is then analyzed with IBM's Infosphere Streams software to apply appropriate data analysis algorithms [1]. Leveraging the Infosphere

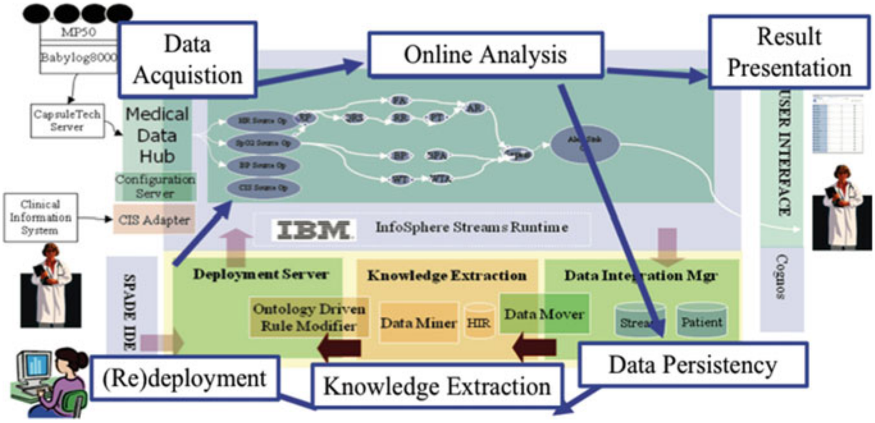


Fig. 1 The artemis framework [1]

Streams software, “enables the real-time deployment of clinical rules representing correlations between behaviours of interest in physiological streams for each condition that is being monitored” [1].

Knowledge extraction of stored data is then done via Service based Multidimensional Temporal Data Mining [3]. This can be performed on both real time and historical patient data. Extracted knowledge is then ready to be presented to researchers for further analysis. Finally, data is stored persistently for the long term to assist future knowledge extraction processes [1].

Artemis Cloud

In order to further the effectiveness of Artemis, McGregor proposed a cloud-based software-as-a-service model that would allow hospitals to interact with the Artemis framework by consuming various web services [1]. Furthermore, hospitals would have persistent access to long-term data and knowledge extraction stored in the Artemis cloud, also accessible using a data-as-a-service storage service [1]. This can be seen in Fig. 2.

Using various proposed web services, hospitals would be capable of initializing an IBM Infosphere Streams environment with custom rule sets and knowledge extraction algorithms [1]. Patient data can then be streamed into the appropriate web interface, encoded in the proposed XML protocol format [4], and coupled with existing patient data in the HL7 V3 format. Ultimately this setup would allow

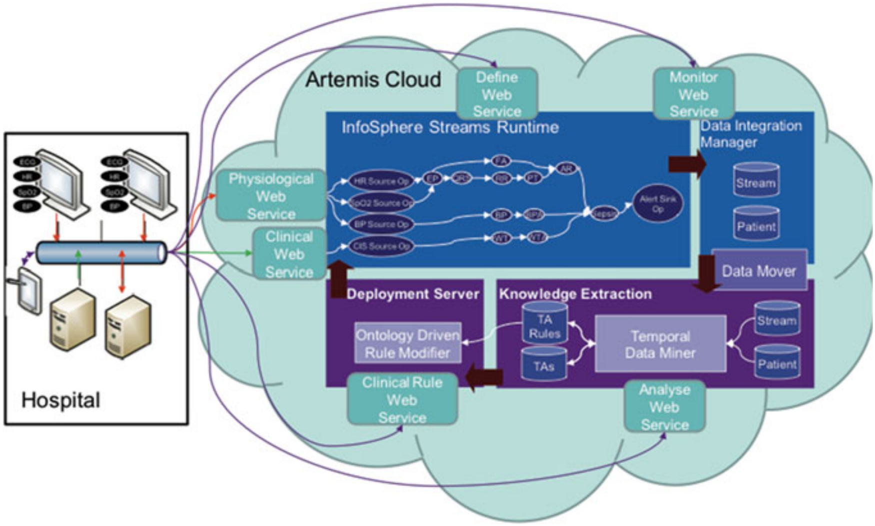


Fig. 2 The artemis cloud framework [1]

clinicians to deploy their own knowledge extraction rules to effectively monitor patient changes in real time and perform extensive clinical research.

Although this framework seems to fit the needs of both researchers and clinicians, McGregor states the "plausibility of using cloud environments for robust healthcare monitoring systems is unknown" [1].

Providing PaaS for Artemis Cloud Deployments

As we have seen in the above Artemis cloud framework, each participating hospital must be able to provision a private Infosphere Streams instance capable of conducting knowledge extraction through customized rule sets. The real challenge of successfully deploying the Artemis cloud lies in the provisioning of underlying resources needed by each hospital. If each component of the cloud architecture in Fig. 2 were run entirely on individual servers, eventually computational resources would be exceeded. In addition, the provisioning and addition of extra hardware resources would be difficult and could result in unneeded downtime. In this section we will define the requirements for an Artemis cloud deployment and show how the use of an underlying platform-as-a-service model can be effectively leveraged by McGregor's proposed web service framework.

Requirements

The underlying PaaS deployment leveraged by the Artemis cloud framework must meet various requirements to be effective. These range from high availability to stringent security policies. The most fundamental set of requirements is as follows:

1. Hospitals must be able to securely interact with the Artemis web endpoint without worrying about data leakage or plaintext data transfers. This can be done using various virtual private network (VPN) technologies. VPN security settings should be chosen based on regional data privacy requirements.
2. Each hospital must have access to its own instance of IBM's Infosphere Streams software via their Artemis web service interface.
3. Each hospital's patient data must be isolated as per regional regulatory standards and must be stored long-term for later knowledge extraction. Long-term data storage should have high redundancy and zero loss, even during hardware failure.
4. The Artemis cloud must suffer no downtime. 99.999 % availability or better should be provided at all times, regardless of hardware failure.
5. Underlying physical resources can be added or removed at will with zero downtime. This ensures the expandability of computation and data storage resources in the future.

6. The exploitation of a single hospital's web service should under no circumstances provide access to, or affect, another hospital's patient data.
7. The Artemis web service should be capable of interfacing with underlying platform-as-a-service components using a standardized API. This facilitates the deployment of Infosphere Streams and customized rule sets via the Artemis web services interface.

High Level PaaS Architecture

In order to meet the requirements outlined above, we first analyzed the deployment requirements for each hospital leveraging the Artemis cloud. After breaking down underlying hardware and software components, we concluded that each hospital's web service deployment constituted only four underlying components:

1. A VPN endpoint capable of creating a layer-2 (or higher) interface between a hospital's medical data hub (MDH—see Fig. 1) and the Artemis data captor software.
2. A data hander interface capable of collecting real time data from a hospital's MDH, isolating it, and passing it to the hospital's IBM Infosphere Streams software.
3. An isolated instance of IBM's Infosphere Streams software—known as the *Online Analysis* component—which can be used to extract relevant information from patient data.
4. A shared centralized DB2 database providing isolated and long-term data storage for each hospital.

This logical architecture can be seen in Fig. 3.

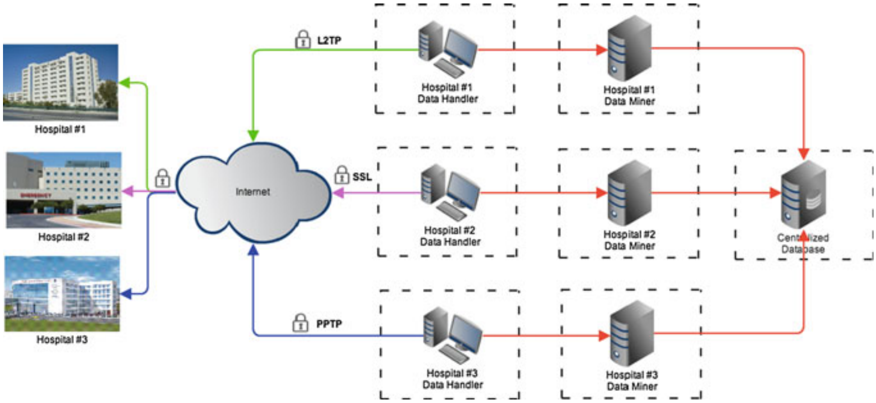


Fig. 3 Logical PaaS deployment model

As we can see, each hospital can be deployed via an underlying set of PaaS resources capable of fulfilling the software requirements needed by the Artemis web interface. Once underlying resources have been deployed, the Artemis cloud can interact with each hospital's resources and provide a high-level and standardized web service interface for each deployment.

Underlying PaaS Architecture

In order to provide the ability to deploy the logical PaaS model outlined in the previous section, we were tasked with selecting the underlying software and hardware capable of meeting the requirements outlined in Sect. 5.1 of this paper. Of upmost importance was ensuring the selected underlying software was capable of utilizing a plethora of hardware in multi-vendor environments while providing a standardized API interface capable of interacting securely with virtualized resources.

After much research, we decided to leverage VMware's vSphere hypervisor to meet our underlying software requirements. vSphere, a low level hypervisor software, allows us to pool any number of hardware resources into a logical cluster capable of providing high availability and disaster recovery [8]. This virtualized datacenter is easily managed and monitored via VMware's vCenter product and is able to provision Artemis cloud resources while providing a robust and mature API interface for interacting with said components [8].

The vSphere product, when coupled with vCenter can provide [8]:

1. High availability in the datacenter, capable of avoiding downtime even during hardware failure.
2. The ability to add or remove hardware resources at any time with zero downtime.
3. Deployment of virtualized Artemis assets extremely fast based on pre-defined service templates.
4. Access to Artemis resources using robust and well test API interfaces.
5. Automatic provisioning of hardware resources to ensure a single hospital does not utilize all computation resources.
6. Host isolation capable of reducing the attack surface of the Artemis cloud and ensuring hospitals are 100 % isolated from all other internal components.
7. Pooling of underlying hardware and storage resources to prevent unexpected data loss, even during hardware failure.
8. Snapshotting of instances in time to facilitate system roll-back, recovery, or knowledge extraction.
9. A simple and secure administrative web interface capable of safely manipulating and monitoring underlying hardware resources in real time.

As we can see, the vSphere and vCenter product exceed our requirements for an underlying PaaS platform that can be leveraged by the Artemis cloud framework.

Conclusions and Future Works

Thanks to recent advances in cloud technologies, like VMware's vSphere and vCenter products, it is possible to provide a robust and security underlying PaaS framework that can be leveraged by the Artemis cloud proposed by McGregor. Such a PaaS architecture can ensure high availability and seemingly infinite resource expansion capable of facilitating real time analysis of remote patient data. Furthermore such a robust architecture, when coupled with McGregor's Artemis cloud design, can be used to provide effective and secure web services that can be leveraged by hospitals worldwide.

Currently the Artemis cloud architecture, and the proposed underlying PaaS model, is being deployed at the University of Ontario Institute of Technology. Future work will focus on tightly integrating the proposed PaaS framework with McGregor's Artemis web service interface to ensure the effective provisioning and management of virtualized hospital resources.

By coupling Artemis with our proposed PaaS model, we can showcase the plausibility and effectiveness of leveraging cloud-based resources to provide real-time knowledge extraction from network-based patient monitoring devices.

References

1. McGregor C (2011) A cloud computing framework for real-time rural and remote service of critical care. Presented at computer-based medical systems (CBMS), 24th international symposium on 2011
2. McGregor C (2005) e-Baby web services to support local and remote neonatal intensive care. In: HIC 2005 and HINZ 2005: proceedings, health informatics society of Australia, p 344
3. McGregor C (2011) System, method and computer program for multi-dimensional temporal data mining. US Patient Office
4. McGregor C, Heath J, Wei M (2005) A web services based framework for the transmission of physiological data for local and remote neonatal intensive care. Data Management, Published by the IEEE Computer Society
5. Savolainen E Cloud service models
6. Finkelstein SM, Speedie SM, Potthoff S (2006) Home telehealth improves clinical outcomes at lower cost for home healthcare. Telemedicine J e-Health 12(2):128–136
7. CareCloud (2013) Carecloud, web-based medical practice management software, vol 2013, p 11
8. Lowe S (2011) Mastering VMware VSphere 5
9. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I (2010) A view of cloud computing. Commun ACM 53(4):50–58
10. Blount M, Ebling M, Eklund JM, James A, McGregor C, Percival N, Smith KP, Sow D (2010) Real-time analysis for intensive care. IEEE Eng Med Biol Mag 29(2):110–118
11. Mell P, Grance T (2011) The NIST definition of cloud computing (draft). NIST Special Publication 800, p 145

MABLIPS: Development of Agent-Based Vehicle Location Tracking System

Yun-Yao Chen, Yueh-Yun Wang and Chun-Wei Lin

Abstract Location positioning and information delivering techniques have recently become important research issues with the increasing of Internet Map service and GPS technology. Similarly, location-aware technologies for bike riders have also become a new issue for bike positioning. In this paper, a mobile agent technology system for locating bicycle position is thus developed. The proposed system can also automatically push and pull the message for location-aware. Combining multi-agent technology with push technique in the proposed system, it can thus enhance the intelligence of bike location tracking and route planning.

Keywords Mobile agent · Location positioning system · Pushing system

Introduction

Agent technology has shown great potential for solving problems in large scale distributed systems. Wireless communication and network services have substantially changed the landscape of communications. The associated facilities of

Y.-Y. Chen
Institute of Manufacturing Information and Systems, National Cheng Kung University, Tainan City, Taiwan, Republic of China
e-mail: yychenzx@gmail.com

Y.-Y. Wang
Department of Leisure and Sports Management, Far East University, Tainan City, Taiwan, Republic of China
e-mail: pamwang@cc.feu.edu.tw

C.-W. Lin (✉)
Innovative Information Industry Research Center (IIIRC), Shenzhen Key Laboratory of Internet Information Collaboration, School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Harbin, China
e-mail: jerrylin@ieee.org

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_132, © Springer Science+Business Media Dordrecht 2014

the leisure bike industry have been developed rapidly in recent years due in part to the advocacy of the government in Taiwan. Therefore, information technologies used in the leisure bike industry have substantially increased. With the huge popularity of mobile devices and remarkable advances in mobile communications, mobile devices have gained much attention from bicycle riders. It allows bicycle riders and users to pinpoint their location and find relevant information on demand. Push technology comprises computers that regularly and actively transmit information via the Internet through certain standards and protocols according to the users' demands [1]. Combining push technology and agent technology with mobile devices, a more convenient interface for its users and improvements to the shortcomings of traditional push technologies can be produced. Many related studies have focused on the efficiency of delivering information to users [2–7]. Many agent standards [8, 9] have been also proclaimed by several organizations [10, 11]. The intelligent agents for message delivery and push/pull research issues are, however, still in urgent need of solutions.

In this paper, a mobile agent-based bicycle positioning system with a mobility agents to free travel amongst the hosts in the network is proposed. The states and codes can be transported to other execution environments in the network, thus reducing the time in a batch process with the resume mechanism [12, 13]. The proposed active push model can also reduce the resource requirements for delivering messages [14].

System Architecture

In the proposed architecture, bike riders are assumed to be equipped with mobile devices. Mobile devices with GPS modules continuously and automatically update the location of the rider to the server through a 3G network. Users who subscribe to the pushing services can receive location information from the bike rider's mobile device. The operation steps are shown in Fig. 1.

For the clients who wish to receive the bike riders' location information must subscribe to the pushing services from the push server by using their devices such as PC, TPC, or laptop. The server provides a firewall system and anti-hack techniques for preserving and securing user information. The Mobile devices carried by bike riders can keep receiving GPS signals continuously. The GPS signals may disappear in some locations but the system will automatically restart the signal searching functions to keep it working continuously. Applications on mobile devices can handle the data and display the latitude, longitude and speed information of the bike by using Universal Description Discovery and Integration UDDI to search for web services on the web server, which are stored with the Web Services Description Language WSDL message in the directory. Received messages can be parsed as readable information for users and displayed on the screen of the mobile device.

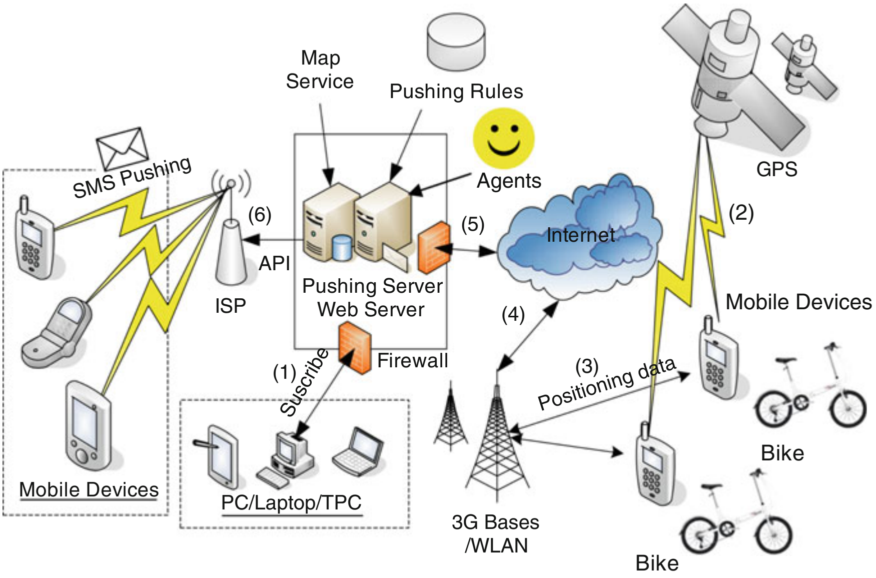


Fig. 1 The proposed architecture

All the information can be transferred by 3G network or Wireless Local Area Network LAN (WLAN). The devices carried by bike riders must able to connect with a 3G network and wireless LAN. When the network disconnects for some unpredictable reason, the system will reconnect to the network and finish any unfinished jobs dispatched by the user.

To access web services on a web server, all applications from a bike rider's mobile device must be identified and obtained an authorization. Some push rules such as certain time interval or certain date/time will trigger the SMS pushing service and mail pushing service. The listening service on the pushing server keeps listening to the contents and is triggered when the rules are satisfied the defined condition.

MABLIPS Framework

The framework of the proposed mobile agent-based bicycle location and information pushing system (MABLIPS) has multiple levels shown in Fig. 2.

The Information Aggregation Layer (IAL) level is composed of various modules to aggregate the proper information needed in the system. The lowest level also contains agencies for acquiring information from different modules. The Middleware Layer (ML) includes BTMC Stationary Agency (BSA), which is responsible for coordinating all of the agencies in IAL. The upper layer is

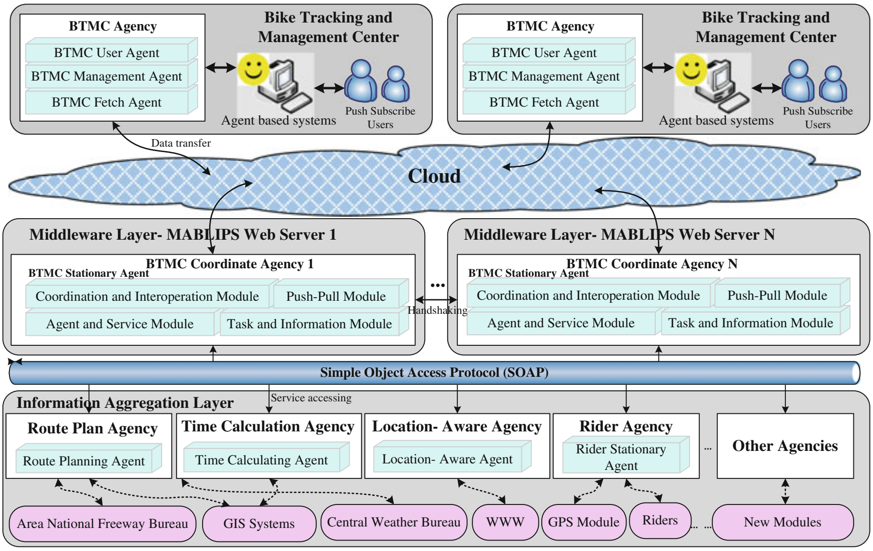


Fig. 2 MABLIPS architecture

described as the Bike Tracking and Management Center (BTMC), which includes the BTMC Agency responsible for delegating tasks to the lower level agents and analyzing the information from these agents. It can also dispatch mobile agents to lower level agencies to fulfill unforeseen tasks, and is a gateway between the Push Subscribe Users (PSU) and agent-based systems proposed in this research. All send/receive information is based on the standard of Simple Object Access Protocol (SOAP) with the Extensible Markup Language (XML) data type.

Architecture Design

Bike Tracking and Management Center

The BTMC consists of a Bike Tracking and Management Center Agency (BTMCA), the Agent-Based Systems (ABS) and the Push Subscribe Users (PSUs). The BTMCA comprises the BTMC User Agent (BUA), the BTMC Management Agent, and the BTMC Fetch Agent (BFA) for performing the tasks as follows:

- Create mobile agents and dispatch them to different agencies.
- Specialize the information from lower level agents and generate reports or control proposals to the BTMC users.
- Generate tasks dynamically and assign these tasks to lower level agents.
- Accept human commands from push subscribe users through agent based systems.

Middleware Layer: MABLIPS Web Server

The ML is based on BTMC Agencies (MTMCA) for performing the tasks as follows:

- Group lower level agents into a cluster according to the assigned task and coordinate these agents to dynamically accomplish the task.
- Serve as agent server and maintain the available services of agents.
- Decompose the assigned tasks by the BTMC to the IAL. Track and monitor those tasks and integrate the information from lower level agents and report to the BTMC agents.
- Interoperate with other agents to solve inter-network problems.

Information Aggregation Layer

The IAL mainly consists of the Route Plan Agency (RPA), the Time Calculation Agency (TCA), the Location-Aware Agency (LAA), and the Rider Agency (RA). The other agent models can be also integrated into the IAL for the future needs. The tasks can be performed as follows:

- Process real-time data from GIS or other data aggregation modules.
- Realize vehicle re-identification, estimate travel time, detect incidents, and forecast weather.
- Improve detection algorithms and select proper algorithms dynamically.
- Track real-time traffic conditions and alert predefined events to BTMC opera.

Agent Communication

In this research, the partial dynamic hierarchical control architecture [13] is adopted to dynamically organize different agents based on task decomposition. The numbers of agents are grouped into virtual clusters. The agent communication and control model are shown in Fig. 3, indicating the agent communication flow and the grouping clusters of agents.

When the RA receives a request from users, it automatically parses the request into certain information data types. The agents can be created to dispatch the information to BCAs. The BA is capable of receive the delay of the heavy network transmitting by the users. While the BTMC Coordinate Agency BCAs accept the data sent by BA, it automatically decomposes tasks assigned by the BA, and then starts to respectively aggregate the information provided by RPA, TCA, RA and RSA. The same data types of information can be recognized and adopted by BCAs.

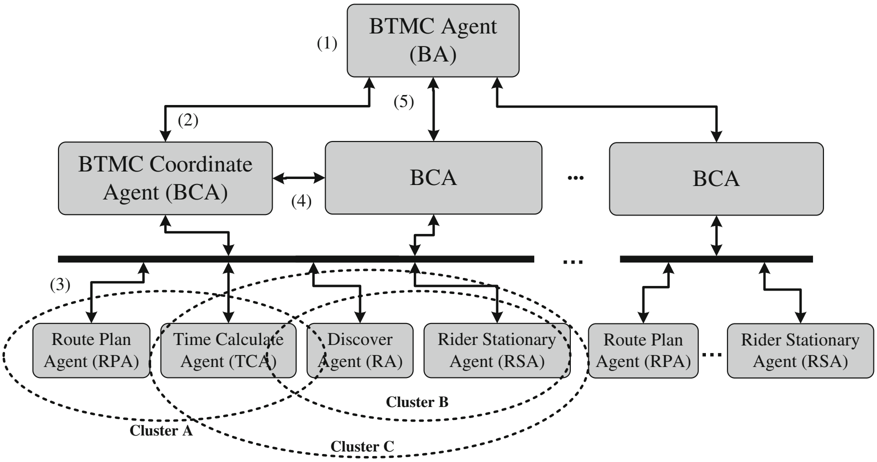


Fig. 3 Agent communication model

Meanwhile, a number of agents can be dynamically grouped into virtual groups and interact with each other to perform the same tasks. These virtual groups can interact through the same interface and same data types. A BCA can coordinate these agents in groups, and interact with other BCAs. The BCAs are able to manage different connections between each other and can handle and avoid the conflicts of reputational data transmitting. After that, the final suggested results return to BA, which displays the best processed results and the best descriptions for users. The proposed model can focus on interactions between different agents. Push subscribe users can receive the best results to locate themselves, obtaining the best routes for riding.

System Development

The system is developed by Apache 2.3 on Microsoft Windows XP Professional operating system with MySQL 5.0.51a database system. The mobile device application was implemented by Microsoft ASP.NET/C# and using .NET Framework on Windows Mobile 6 (HTC Touch Pro 2) mobile device operating system. Based on the defined pushing rule, information can be transmitted to end users in real-time. A few pushing rules are triggered to push SMS message or send the email with the biker's positioning information, which will be displayed on both a biker's mobile device and the installed application by the end user. The predefined rules may include the following conditions.

1. When bike riders ride through certain regions, i.e. the bike rider rides into a certain city (Taipei or ILan).

2. When mobile devices equipped by bike riders have no response for a certain time interval, i.e. the mobile device is unavailable to track for 30 min.
3. When the bike riders manually send out alarms, i.e. the bike rider uses their mobile device to send out emergency information.
4. When the bike rider's route distance exceeds a certain value, i.e. riding for 300 m.

Conclusion

In this paper, the concept of agent technology and an intelligent agent approach for the interaction with standard protocol web services are thus proposed. The framework of MABLIPS is also stated in details. The proposed architecture can be concluded with the following advantages: (1) Mobile agents can enhance the ability of push systems to deal with uncertainty in dynamic environments. (2) MABLIPS integrates multiple modules to enable the comprehensive support of the push system for bike positioning. (3) The open architecture with standard information transmission protocols allows new detection systems to be added by incorporating them into sub-agent systems.

Acknowledgments This research was partially supported both by Shenzhen peacock project, China, under contract No. KQC201109020055A and Shenzhen Strategic Emerging Industries Program under Grants No. ZDSY20120613125016389.

References

1. Li S, Li G, Chen Y (2008) A study on information push in personalized education system. In: International symposium on knowledge acquisition and modeling, pp 534–537
2. Carzaniga A, Wolf AL (2001) Content-based networking: a new communication infrastructure. In: NSF workshop on an infrastructure for mobile and wireless systems, pp 1–5
3. Cao J, Feng X, Lu J, Chan H, Das SK (2002) Reliable message delivery for mobile agents: push or pull. In: International conference on parallel and distributed systems, pp 314–320
4. Chen C, Ding J, Hua G, Chen Y (2009) Design and implementation of SMS employment agent based on ontology. In: World congress on software engineering, pp 489–492
5. Franklin M, Zdonik S (1998) Data in your face: push technology in perspective. In: ACM SIGMOD international conference on management of data, pp 516–519
6. Jacobsen HA (2001) Middleware services for selective and location-based information dissemination in mobile wireless networks. In: The advanced topic workshop on middleware for mobile computing, pp 154–169
7. Kozina B, Glavinic V, Kraljevic L (2006) Agent-based messaging system for m-learning. In: IEEE mediterranean electrotechnical conference, pp 1213–1216
8. Brewington B, Gray R, Moizumi K, Kotz D, Cybenko G, Rus D (1993) Mobile agents for distributed information retrieval. In: Intelligent information agents, chapter 15, pp 355–395
9. Zhang Y, Liu J (2003) Mobile agent technology. Tsinghua University Press, Beijing, pp 9–11
10. Mobile agent system interoperability facilities (MASIF) http://www.omg.org

11. The foundation for Intelligent Physical Agents http://www.fipa.org
12. Lange D, Oshima M (1998) Programming and developing java mobile agents with aglets. Addison Wesley Longman Inc., Botson
13. Lange DB (1998) Mobile objects and mobile agents: the future of distributed computing. In: European conference on object-oriented programming, pp 1–12
14. Sun T, Wang Y (2005) Implementation of a chat system with push and mobile agent technologies. In: International conference on intelligent agents, web technologies and internet commerce, international conference on computational intelligence for modeling, control and automation, vol. 2, pp 336–360
15. Du J, Tian Y, Zuo M, Zhou Y (2007) The realization of push and pull model in the agent-based electronic commerce platform. In: International conference on control, automation and systems, pp 17–20

Data Sensing and Communication Technology for the IoT-IMS Platform

Tin-Yu Wu and Wen-Kai Liou

Abstract When the Future Internet is defined as a dynamic global network infrastructure, the Internet of Things (IoT), an integrated part of it, is viewed as one of the most important technologies. Distinct from former definition and applications of IoT and the IoT based on *Radio Frequency IDentification* (RFID), the current IoT means the interconnection of devices and appliances via the Internet. However, because the development of the IoT technologies varies, there is neither an unified standard nor a standardized interoperability to provide an unified network software architecture for the IoT. In this paper, we propose a method for integrating the data detected by sensors in the IoT-IMS communication platform with specifications, including Zigbee, *Wireless Sensor Network* (WSN), RFID and IEEE 1451, and converting the data to *Sensor Be One Code* (SBO Code), a coding format that supports all kinds of sensors. Through this system, the data gathered by different sensors can be unified to meet the IoT requirement.

Keywords IoT · WSN · Intelligent sensor · IEEE 1451

Introduction

With the popularization of broadband network and the enhancement of transmission rate, users' requests for novel and diverse multimedia services and applications are never satisfied and the notion of ubiquitous services therefore

T.-Y. Wu (✉) · W.-K. Liou
Department of Computer Science and Information Engineering, National Ilan University, Ilan City, Taiwan, Republic of China
e-mail: tyw@niu.edu.tw

W.-K. Liou
e-mail: kevin1990115@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_133, © Springer Science+Business Media Dordrecht 2014

becomes very important. To meet user demand for network services, the *Internet Service Providers* (ISPs) keep making efforts in presenting a novel integrated infrastructure in response to future broadband services. Moreover, this demand has initiated the idea of the *Next Generation Network* (NGN) and the *IP Multimedia Subsystem* (IMS) under the NGN framework [1]. In addition to the solutions for the next generation core networks, the concept of the *Internet of Things* (IoT) takes shape: RFID tags, sensors or QR code embedded in the objects are linked via the Internet to endow the objects with the intelligence or ability to communicate with human beings and other objects. Therefore, the IoT enables not only machine/device/employee management/control, but also remote monitoring of home appliances and vehicles.

To cope with such a trend, this paper focuses on the IoT-IMS based data sensing and communication technology. Based on the IoT-IMS communication platform, the Smart Client Platform gathers the data detected by different sensors (QRCode, RFID, Barcode and RuBee, for example), converts the data to *Unique IDentifiers* (UID) by EPC coding scheme, and writes the UID on smart tags. By this method, objects and devices can be easily interconnected via the IoT. Furthermore, *EPC Information Services* (EPCIS) of EPC Global Network are built on the Smart Client Platform for information and data transmission while the All-IP heterogeneous network server is established to integrate different transmission standards and incompatible data to EPC Sensor Profile.

Based on EPCglobal [2], this paper presents *Sensor Be One code* (SBO Bode), a novel coding scheme to integrate data formats of heterogeneous sensors and allow data transmission in either Zigbee, RFID, Bluetooth or Wi-Fi. Two main research directions of this paper include:

- Establish IMS-URI to integrate IMS with the IoT.
- Present SBO Code to integrate data formats and transmission interfaces of IEEE 1451 standards and other standards.

Related Work

Introduction to the EPCglobal Network

For wireless radio system integration, the EPCglobal Network [3–11] was first developed by the Auto-ID Center, an unique partnership between almost 100 global companies and labs of seven leading research universities: the Massachusetts Institute of Technology in the US, the University of Cambridge in the UK, the University of St. Gallen in Switzerland, the University of Adelaide in Australia, Fudan University in China, Keio University in Japan, and Information and Communication University in South Korea; the labs were later renamed the Auto-ID Labs. Auto-ID solutions presented by the Auto-ID labs include *Electronic*

Table 1 EPC data structure

Header	EPC manager	Object class	Serial number

Product Code (EPC), proposal for network elements and passive tags and readers. In 2003, the research results produced by the Auto-ID Labs promoted the establishment of EPCglobal Inc., an international nonprofit joint venture between GS1 and GS1 US. As the commercial successor of the Auto-ID Labs, EPCglobal Inc. aims at architecting open and autonomous EPCglobal standards and promoting wider adoption of EPC technology.

Electronic Product Code

The EPC number is either a 64 bits or 96 bits identifier, a string of data in four partitions: header, EPC manager, object class and serial number (Table 1). The 96 bits version of the EPC, the EPC-96, can identify approximately 268 million manufacturers, each of which can have 16 million object classes and 68 billion serial numbers for every unique item. Table 2 displays the coding scheme of EPC-96 Type I based on GID.

- Header (assigned by EPCglobal): Identifies the length, type, structure, version, and generation of the EPC. GS1 assigns the keys to each class of object.
- EPC Manager (assigned by EPCglobal): Identifies the company that manages the data, usually the manufacturer of the product the EPC is attached to.
- Object Class (managed by the company): Identifies the class or type of product.
- Serial Number (managed by the company): Identifies every unique item of the object class.

Conversion Between EPC and Bar Code

GS1 has been recently making great efforts in integrating the EPC code with QR code and thus there is no clear explanation for the conversion between the EPC and QR code currently. Therefore, only the conversion between the EPC and bar code is mentioned here. To convert bar code to the EPC correctly, we need the following information: header, filter value, partition value and EPC manager number.

Table 2 EPC-96 I type coding scheme

Header	EPC manager	Object class	Serial number
8 bits	28 bits	24 bits	36 bits

The *Uniform Resource Identifier* (URI) coding scheme is usually adopted by RFID to enable data interchange in large-scale systems. To inform the operators who read and write about the tag format, the URI is a string of characters: urn : epc : tag : Enc Name: Encoding Specific Fields, where Enc Name refers to the coding format of the tag.

Internet of Things

The *Internet of Things* (IoT) means that through RFID, infrared sensors, *Global Positioning System* (GPS), laser scanners, etc., all independent devices can be connected via wireless links using specific protocols. Therefore, the devices are endowed with the ability to communicate with each other and intelligent identification, positioning, tracking, monitoring, controlling and management thus can be realized. Features of the IoT as below.

- In the IoT, the information about the objects is retrieved from RFID tags, sensors, QR code, or other means.
- The information about the objected is transmitted via the Internet because the IoT is connected to the Internet.
- By using various intelligent computing techniques, the IoT can process large amounts of data and accomplish intelligent control of objects.

Architecture of the Internet of Things

IoT is mainly divided into three layers: perception layer, network layer and application layer (Fig. 1). The main function of the perception layer is to identify objects and collect information. The key techniques in this layer include RFID and WSN (Wireless Sensor Network) [12–15], and the sensing devices include sensors, readers, IP Cams and *Micro Electro Mechanical Systems* (MEMS). The main function of the network layer is information transmission and processing, including integration of heterogeneous networks, center for intelligent information processing, information center, and management center. The primary techniques in the network layer include 2G/3G, Wi-Fi, WiMax and Zigbee. The main function of the application layer is to control input/output and provide services in cloud service platforms. The application support layer provides interfaces between the application layer and the network layer, integrates different networks and supports applications.

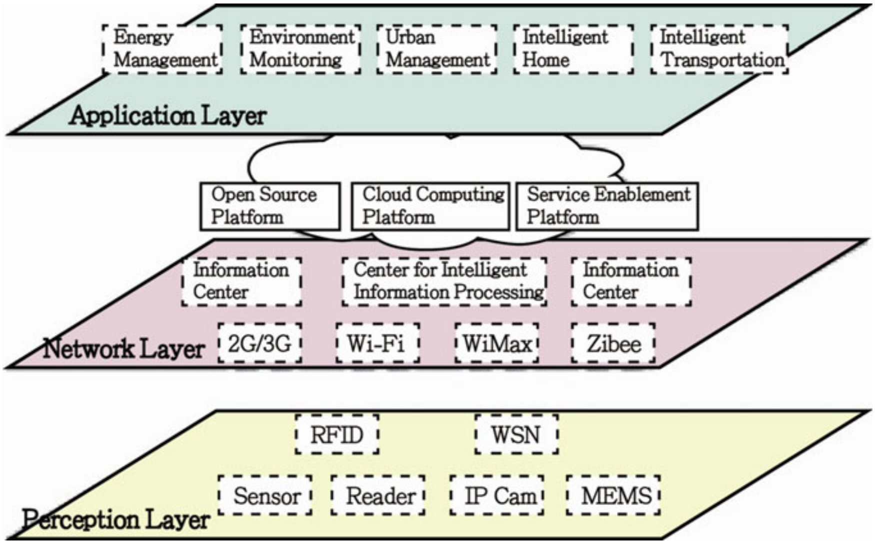


Fig. 1 Architecture of the internet of things

IP Multimedia Subsystem

As part of 3GPP Release 5, the *IP Multimedia Subsystem* (IMS) [1] is defined as a collection of core network functions in *Next Generation Networks* (NGN) that support multimedia services delivery on mobile devices. Based on the 3GPP Release 4 packet-switched domain, Release 5 adds an IP-based multimedia subsystem that uses the *Session Initiation Protocol* (SIP) to control calls or manage data via packet-switching functions. Therefore, the IMS provides support for mobile applications. Though independent of circuit-switched systems, the IMS can connect to circuit-switched networks and satisfy requirements for security, billing, roaming and QoS. In addition to the standardization of IMS by 3GPP, the IMS architecture is also adopted by 3GPP2 for IP multimedia services. Moreover, the IMS has been adopted, adapted and optimized by ITU-T.

Responsible for processing SIP signaling and responses, the IMS core comprises three kinds of *Call Session Control Functions* (CSCF): *Proxy-CSCF* (P-CSCF), *Serving-CSCF* (S-CSCF) and *Interrogating-CSCF* (I-CSCF). As the most important component in the IMS architecture, the CSCF is responsible for multimedia session processing: multimedia session control, address translation, service negotiation, and so on.

IEEE 1451 Smart Transducer Interface Standard

To unify sensor interface standards, *IEEE Technical Committee* 9 (TC-9) in 1993 discussed the possibility of developing a smart sensor communication interface standard and founded the IEEE 1451 smart transducer interface standards. A series of workshops have been established and defined a set of standards for intelligent sensors.

A Transducer Electronic Data Sheet, TEDS, refers to a set of electronic data in a standardized format stored in an EEPROM attached to the STIM. The TEDS includes the manufacturer ID, model number, serial number, measurement range, electrical output range, sensitivity, power requirements, and calibration data. The TEDS allows the self-identification, self-description, automatic setting of parameters (range, sensitivity, proportion, for example), self-diagnosis and self-calibration of the sensors (Table 3).

Proposed Architecture

The proposed architecture of this paper is based on the IoT-IMS communication platform, in which the data from different sensors is gathered by a *Network Capable Application Processor* (NCAP) and converted to SBO Code based on the EPC coding format. Objects therefore can be interconnected via the IoT. Here, we focus on how to transmit these messages and information to the cloud network.

Table 3 The IEEE 1451 family of standards

IEEE 1451.0 (2007) (in balloting)	Defines the common functions, communication protocols, and transducer electronic data sheet (TEDS) formats for IEEE 1451.X standards
IEEE 1451.1 (1999) (published standard)	Defines a common object model for smart transducers and interface specifications
IEEE 1451.2 (1997) (published standard)	Defines a TEDS and a point-to-point digital interface and communication protocols between the smart transducer interface module (STIM) and the network capable application processor (NCAP)
IEEE 1451.3 (2003) (published standard)	Defines a transducer interface standard for distributed multi drop systems: transducer bus interface module (TBIM) and transducer bus controller (TBC)
IEEE 1451.4 (2004) (published standard)	Defines a mixed-mode (analog/digital) transducer interface (MMI)
IEEE 1451.5 (2007) (in balloting)	Defines a transducer-to-NCAP interface and TEDS for wireless transducers
IEEE 1451.6 (in progress)	Defines a transducer-to-NCAP interface and TEDS using the CANopen protocol
IEEE 1451.7 (2010) (in progress)	Defines an interface and communication protocol between transducers and RFID systems

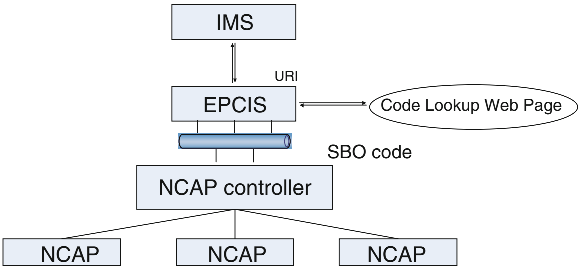


Fig. 2 IMS-IoT network architecture

To integrate different standards of wireless communication and compatible data, we refer to *EPC Information Services* (EPCIS) in the EPCglobal Network. For heterogeneous sensor integration, our proposed architecture enables All-IP heterogeneous networks to acquire the names and addresses of smart tags via the IoT for sharing data in different standards through the platform. Finally, by using previous techniques together with the IoT-IMS communication platform, we can create associations between objects: the Family Tree-ID Relationship will be combined with IMS-URI to form associations between objects in the execution process of the research project. The project execution will achieve the integration of the identifiers under the IoT-IMS framework.

With the aim of heterogeneous sensor integration, this paper first builds a Smart Client Platform based on EPC SGTIN coding scheme. In the platform, we propose *Sensor Be One code* (SBO code) to integrate data formats of heterogeneous sensors.

SBO code is divided into several partitions to mark the transmission method, transmission content, sensor manufacturer, sensor serial number and so on. By using SBO coding scheme, we can integrate sensor data of various WSN as the basis of WSN for the IoT.

In this architecture, the NCAP gathers the data from different sensors and uploads it to the cloud database for remote users to access data. Figure 2 displays our proposed IMS-IoT network architecture. Based on IEEE 1451 architecture, the NCAP at the bottom are responsible for receiving data returned by sensors within a fixed range and uploading the data to the NCAP Controller, the large server in the range that is responsible for data management and relaying. The NCAP Controller converts the data uploaded by the NCAP to SBO code and stores it in EPCIS for users to access. At last, we integrate the IMS with EPCIS so that the IMS server can connect to EPCIS through URI.

Based on EPC SGTIN-96 coding scheme, the 256-bit code is designed as shown in Fig. 4. The code includes a Header for 14 bits, Manufacturer ID for 29 bits and Sensor Class for 11 bits, 40 bits in subtotal. The rest 202 bits are Sensor Serial Number for 24 bits and Transmission Content for 178 bits. The code thus is converted into the format: urn : sbo : id : transmission method : manufacturer code : sensor class : sensor serial number : transmission content. As shown in Fig. 3, we can convert ID to IMS-URI to complete the network deployment for the IoT.

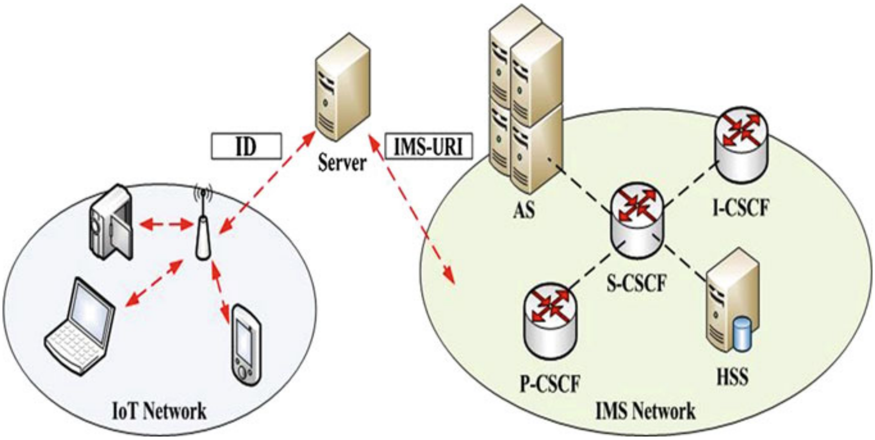
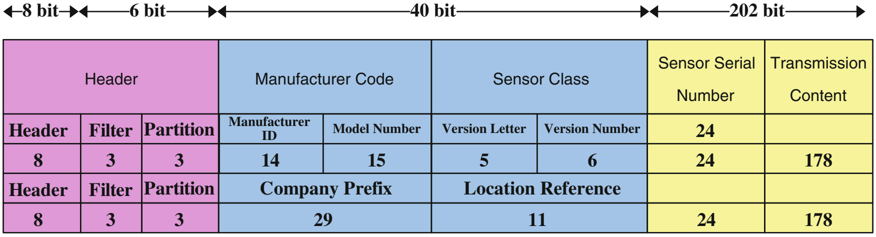


Fig. 3 Converting ID to IMS-URI


8 bit
6 bit
40 bit
202 bit

Header			Manufacturer Code		Sensor Class		Sensor Serial Number	Transmission Content
Header	Filter	Partition	Manufacturer ID	Model Number	Version Letter	Version Number	24	
8	3	3	14	15	5	6	24	178
Header	Filter	Partition	Company Prefix		Location Reference			
8	3	3	29		11		24	178

Fig. 4 SBO coding scheme

Achievements

The equipment adopted in this research project includes cc2530 Zigbee module, different types of sensors, NI PCI signal acquisition card to acquire Zigbee transmission, and NI LabVIEW 2011 software to convert data from Zigbee sensors to SBO code and relay data to remote SQL database for storage. As shown in Fig. 5, to construct and examine our proposed architecture and to test the differences between intelligent and traditional sensors, we use IEEE 1451-based intelligent sensors as the sensor nodes in the IoT. The major difference between intelligent and traditional sensors is that an intelligent sensor has a EEPROM to store the TEDS for plug and play data measurement.

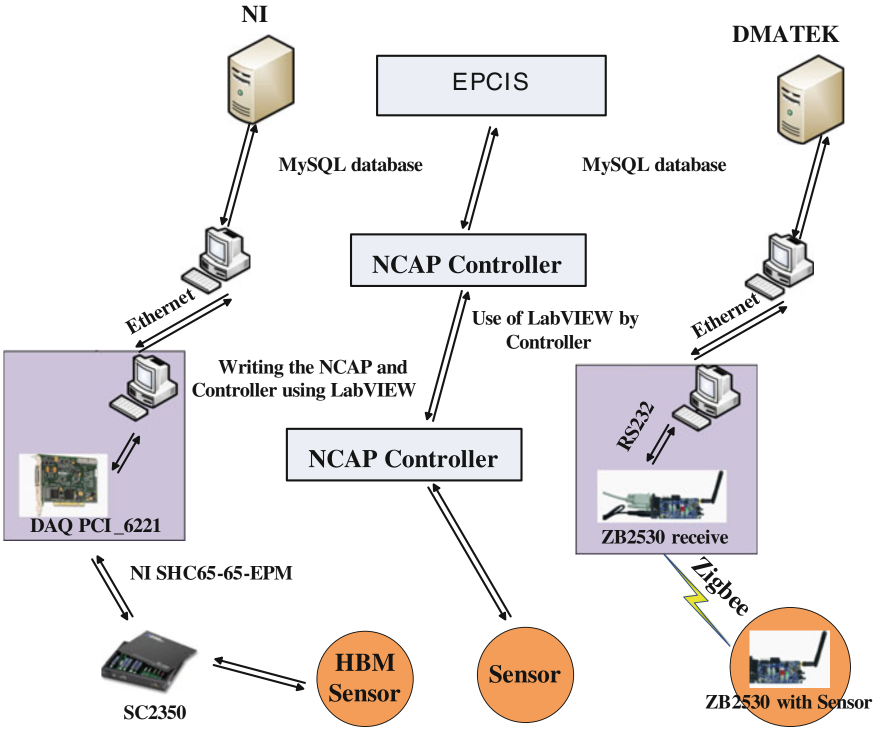


Fig. 5 Hierarchical equipment structure

Conclusion

This paper points out a possible direction for the IoT: collecting the data from Zigbee, Bluetooth, Wi-Fi, integrating the data into the same SBO Code and implementing an operable IoT system. Our future objective is to assign an independent IPv6 to each object. Due to hardware limitations, computing capability and power of sensors, the biggest challenge and also the ultimate goal for the future IoT is that every object has an IP address, which is waiting to be achieved by international standards and related equipment providers.

Acknowledgments This study was supported by the National Science Council, Taiwan, under grant No. NSC 101-2219-E-197-003.

References

1. Hwang J, Kim N, Kang S, Koh J (2008) A framework for IMS interworking networks with quality of service guarantee. In: The 7th international conference on networking 2008, ICN 2008, pp 454–459, 13–18 Apr 2008
2. Epcglogal Tag Data Standard v. 1.5 http://www.epcglobalinc.org/standards/tds/tds_1_5-standard-20100818.pdf
3. Zhang DY, Zhu YL, Chen HN (2008) An algorithm for deployment of RFID readers in EPC network. In: The 4th international conference on wireless communications, networking and mobile computing 2008, WiCOM '08, pp 1–4, 12–14 Oct 2008
4. Lee SD, Shin MK, Kim HJ (2007) EPC vs. IPv6 mapping mechanism. In: The 9th international conference on advanced communication technology, pp 1243–1245, 12–14 Feb 2007
5. Gao J, Liu F, Ning H, Wang B (2007) RFID coding, name and information service for internet of things. In: IET conference on wireless mobile and sensor networks 2007, (CCWMSN07), pp 36–39, 12–14 Dec 2007
6. Chang YC, Chen JL, Lin YS, Wang SM (2008) RFIPv6: a novel IPv6-EPC bridge mechanism. In: international conference on consumer electronics 2008, ICCE 2008. Digest of technical papers, pp 1–2, 9–13 Jan 2008
7. Thiesse F, Floerkemeier C, Harrison M, Michahelles F, Roduner C (2009) Technology, standards, and real-world deployments of the EPC Network. Internet computing, IEEE, pp 36–43, Mar–Apr 2009
8. Sanchez Lopez T, Kim D (2007) A context middleware based on sensor and RFID information. In: The 5th annual IEEE international conference on pervasive computing and communications workshops 2007, PerCom Workshops '07, pp 331–336, 19–23 Mar 2007
9. Huifang Deng, Haiyan Kang (2010) Research on high performance RFID code resolving technology. In: The 3rd international symposium on intelligent information technology and security informatics (IITSI) 2010, pp 677–681, 2–4 Apr 2010
10. Chalasani S, Boppana R (2007) Data architectures for RFID transactions. Industrial informatics, IEEE transactions, pp 246–257, Aug 2007
11. Pathak R, Joshi S (2009) Recent trends in RFID and a java based software framework for its integration in mobile phones. In: The first Asian himalayas international conference internet 2009, AH-ICI 2009, pp 1–5, 3–5 Nov 2009
12. Choi YS, Jeon YJ, Park SH (2010) A study on sensor nodes attestation protocol in a Wireless Sensor Network. In: The 12th international conference on advanced communication technology (ICACT) 2010, vol 1, pp 574–579
13. Yueqing R, Lixin X (2010) A study on topological characteristics of wireless sensor network based on complex network. In: The international conference on computer application and system modeling (ICCASM) 2010, vol 15, pp 486–489
14. Ilyas MU, Radha H (2006) End-to-end channel capacity of a wireless sensor network under reachback. In: The 40th annual conference on information sciences and systems 2006, pp 1713–1718
15. Choi SH, Kim BK, Park J, Kang CH, Eom DS (2004) An implementation of wireless sensor network. IEEE Trans Consum Electron 50(1):236–244

Adaptive Multi-Hopping MAC Mechanism for WSN Scheduling

Lin-Huang Chang, Shuo-Yao Chien, H. F. Chang and Tsung-Han Lee

Abstract In this paper, we extend the routing enhanced MAC (RMAC) protocol to combine wireless sensor network (WSN) technology with energy-efficiency and quality of service (QoS) guarantee transmission scheduling mechanisms. This energy-efficient QoS-aware scheduling over multi-hopping WSNs is called adaptive multi-hopping MAC (AMH-MAC). Based on the RMAC scheme, we further adjust the transmitted power and the number of multi-hopping nodes dynamically to reduce the transmission latency and energy consumption. We conduct the simulation using network simulator 2 (ns-2) to show the improvement of the performance in terms of end to end latency and energy consumption as compared with RMAC scheme.

Keywords RMAC · Adaptive transmitted power · Scheduling · Multi-hop

Introduction

With the characteristics of low cost and low power consumption, wireless sensor network (WSN) communication is suitable for real-time and emergency services, such as mountain emergency rescue, mining or farm monitoring.

L.-H. Chang · S.-Y. Chien · T.-H. Lee (✉)
Department of Computer Science, National Taichung University, Taichung, Taiwan, Republic of China
e-mail: thlee@mail.ntcu.edu.tw

L.-H. Chang
e-mail: lchang@mail.ntcu.edu.tw

H. F. Chang
Department of Computer and Communication Engineering, Taipei Chengshih University of Science and Technology, Taipei, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_134, © Springer Science+Business Media Dordrecht 2014

The idle listening for the well-known duty-cycling schemes in WSN MAC protocol is inefficient and wastes significant energy. To reduce this energy consumption of idle listening, some duty-cycle MAC protocols, such as S-MAC [1], T-MAC [2], and RMAC [3], have been introduced. The S-MAC and T-MAC mitigate the energy consumption of idle listening by using fairness contention to transmit packets in a periodic synchronized listen/sleep schedule. The increase of contention probability, due to the synchronized wake-up time at the sensor node's signal coverage during the same short period, and end to end delivery latency, due to the single hop characteristic in one duty cycle, would deteriorate the performance of S-MAC and T-MAC mechanisms. Therefore, the RMAC protocol, with contention-based and synchronization-based mechanisms, exploits cross-layer routing information to allow multi-hopping transmission within a single duty-cycle. Low power listening (LPL) is another well-known mechanism which employs a preamble in front of each outgoing packet from physical layer for transmission detection. These mechanisms, such as B-MAC [4] and X-MAC [5], could reduce the energy consumption, however, the latency issue due to the asynchronous mechanisms would be a problem.

In this paper, we take into account the end to end latency issue by employing the RMAC protocol with synchronization and multi-hopping scheduling mechanisms. Furthermore, with the LPL mechanism for reducing energy consumption during the RMAC idle time in mind, we dynamically control the transmitted power to reduce transmission energy consumption and dynamically adjust the number of multi-hopping in one duty cycle to properly increase the hopping numbers and consequently reduce the end to end latency. The proposed adaptive multi-hopping MAC (AMH-MAC) mechanism is expected to provide duty cycle scheduling in WSN MAC protocol with energy efficiency and low end to end latency.

Related Work

The S-MAC of well-known WSN MAC protocol, similar to IEEE 802.11 DCF mode, is designed with periodic synchronized listen/sleep schedule. The listening period, with sensor nodes' radio being set enabled, consists of SYNC period to synchronize the sensor nodes' clocks and DATA period to deliver packets with RTS/CTS mechanism for the reservation of transmission scheduling during SLEEP period. During the SLEEP period, every sensor node goes to sleep to save energy unless it is scheduled to send or receive data.

The duty cycle S-MAC protocol provides scheduling mechanism with energy efficiency as compared with traditional MAC protocols. However, the end to end delivery latency could be increased because a packet is forwarded over only a single hop for each duty cycle. The RMAC is therefore designed to forward a data packet with multiple hops in one duty cycle to reduce the end to end delivery latency.

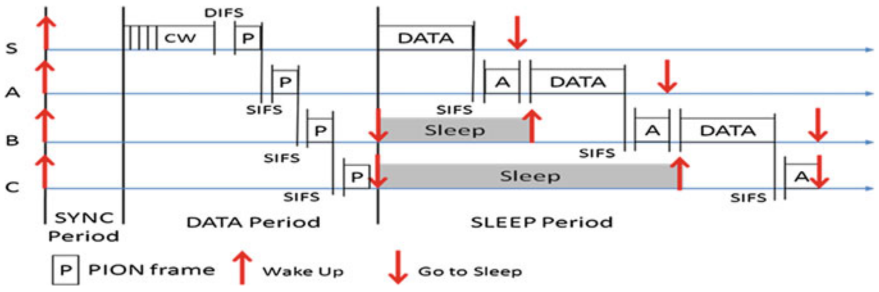


Fig. 1 The overview of RMAC operational and scheduling mechanism

The overview of RMAC operational and scheduling mechanism is shown in Fig. 1. RMAC uses a small control frame, called pioneer frames (PIONs), during the DATA period across multiple hops to allow all nodes along the path learn when to be awake. A PION control packet is employed to reserve and confirm communication, such as RTS and CTS respectively, for the corresponding data transmission schedules in SLEEP period within a single duty cycle. When a relaying node, such as node A or B which receives the data packet from the upstream nodes S or A and forwards it to the downstream nodes B or C respectively, receives PION frame, it sets its next wakeup time according to the hop count information in the PION so that it can wake up in its turn to receive or transmit data packet during SLEEP period.

The Hybrid MAC (HMAC) mechanism is proposed by Wang et al. [6] by combining the TDMA scheme to the multi-hopping scheduling of RMAC scheme to achieve energy saving and low delivery latency. However, HMAC mechanism uses fixed time slot for packet transmission which could limit the scheduling to certain number of nodes. Liu et al. [7] proposed the Q-MAC mechanism for traffic classification with the cut-off of contention window to achieve QoS service demand. But the packet transmission of Q-MAC did not support multi-hopping issue.

Some mechanisms and issues related to the QoS services of WSN MAC protocols are addressed above. Based on the RMAC protocol, in this paper, we propose an AMH-MAC mechanism to dynamically control the transmitted power to reduce transmission energy consumption and dynamically adjust the number of multi-hopping in one duty cycle to further reduce the end to end latency.

Design of Adaptive Multi-Hopping MAC Protocol

In RMAC mechanism, each node has to wake up in DATA period for listening even if it will not be scheduled to receive or forward data during this round of duty cycle. This results in the energy waste during this wake up for idle listening.

On the other hand, the number of hops, N, in a single duty cycle for RMAC multi-hopping scheme could be a factor affecting the transmission cycles for a packet travelling from a source node to the destination node. The research from Cho and Bahk [8] proposed a hop extended MAC (HE-MAC) scheme to optimal wake up duration with respect to packet latency along with power consumption. However, they did not provide any dynamic adjustment of N value for different transmission paths.

The operational mechanism and power transmission of AMH-MAC is shown in Fig. 2. In AMH-MAC mechanism, the SYNC packet, carrying N value and path ID, informs all nodes in the path during SYNC period. All nodes in the path will adjust the wakeup/sleep time according to the updated N value during the DATA period instead of waking up and idling for all DATA period. The dynamical adjustment of N value carried in SYNC packet on the one hand will relay PION frame to more nodes, and on the other hand, reserve the power consumption of nodes, which will not participate in the transmitting or receiving packets, by keeping them sleep during this duty cycle.

The path ID is used for nodes to determinate whether they will participate in this DATA period. For those with different path ID will sleep in DATA period till next SYNC period to save more energy.

Besides adjusting the number of multi-hopping nodes dynamically, the proposed AMH-MAC mechanism also dynamically adjusts the transmitting power to the receiving nodes during the DATA period. The research on the location-based routing protocol from He et al. [9] and Shang et al. [10] revealed that sensor nodes learn the position and distance of each other with RSSI signal. In AMH-MAC, the SYNC packet is broadcasted with maximum radio power to as many covered nodes as possible to reduce the number of broadcast while maintaining the synchronization functionality. However, each node uses point to point communication to its receiving neighbors with optimized transmitted power during the DATA period, as shown for the transmission of the unicast PION/data packets in Fig. 2. This dynamic adjustment of transmitted power will reduce the power consumption for nodes participating in DATA period.

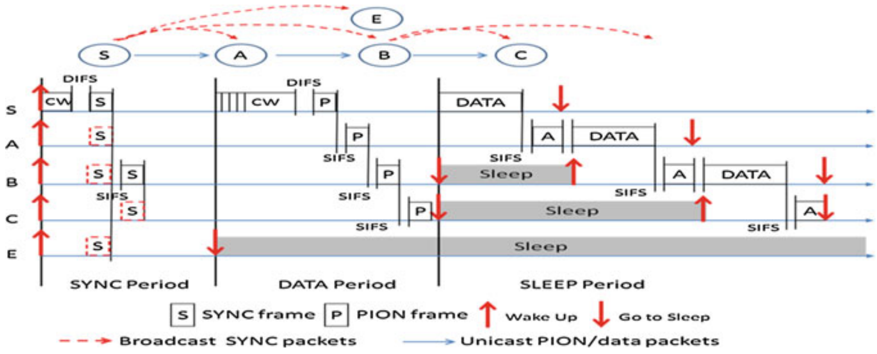


Fig. 2 The operational mechanism and power transmission of AMH-MAC

Performance Evaluation

We conduct the ns2 simulation with free space model to analyze the performance of AMH-MAC as compared with S-MAC and RMAC protocols. Due to the limit of the page length in this paper, we can only provide partial simulated results, such as the case with N equals to 4 in Table 1, conducted in the research.

In AMH-MAC, we add a long sleep time parameter, which equals to DATA period plus SLEEP period, in the SYNC packet to indicate long sleep without wakeup until the next duty cycle when a node does not participate any receiving or transmitting packets. The chain topology shown in Fig. 3 with nodes 1–24, separated from 100 m each, is used to simulate the experiment with constant bit rate (CBR) traffic loads of 100 packets at rate of 0.02 packet/sec.

The simulation results of energy consumption distribution for S-MAC, RMAC and AMH-MAC mechanisms are illustrated in Fig. 4a–c, respectively, where the total energy consumption is divided into receive (rx), transmit (tx), sleep and idle energy. As shown in Fig. 4a, b for S-MAC and RMAC schemes respectively, the idle listening, which contributes significant ratio of energy consumption, consumes much more energy as compared with our proposed AMH-MAC mechanism in Fig. 4c. The S-MAC scheme, with only one hop per single duty cycle, suffers from the idle listening power consumption significantly. On the other hand, the dynamic adjustment of transmitted power and modification of SYNC packet, such as path ID which keeps unnecessary wakeup for nodes not participating in transmission

Table 1 Simulation parameters

Parameters	S-MAC	RMAC	AMH-MAC
Bandwidth		20 Kbps	
Rx power		0.5 W	
Idle power		0.45 W	
Sleep power		0.05 W	
Sync range	100 m	100 m	200 m
Data/PION range	100 m	100 m	100 m
Transmitted power (200 m)	Null	Null	24.5 dBm
Transmitted power (100 m)		21.49 dBm	
SYNC period (ms)		55.2	
DATA period (ms)	104.0	168.0	168.0
SLEEP period (ms)	3025.8	4241.8	4241.8
Long SLEEP period (ms)	Null	Null	4409.8
Total cycle (ms)	3158.0	4465.0	4465.0
Duty cycle		5 %	
N value	Null	4	4

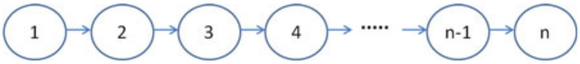


Fig. 3 The chain topology with multiple hops

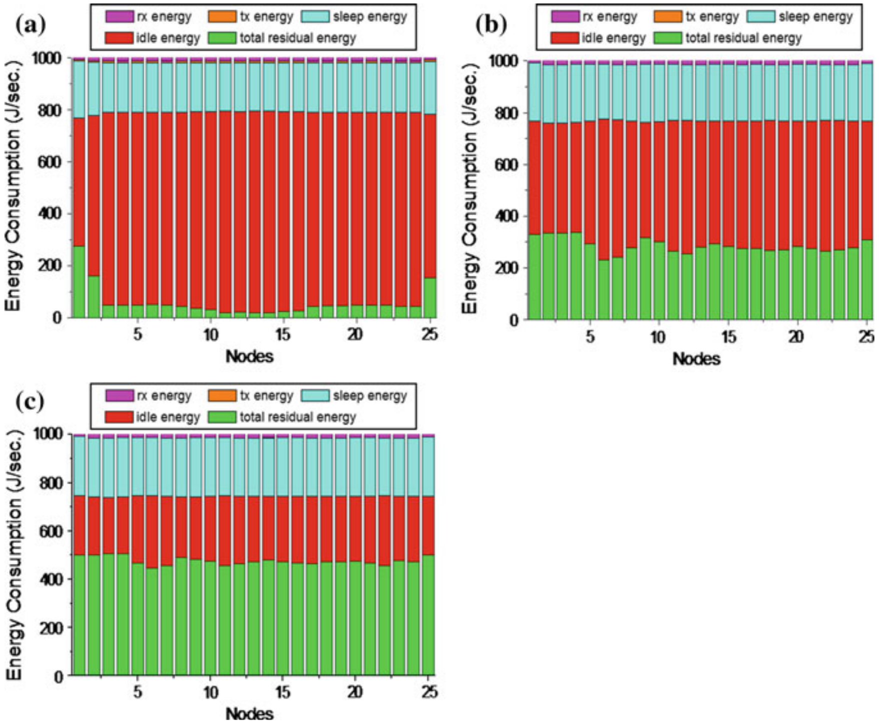
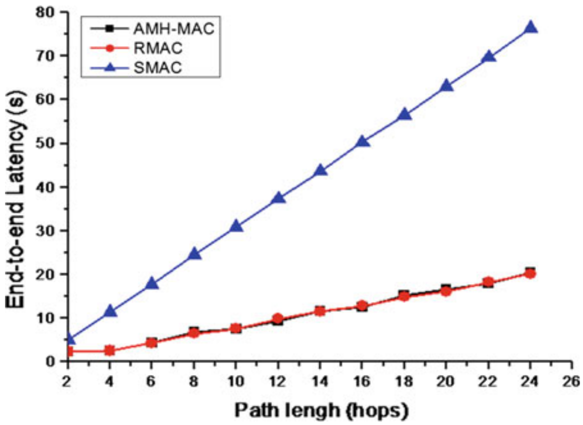


Fig. 4 **a** S-MAC, **b** RMAC, **c** AMH-MAC energy consumption distribution

during the duty cycle, in AMH-MAC further reduce the idle listening energy consumption as compared with RMAC scheme. The results reveal that significant improvement in total residual energy in our proposed AMH-MAC mechanism as compared with S-MAC and RMAC schemes.

The simulation result for end to end deliver latency is shown in Fig. 5 for three different schemes. As noted, the multi-hopping characteristic plays a significant role in latency. Therefore, due to the one hop only ability per single duty cycle for S-MAC, the latency is much higher than other schemes. The latency of AMH-MAC is very close to the RMAC result because we only present the result with the same N value but not the dynamic N cases in AMH-MAC due to the limit of page length. Although the proposed AMH-MAC mechanism with fixed N value shows about the same latency as compared with RMAC, however, it significantly improves the energy consumption.

Fig. 5 End to end latency for different schemes



Conclusion

In this paper, we have proposed an adaptive multi-hopping MAC (AMH-MAC) for multi-hopping transmission scheduling mechanism, extended from RMAC protocol, with dynamical adjustment of transmitted power and number of multi-hopping nodes. The simulation results show that the proposed AMH-MAC mechanism provides significant improvement in terms of energy consumption and end to end deliver latency as compared with S-MAC and RMAC, though the results for the dynamic adjustment of N values were not shown in performance evaluation session due to the limitation of page length.

Acknowledgments The authors would like to acknowledge the support from the National Science Council of Taiwan (No. NSC 101-2221-E-142-003, NSC100-2221-E-142-008) and National Taichung University regarding the MoE project (No. 1020035480A).

References

1. Ye W, Heidemann J, Estrin D (2002) An energy-efficient MAC protocol for wireless sensor networks. In: Proceedings of INFOCOM, June 2002, pp 1567–1576
2. van Dam T, Langendoen K (2003) An adaptive energy-efficient MAC protocol for wireless sensor networks. In: Proceedings of first international conference on embedded networked sensor systems (SenSys 2003), Nov 2003, pp 171–180
3. Du S, Saha AK, Johnson DB (2007) RMAC: a routing-enhanced duty cycle MAC protocol for wireless sensor networks. Proceedings of INFOCOM, May 2007, pp 1478–1486
4. Polastre J, Hill J, Culler D (2004) Versatile low power media access for wireless sensor networks. In: Proceedings of the second ACM conference on embedded networked sensor systems (SenSys), Nov 2004, pp 95–107
5. Buettner M, Yee GV, Anderson E, Han R (2006) X-MAC: a short preamble MAC protocol for duty-cycled wireless sensor networks. In: Proceedings of the 4th international conference on embedded networked sensor systems (SenSys) 2006, pp 307–320

6. Wang H, Zhang X, Abdesselam FN, Khokhar A (2010) Cross-layer optimized MAC to support multihop QoS routing for wireless sensor networks. IEEE Trans Veh Technol 59(5):2556–2563
7. Liu Y, Elhanany I, Qi H (2005) An energy-efficient QoS-aware media access control protocol for wireless sensor networks. In: Proceeding of mobile adhoc and sensor systems conference 2005
8. Cho KT, Bahk S (2012) Optimal hop extended MAC protocol for wireless sensor networks. Comput Netw 56(4):1458–1469
9. He T, Huang C, Blum BM, Stankovic JA, Abdelzaher T (2003) Range-free localization schemes for large scale sensor networks. In: Proceeding of the 9th annual international conference on mobile computing and networking (MobiCom'03) 2003, pp 81–95
10. Shang Y, Ruml W, Zhang Y, Fromherz M (2004) Localization from connectivity in sensor networks. IEEE Trans Parallel Distrib Syst 15(11):961–974
11. McCanne S, Floyd S ns Network Simulator. http://www.isi.edu/nsnam/ns/

Avoiding Collisions Between IEEE 802.11 and IEEE 802.15.4 Using Coexistence Inter-Frame Space

Tsung-Han Lee, Ming-Chun Hsieh, Lin-Huang Chang, Hung-Shiou Chiang, Chih-Hao Wen and Kian Meng Yap

Abstract Recently, more and more wireless networks have been deployed and start provide varies services to customers. Such as smart-phones integrate heterogeneous wireless devices, it may cause unintended interactions between multiple radios using difference radio access technologies. Thus, heterogeneous wireless network will be the trend of future network. The Co-channel interference problem in heterogeneous wireless networks is more and more important. This paper focus on the coexistence problem between IEEE 802.11b/g/n and IEEE 802.15.4 protocols in the ISM 2.4 GHz band. The performance impact on IEEE 802.15.4 network under IEEE 802.11 wireless network has been analysis, and results show that IEEE 802.15.4 has serious collision problem if there has no appropriate scheduling mechanisms among two wireless protocols. A CIFS (Coexistence Inter-Frame Space) has been proposed in this paper to effectively enhance the IEEE 802.15.4 transmission probability which is implemented in IEEE 802.11 nodes to

T.-H. Lee · M.-C. Hsieh · L.-H. Chang (✉) · H.-S. Chiang · C.-H. Wen
Department of Computer Science, National Taichung University of Education,
Taichung, Taiwan, Republic of China
e-mail: lchang@mail.ntcu.edu.tw

T.-H. Lee
e-mail: thlee@mail.ntcu.edu.tw

M.-C. Hsieh
e-mail: BCS100101@gm.ntcu.edu.tw

H.-S. Chiang
e-mail: bcs100110@gm.ntcu.edu.tw

C.-H. Wen
e-mail: BCS100102@gm.ntcu.edu.tw1

K. M. Yap
Department of Computer Science and Networked Systems, Sunway University,
Bandar Sunway, Malaysia
e-mail: kmyap@sunway.edu.my

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_135, © Springer Science+Business Media Dordrecht 2014

observe the transmission opportunity for IEEE 802.15.4 under wireless coexistence environment.

Keywords IEEE 802.11 · IEEE 802.15.4 · Coexistence · Co-channel interference

Introduction

Increasingly devices are equipped with multiple radio interfaces. This is due to the decreasing cost and size of radio embedded technology and the increasing demand from end-users. Thus, heterogeneous wireless networks are currently receiving a significant amount of interest in research. This paper focuses on the co-channel interference between IEEE 802.11 b/g/n [1–3] and IEEE 802.15.4 [4] protocols in the ISM 2.4 GHz band. In Fig. 1, the IEEE 802.11b/g has eleven channels between 2.412 and 2.462 GHz [5]. Each channel occupies 22 MHz and up to 3 separate channels can be simultaneously used without any mutual interference. In the other hand, the IEEE 802.15.4 has sixteen channels between 2.4 and 2.4835 GHz. Each channel occupies 3 MHz. The IEEE 802.15.4 will suffer serious interference problems from IEEE 802.11 if there is no proper arrangement for the mutual scheduling mechanism. Thus, the goal of this research is try to improve the performance of IEEE 802.15.4 in the heterogeneous wireless network environment.

Furthermore, it can explore to three different interference situations depend on the IEEE 802.15.4 transmission range. Thus, the interference circumstances can be divided into three ranges which are shown in Fig. 2.

- Range 1: a range in which IEEE 802.15.4 nodes and IEEE 802.11 nodes can sense each other. Such as the range between AP and Zigbee C.

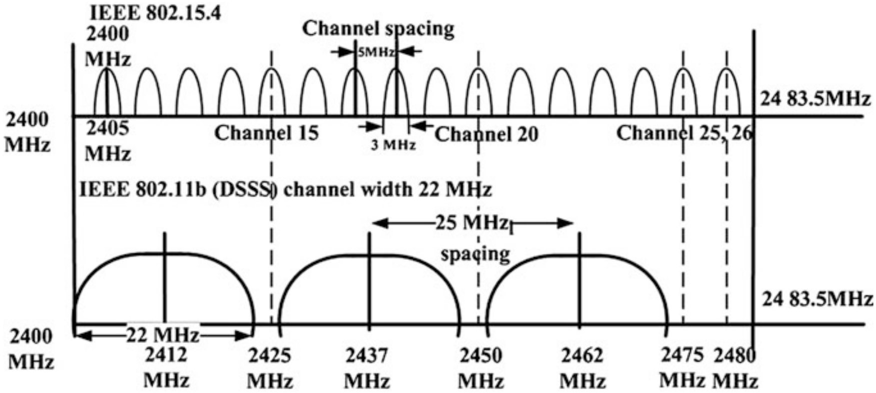


Fig. 1 The channel allocation in the 2.4 GHz band

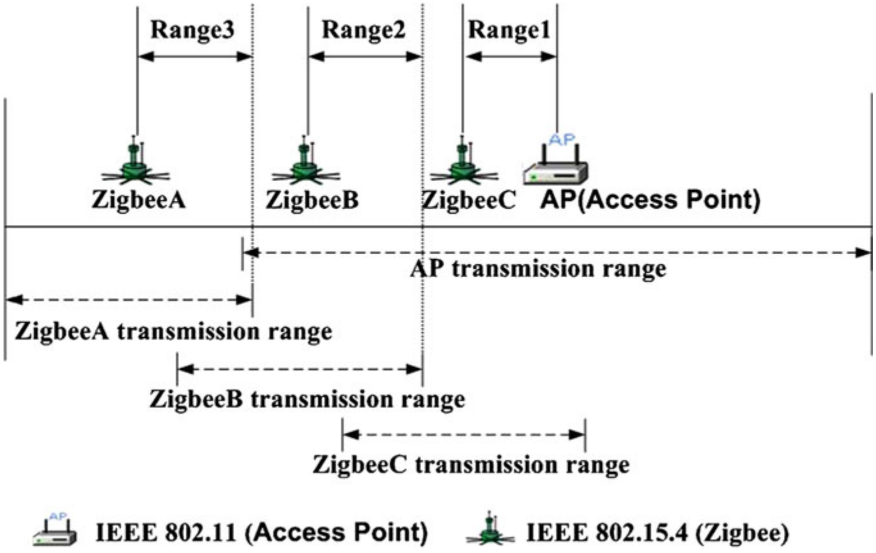

Range3
Range2
Range1
AP
ZigbeeA
ZigbeeB
ZigbeeC
AP(Access Point)
AP transmission range
ZigbeeA transmission range
ZigbeeB transmission range
ZigbeeC transmission range
IEEE 802.11 (Access Point)
IEEE 802.15.4 (Zigbee)

Fig. 2 The coexistence cases of IEEE 802.11 and IEEE 802.15.4

- Range 2: a range in which IEEE 802.15.4 nodes can sense IEEE 802.11 nodes, but IEEE 802.11 can't sense IEEE 802.15.4. Such as the range between AP and Zigbee B.
- Range 3: a range in which cannot sense each other. Such as the range between AP and Zigbee A.

The rest of this paper is organized as follows. Related Work will discuss related works for co-channel interference between IEEE 802.11. and IEEE 802.15.4. Packet Error Analysis of IEEE 802.15.4 Under IEEE 802.11 Coexistence describes the co-channel interference model in wireless IEEE 802.15.4 and IEEE 802.11 coexistence networks. The Coexistence Inter-Frame Space describes the proposed Coexistence Inter-Frame Space mechanism. Performance Evaluation evaluates the performance of the proposed CIFS by PRISM simulation. Finally, summarizes conclusions of this paper and give suggestions for future work in Conclusion.

Related Work

In [6–8], authors proposed different interference models. Simulation results can reduce the bit error rate (BER) and the packet error rate (PER) effectively. In [9, 10], interference avoidance is achieved by means of energy detection (ED); Yi et al. [9], both using "Safe Distance" and "Safe Offset Frequency" to reduce the BER and PER. In [10], author builds a neighbor table to allow nodes switch to relatively low-interference channel to reduce the packet loss rate. In [11] authors employs a separate signaler node between two IEEE 802.15.4 nodes to emit a carrier signal (busy-tone), thereby enhancing the IEEE 802.15.4's visibility to IEEE 802.11.

The reason of interference has been studied in [12, 13]. Authors try to avoid the interference by using packet scheduling mechanisms in the MAC layer. In [12], the authors use the channel switching and routing synchronization to avoid interference of IEEE 802.11. Yuan [13], the authors take into account difference interference from Range 1 to Range 3. However, all above literatures use packet scheduling mechanism to avoid interference in the MAC layer. In [14], MCSP (Multi-Canal Scheduling Protocol) has been proposed to avoid the lighthouse collision. Shin [15] was modified the state transition probabilities in IEEE 802.15.4 Markov process model. Simulation results show the proposed model match to the theoretical expressions. The longer distance between IEEE 802.11 and IEEE 802.15.4 gains higher throughput. However, IEEE 802.11 nodes are not able to affect the throughput of IEEE 802.15.4 when IEEE 802.15.4 nodes outside the IEEE 802.11's transmission range, such as Range 3 in Fig. 2.

In [12–14] was focus on the MAC layer arrangement for the mutual scheduling mechanism. In [6–10] was focus on the Range 1 which IEEE 802.15.4 nodes and IEEE 802.11 nodes can sense each other, but there were inconspicuous display problem. Most of literatures above focus on the Range 2 co-channel interference, and most of literatures ignore the impact on Range 1. To alleviate the interference problem, the IEEE 802.15.4 employs the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) mechanism. However, IEEE 802.15.4 still has serious collision problem if there has no appropriate scheduling mechanisms for Range 1 co-channel interference. Therefore, we consider the performance enhancement of IEEE 802.15.4 for Range 1 co-channel interference in this paper.

Packet Error Analysis of IEEE 802.15.4 Under IEEE 802.11 Coexistence

In this section, we investigate the interference between IEEE 802.11 and IEEE 802.15.4 in the Range 1. A simulation environment was developed using the PRISM [16] simulator as shown in Fig. 3; Table 1 shows the related parameters for the simulation. In Table 1, the length of Slot-Time is different between IEEE 802.11 and IEEE 802.15.4. IEEE 802.15.4's Slot-Time is sixteen times to IEEE 802.11. IEEE 802.11 has shorter DIFS (DCF, Distributed Coordination Function IFS). It cause IEEE 802.11 always has higher transmission opportunity than IEEE 802.15.4. The unfairness transmission opportunity will resulting packet dropped in IEEE 802.15.4 due to reach the maximum number of back-off restriction.

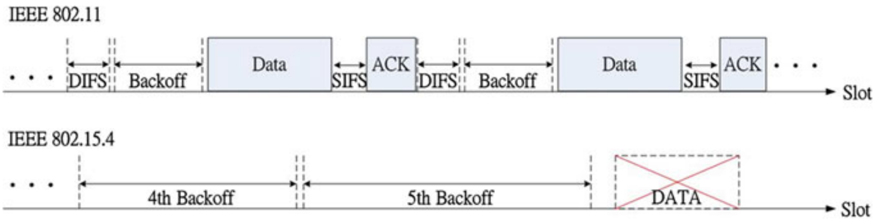


Fig. 3 Original IEEE 802.15.4 and IEEE 802.11 packet transmission in range 1

Table 1 Simulation configuration and parameters

	IEEE 802.15.4	IEEE 802.11b
Packet length (byte)	133	1,500
Data rate (bps)	250 k	2 M
Retry limit	3	7
CWmin	7(2^3–1)	31(2^5–1)
CWmax	31(2^5–1)	1,023(2^{10}–1)
Backoff-max	5	6
A slot time	320 µs	20 µs
DIFS		50 µs
SIFS	192 µs	10 µs

The Coexistence Inter-Frame Space

This section, a Coexistence Inter-Frame Space (CIFS) has been proposed to enhance the IEEE 802.15.4 transmission probability under the IEEE 802.11 network. In Fig. 4, the CIFS is implemented in IEEE 802.11 nodes, which try to observe the transmission opportunity to IEEE 802.15.4 under wireless coexistence environment. Figure 5 shows the IEEE 802.15.4 transmission probability in different CIFS. The result shows the shorter CIFS cause lower transmission probability in IEEE 802.15.4. The IEEE 802.15.4 transmission probability reaches to 100 % when the CIFS around 1.2 ms. Moreover, result shows the length of CIFS is proportional to the transmission probability under wireless coexistence environment.

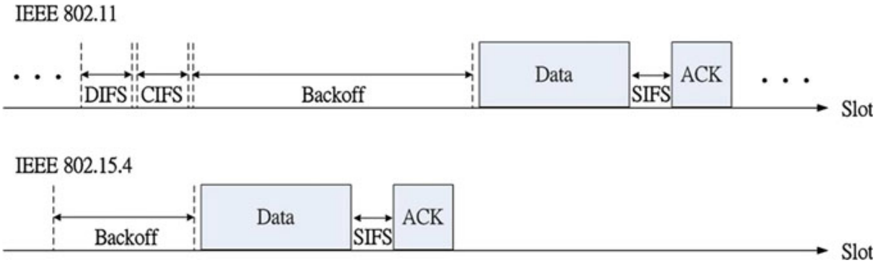


Fig. 4 Implement the CIFS to postpone the IEEE 802.11 packet transmission in range 1

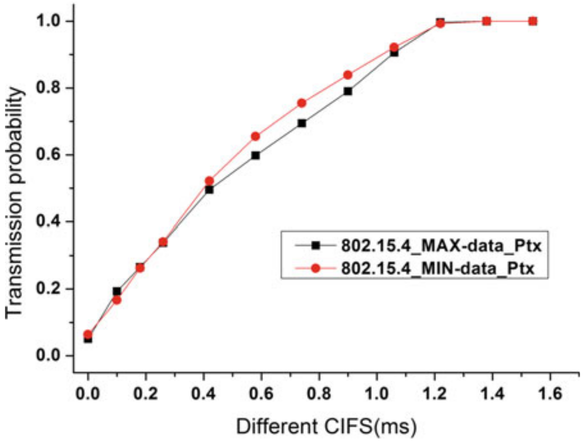


Fig. 5 The transmission probability of IEEE 802.15.4 in different CIFS

Performance Evaluation

In order to demonstrate the relationship between the payload size and transmission probability in IEEE 802.15.4 and IEEE 802.11 coexistence networks. Our probabilistic model was emulated by PRISM [16] simulator. The simulation environment is summarized in Table 1, which considers the operation of IEEE 802.15.4 with one pair of source and destination nodes in the multiple IEEE 802.11b WLANs coexistence environment.

We obtain transmission throughput from Eqs. (1) and (2) for IEEE 802.11 and IEEE 802.15.4 respectively.

$$S_{11} = \frac{L_{11}}{slot_{11} * IEEE\ 802.11\ \mathrm{A\ Slot\,Time}} \tag{1}$$

$$S_{15.4} = \frac{L_{15.4}}{slot_{15.4} * IEEE\ 802.15.4\ \mathrm{A\ Slot\,Time}} \tag{2}$$

The S_{11} presents the IEEE 802.11 throughput in coexistence environment, and L_{11} is the packet length of IEEE 802.11. $slot_{11}$ is the total slots to transmit a data packet in IEEE 802.11. $S_{15.4}$ is IEEE 802.15.4 throughput in coexistence environment. $L_{15.4}$ is IEEE 802.15.4 packet length, and $slot_{15.4}$ is total slots to transmit an IEEE 802.15.4 packet.

$$DRU_{11} = \frac{S_{11}}{IEEE\ 802.11\ Data\ Rate} \tag{3}$$

$$DRU_{15.4} = \frac{S_{15.4}}{IEEE\ 802.15.4\ Data\ Rate} \tag{4}$$

Finally, from Eqs. (3) and (4) will obtain the utilization on desire data rate for IEEE 802.11 and IEEE 802.15.4 respectively.

Fig. 6 IEEE 802.11 and IEEE 802.15.4 throughput in different CIFS

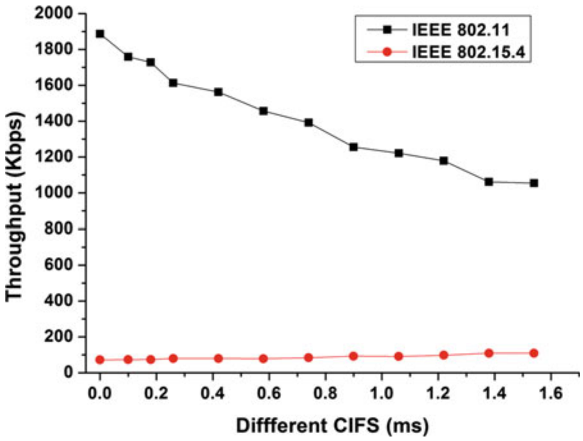


Fig. 7 The utilization on desire data rate in IEEE 802.11 and IEEE 802.15.4 data in different CIFS

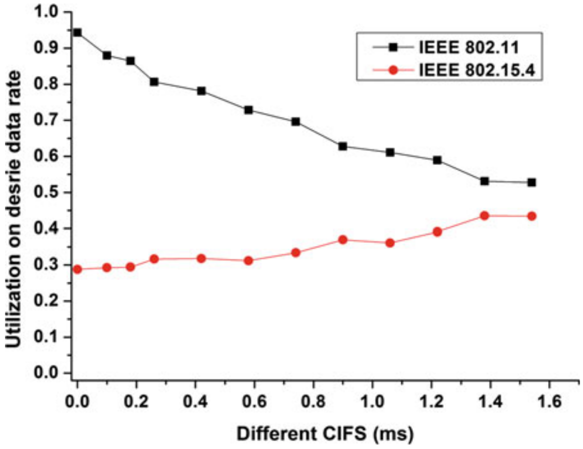


Figure 6 shows IEEE 802.15.4 with CIFS has better probability against to original IEEE 802.15.4. It is because that the CIFS will postpone the data transmission in IEEE 802.11 to observe the transmission opportunity for IEEE 802.15.4. Thus, the throughput of IEEE 802.15.4 is proportional to the length of CIFS. However, the throughput of IEEE 802.11 is inversely proportional to the length of CIFS.

In Fig. 7, it shows a trade-off between the utilization on desire data rate in IEEE 802.11 and IEEE 802.15.4. However, coexistence among IEEE 802.11 and IEEE 802.15.4 in the 2.4 GHz band is an important issue in order to ensure that each wireless service maintains its desired performance requirements.

Conclusion

In this paper, the performance impacts on the IEEE 802.15.4 network under the IEEE 802.11 wireless network have been investigated. Simulation results show the IEEE 802.15.4 has serious interference problems in the Range 1 if there are no appropriate scheduling mechanisms among two protocols. The result shows, the proper CIFS (around 1.2 ms) in IEEE 802.11 nodes will increase IEEE 802.15.4 transmission probability. We have improved the performance of IEEE 802.15.4 networks from co-channel interference in the ISM band. In addition, the short IEEE 802.15.4 packet size may take relief from the co-channel interference from IEEE 802.11. Thus, a dynamic packet adjustment via a Channel ranking scheme is consider as the future work.

Acknowledgments The authors would like to acknowledge the support from the National Science Council of Taiwan (No. 101-2119-M-142-001, 100-2221-E-142 -002) and National Taichung University regarding the MoE project (No. 1020035480A).

References

1. IEEE 802.11 Work Group (1999) Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specification: high-speed physical layer extension in the 2.4 GHz Band, ANSI/IEEE Std 802.11b
2. IEEE 802.11 Work Group (1999) Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications: further higher data rate extension in the 2.4 GHz band, ANSI/IEEE Std 802.11g
3. IEEE 802.11 Work Group (1999) Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications: further higher data rate extension in the 2.4 GHz band, ANSI/IEEE Std 802.11n
4. IEEE 802.15 Work Group (2003) Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs), ANSI/IEEE Std 802.15.4
5. Musaloiu R, Terzis A (2007) Minimising the effect of wifi interference in 802.15.4 wireless sensor networks. Int J Sen Netw 3(1):43–54
6. Tamilselvan GM, Shanmugam A (2010) Inter and intra cluster scheduling for performance analysis of coexistence heterogeneous networks. Int J Comput Appl 1(8):0975–8887
7. Tamilselvan GM, Shanmugam A (2011) A cluster based interference mitigation scheme for performance enhancement in IEEE 802.15.4. J Sci Ind Res 70(9):756–761
8. Tamilselvan GM, Shanmugam A (2011) Multi hopping effect of ZigBee nodes coexisting with WLAN nodes in heterogeneous network environment. 978-1-4577-0183-2/09/$26.00 2011 IEEE
9. Yi P, Iwayemi A, Zhou C (2011) Developing ZigBee deployment guideline under WiFi interference for smart grid applications. IEEE Trans Smart Grid 2(1):110–120
10. Won C, Youn JH, Sharif HA, Deogun J (2005) Adaptive radio channel allocation for supporting coexistence of 802.15.4 and 802.11b. In VTC, 2005
11. Zhang X, Shin KG (2012) Cooperative carrier signaling: harmonizing coexisting WPAN and WLAN devices. IEEE Trans Netw, no. 99
12. Shin SY, Park HS, Choi S, Kwon WH (2007) Packet error rate analysis of ZigBee under WLAN and bluetooth interferences. IEEE Trans Wireless Commun 6(8):2825–2830

13. Yuan W (2011) Coexistence of IEEE 802.11b/g WLANs and IEEE 802.15.4 WSNs: modeling and protocol enhancements. Technische Universiteit Delft, PhD Thesis
14. Sahraoui M (2012) Collisions avoidance multi-channel scheme for the protocol IEEE 802.15.4. International conference on information technology and e-Services (ICITeS), pp 1–9
15. Shin SY (2013) Throughput analysis of IEEE 802.15.4 network under IEEE 802.11 network interference. AEU—Int J Electron Commun
16. Prsim http://www.prismmodelchecker.org/

Implementation and Evaluation of Multi-Hopping Voice Transmission over ZigBee Networks

Lin-Huang Chang, Chen-Hsun Chang, H. F. Chang and Tsung-Han Lee

Abstract The wireless sensor network (WSN) technology has been one of the most active research topics in the last several years. Besides sensing the environmental or physical data, there has been a growing need on supporting voice communication, particularly under emergency conditions, in WSNs. In this paper, we will conduct the implementation of voice codec on embedded system using multi-hopping ZigBee communication. The implementation issues, in terms of codec and transmission rate-dependence consideration, as well as the voice quality with multi-hopping are analyzed and evaluated for the designed test-bed.

Keywords WSN · ZigBee · Speex · Xbee · Bit rate

Introduction

The wireless sensor network (WSN) technology has been one of the most active research topics currently. The ZigBee [1] standard, based on IEEE 802.15.4 physical radio standard [2], is designed for low rate wireless personal area network (LR-WPAN) applications with the deployment of low-cost and low-power consumption wireless sensors or devices. The traditional operation to applications of

L.-H. Chang · C.-H. Chang · T.-H. Lee (✉)
Department of Computer Science, National Taichung University, Taichung, Taiwan,
Republic of China
e-mail: thlee@mail.ntcu.edu.tw

L.-H. Chang
e-mail: lchang@mail.ntcu.edu.tw

H. F. Chang
Department of Computer and Communication Engineering,
Taipei Chengshih University of Science and Technology, Taipei,
Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_136, © Springer Science+Business Media Dordrecht 2014

WSN is restricted on infrequently sensing environmental or physical data, such as temperature, pressure and light information with low duty cycle. Besides sensing and reporting data asynchronously, there has been a growing need on providing real time applications in WSNs. The transmission of voice, particularly under emergency conditions, has been proven to be feasible in WSNs [3]. However, there are still some issues, such as codec tuning and rate-dependent voice quality, need to be resolved for voice transmission using ZigBee communication.

The basic characteristic of ZigBee includes (1) low transmission rate of 250 Kbps, (2) short transmission distance of about 50–100 m (with different transmission power, it could be extended to 300 m), (3) low power consumption, and (4) operation at 2.4 GHz ISM band. It supports point-to-point communication with multi-hopping transmission.

In this paper we will conduct the implementation of Speex [4] voice codec on embedded system with Xbee [5] wireless sensor nodes. We will present the system architecture, voice codec consideration of Speex implementation, and transmission rate-dependence issues for the designed voice transmission test-bed over multi-hopping WSNs. The voice quality is always one important issue to achieve the realization of voice transmission using ZigBee communication. We will perform the voice quality analysis with multi-hopping WSNs in this research.

The rest of the paper is organized as follows. Section Related Work reviews the related works in voice transmission using WSN. The designed test-bed of voice transmission system using multi-hopping ZigBee communication is described in details in section System Design, followed by the performance analysis in section Performance Evaluation. Finally, section Conclusions addresses the conclusion of this research.

Related Work

Although the voice communication over ZigBee networks has attracted much attention currently, there are still some challenges to deliver timeliness guarantees or provide quality of service (QoS) with multi-hopping transmission. The research in [3] has presented the evaluation and implementation of a voice codec using Z-phone project to transmit voice in full-duplex mode over ZigBee networks. They investigated some criteria in selecting the most suitable codec and provided the performance analysis of the Z-phone. However, the system is based on single hop ZigBee scenario. The researches in [6, 7] have implemented the voice over sensor network with multi-hopping. They also provided the performance analyses, such as delay, packet loss and quality of service (QoS) on their test-beds. Unfortunately, the analyzed mean opinion score (MOS) values, which is a key factor to evaluate QoS of voice communication, in these researches are mostly below 3.0, corresponding to some users dissatisfied level.

In [8], the research group used the PCM codec, TLV320AIC1107 voice processing chip with ADPCM coding rates 16 kHz to 64 kHz, for the coding/decoding procedures to implement push-to-talk (PTT) functionality of voice communication across multi-hopping ZigBee networks based on CC2430 architecture. The packet loss of the designed test-bed with more than 3 hops was higher than 10 % or even as high as 40 % or more which may not provide acceptable voice quality.

The research in [9] provided some results on the time analysis in terms of a variety of delays for voice transmission over multi-hopping ZigBee networks. Their measurements revealed that the device probing delay is about 30 ms, including 15 ms for device channel link delay and 15 ms for trigger delay. The research in [10] on the other hand suggested the synchronization of relay nodes to the sending and receiving nodes to reduce the contention delay in the WSN chain topology. This design however, required additional synchronization signal overhead periodically.

In this paper we will conduct the implementation of open source Speex voice codec to transmit voice packets over multi-hopping ZigBee networks using Xbee wireless sensor nodes. The system performance, such as signal strength, delay, and rate-dependent voice quality, is carried out for the implemented test-bed.

System Design

Hardware Components and Design Issues

The system architecture with process flow of the designed test-bed is shown in Fig. 1, where the multi-hopping scenario is represented by the relay node besides sending and receiving nodes. The sending and receiving nodes are designed to be the ZigBee voice gateways which are implemented on the BeagleBone embedded systems with Xbee version Serial 1 wireless sensor nodes. At the sender side, when the analog signal is converted to digital data, it is encoded by Speex encoder, where the coding rates are adjusted to study the rate-dependence on voice quality. The Xbee version Serial 1 ZigBee node, with maximum transmission rate 250 Kbps, can carry 100 bytes data in payload. To avoid the issues and efforts from reassembling audio frames due to packet loss, each Xbee packet should carry complete audio frames instead of audio byte streams in the payload. The Xbee hardware structure and internal data flow diagram is illustrated in Fig. 2 which will be explained in details shortly.

Due to relatively small transmitting rate and limited buffer in Xbee hardware structure, it is unable to handle and process all audio streams from Speex encoder in corresponding time. This will result in significant packet loss and delay during the internal process. Therefore, we use leaky bucket mechanism to implement a flow control in ZigBee voice gateway to modulate the rate difference between Speex encoder and Xbee device. When the voice data is stored in traffic buffer of voice gateway, the flow control module with leaky bucket mechanism will be triggered to read 100 B voice data every 3 ms from traffic buffer to Xbee buffer. The triggered rate is tuned to adapt to the Xbee transmission rate and limited buffer

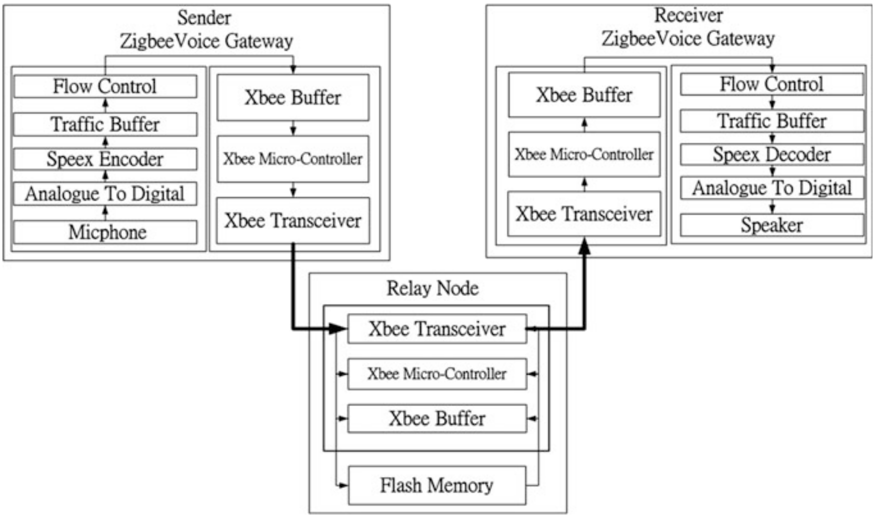


Fig. 1 The implemented system architecture with process flow

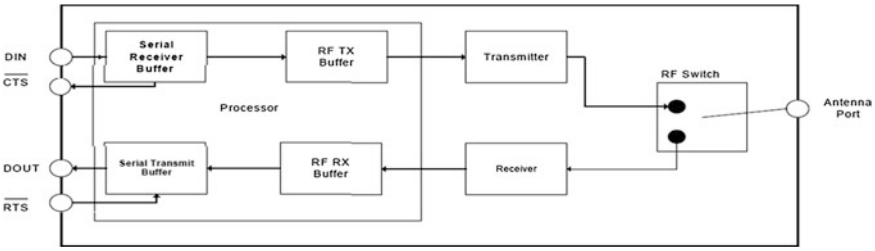


Fig. 2 The Xbee internal data flow diagram

space. On the other hand, the relay node, performing only the forwarding of voice packets, is simply a Xbee sensor node with flash memory. Because of the strict requirement in end-to-end delay for real time voice communication, the relay node, which does not conduct any additional operation or coding process but extend the transmission distance, aims to relay voice packets efficiently.

As shown in Fig. 2, there are two registers, RF TX buffer and RF RX buffer, in Xbee structure. When the RF TX buffer is full, the CTS pin will signal to stop receiving data from DI buffer. The resuming process of CTS signal will cause additional latency, therefore, it is important not to overwhelm the RF TX buffer. That is the significance to design a flow control in sender voice gateway and actually similar situation for the receiving case. The RF switch block provides the transceiver conversion within a very short time between the transmitting and

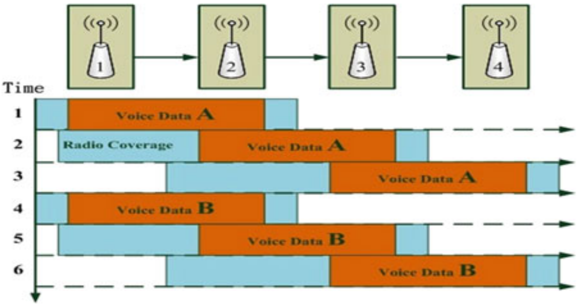


Fig. 3 The multi-hopping topology and transmission flow

receiving antenna ports. The RF switch in relay node is implemented to attach the destination address of the relayed packets into the next hop address.

Figure 3 illustrates the multi-hopping topology as well as the voice data transmission flow in our implementation. Each sensor node, arranged in multi-hopping chain topology, is separated from 50 to 110 m. The measurements of the test-bed are conducted by sending voice data from node 1 relayed via nodes 2 and/or 3 and finally received by node 4.

Audio Codec Using Speex

The relatively low transmission rate, 250 Kbps ideally, of ZigBee wireless networks, and small payload, 100 bytes, using Xbee wireless sensor nodes is a challenge for voice data transmission. One of the top issues to face this challenge is to select a reliable and efficient voice encoder for codec implementation. The research in [3] proposed an implementation of open source Speex voice compression technology over ZigBee network environment. Speex, based on the code-excited linear prediction (CELP) speech coding algorithm, provides a good execution efficiency and voice quality. The fixed-point arithmetic instead of using floating point unit (FPU) for Speex compression will reduce the power consumption and computing time in WSN nodes with low processing ability and low power digital signal processor (DSP). The design goal of Speex is to optimize the voice data with high quality speech and low bit rate.

Speex employs multiple bit rates to support three different modes of operation, including ultra-wideband (32 kHz sampling rate), wideband (16 kHz sampling rate) and narrowband (telephone quality with 8 kHz sampling rate), and some features such as voice activity detection and variable bit rates. Since Speex is robust to loss packets, it is a good candidate for ZigBee communication environment. Due to the lower computational resource requirements and available transmission rate in WSN nodes, the narrowband operation is selected to be implemented in our design. In Speex narrowband mode, there are three major

Fig. 4 Speex 15 Kbps encoding structure

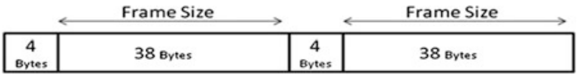


Table 1 Speex frame size

Bit-rate	15 Kbps	11 Kbps	8 Kbps
Sample	300	220	160
Frame size	38 B	28 B	20 B

tasks, including line spectrum pair (LSP) quantization, adaptive codebook search, and fixed codebook search. Depending on the difference of data rates, narrowband provides 8 different sub-modes to handle quantization and un-quantization of the pitch and its gain.

In this paper, we employ three different compression ratios of Speex narrowband mode, including 15, 11 and 8 Kbps, in our designed system. The frame size for Speex narrowband 15 Kbps is 20 ms, which corresponds to 300 samples in a frame. Speex 15 Kbps encoding structure is shown in Fig. 4 where each frame consists of 4 B header to record bit-rate parameters and 38 B voice data payload.

Table 1 lists the corresponding frame sizes for three different bit rates which will decide the number of frames embedded in each packet of Xbee 100 B payload. Figure 5 shows the design of Speex frames in Xbee packets. Without losing the encoding/decoding process and quality for our applications, we design and implement one Speex header only in a Xbee packet with 100 B payload to reduce the overhead of voice data. Therefore, each Xbee packet can carry two, three and four encoded voice frames for 15, 11, and 8 Kbps bit rates, respectively. As notice, the voice quality is affected significantly by the packet loss during transmission and the received signal strength indicator (RSSI) value could be an important and instant indicator related to packet loss. In this study, we will evaluate the relationship between encoding bit rates and RSSI values on voice quality to provide an optimized voice communication over ZigBee networks.

Fig. 5 Xbee voice packet format

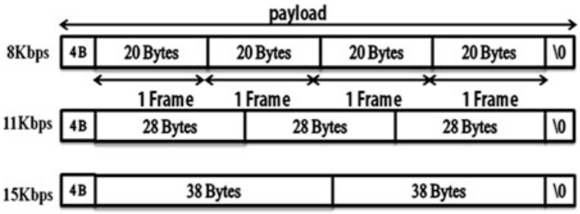


Table 2 Parameters setting

Xbee parameter	Value
Traffic loads	250 Kbps
Serial interface data rate	57,600
Topologies	Point-to-multipoint
Network mode	Beaconless
Data packet size	84, 88, 80 (byte)
Transmit power	1 mW(0 dBm)
Speex parameter	**value**
Mode	Narrowband
BitRate	8, 11, 15 (kb/s)
Quality	4, 6, 8

Table 3 RSSI/LQI measurements

Distance (m)	Relay node (RSSI/LQI)	Receive node (RSSI/LQI)
50 m	−38 dBm/119	−39 dBm/116
80 m	−50 dBm/86	−51 dBm/83
110 m	−56 dBm/68	−58 dBm/63

Performance Evaluation

The implemented parameters for Xbee and Speex setting of the designed test-bed are listed in Table 2.

Network Performance Analysis

In this sub-section, we conduct the RSSI and link quality indicator (LQI) measurements for different transmission distances. For simplicity in test-bed arrangement, the multi-hopping scenario is measured with 3 nodes topology, sending from node 1, relaying via node 2 and receiving at node 3. The RSSI/LQI values detected from relaying and receiving nodes with a series of different distances are measured and three of them, 50, 80 and 110 m, are listed in Table 3. As the transmission distance increase, the LQI value decreases.

The network delay measurements are conducted by processing original audio file using Speex encoding followed by constant packet inter-arrival time for voice data transmission. The measured delays from sending node to relay node, separated in 50, 80, and 110 m respectively, are plotted in Fig. 6. The corresponding delays are 6, 35 and 58 ms. As the transmission distance increases, the delay increases.

On the other hand, the results from sending node (via relay node) to receiving node, separated in 100, 160 and 220 m respectively, are illustrated in Fig. 7, which corresponding to 2-hop transmission. The corresponding measured delays are about 11, 64 and 120 ms.

Fig. 6 The delay from sending node to relay node

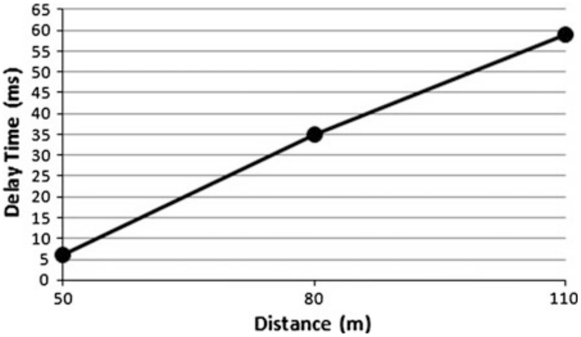
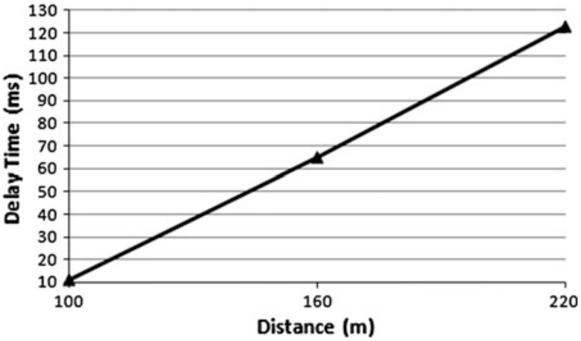


Fig. 7 The end-to-end delay from sending node to receiving node



Voice Quality Analysis

The voice quality in terms of perceptual evaluation of speech quality (PESQ) could deteriorate when the packet loss increases during transmission. As notice, longer transmission distance may incur larger packet loss due to reduction in signal strength and longer delay. In this sub-section, we conduct the experiments on PESQ_MOS voice quality measurements with different LQI values. The voice signal as the input of PESQ measurements is sending from node 1 and extracted from node 3 as the output of PESQ measurements.

The voice quality results are illustrated in Fig. 8. When the LQI value is more than 100, the PESQ_MOS value for 15 Kbps bit rate case is better than those of 11 and 8 Kbps bit rates. The measured PESQ_MOS values are close to the referenced data from Speex project. However, when the LQI drops to 85, the voice quality of 15 Kbps case deteriorates significantly and its PESQ-MOS values is lower than those of 11 and 8 Kbps bit rates. When the LQI even drops to 80 or 70, the voice quality of both 15 and 11 Kbps cases decline dramatically.

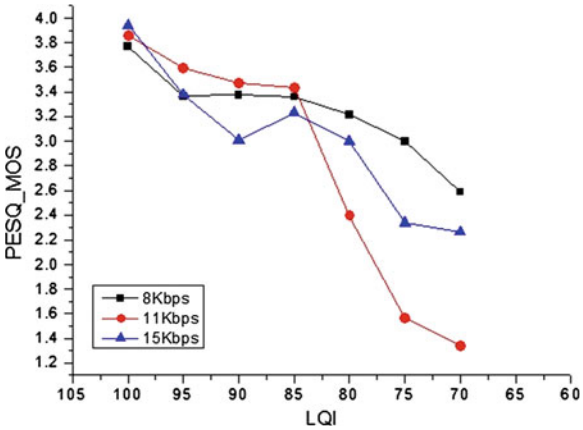


Fig. 8 Voice quality measurement

Conclusions

In this paper, we have designed and implemented the Speex voice codec on embedded system with Xbee WSN nodes using multi-hopping ZigBee communication. The design and implementation issues regarding the Speex codec and Xbee nodes have been addressed and discussed. From our design, we have implemented a multi-hopping ZigBee test-bed for performance analyses in terms of distance-signal strength dependence, network delay and PESQ-MOS voice quality. The design, implementation and evaluation result will provide a test-bed to realize voice transmission using ZigBee communication.

Acknowledgments The authors would like to acknowledge the support from the National Science Council of Taiwan (No. NSC 101-2221-E-142-003, NSC100-2221-E-142-008) and National Taichung University regarding the MoE project (No. 1020035480A).

References

1. Zigbee Alliance (2008) Zigbee specification. Zigbee Alliance. 053474r17 edn
2. IEEE 802.15.4 standard (2003) Wireless medium access control (MAC) and physical layer (PHY) specification for low-rate wireless personal area networks (LR-WPANs)
3. Touloupis E, Meliones A, Apostolacos S (2012) Speech codecs for high-quality voice over zigbee applications: evaluation and implementation challenges. IEEE Commun Mag
4. Speex (2006) a free codec for free speech. Available at http://www.speex.org
5. Xbee Data Sheet (2009). Available at http://www.digi.com
6. Mangharam R, Rowe A, Rajkumar R, Suzuki R (2006) Voice over sensor networks. 27th IEEE international of real-time systems symposium, pp 291–302
7. Song HY, Cho SH (2011) Performances of IEEE 802.15.4 unslotted CSMA-CA for voice communications. The 17th Asia-Pacific conference on communications
8. Brunelli D, Teodorani L (2008) Improving audio streaming over multi-hop Zigbee networks. Proc IEEE ISCC 2008:31–36

9. Brunelli D et al (2008) Analysis of audio streaming capability of zigbee networks. Proc EWSN 2008:189–204
10. Woon W, Wan T (2008) Performance evaluation of IEEE 802.15.4 wireless multi-hop networks: simulation and testbed approach. Int J Ad Hoc Ubiquitous Comput 3(1)

A Lightweight Intrusion Detection Scheme Based on Energy Consumption Analysis in 6LowPAN

Tsung-Han Lee, Chih-Hao Wen, Lin-Huang Chang, Hung-Shiou Chiang and Ming-Chun Hsieh

Abstract 6LoWPAN is one of Internet of Things standard, which allows IPv6 over the low-rate wireless personal area networks. All sensor nodes have their own IPv6 address to connect to Internet. Therefore, the challenge of implementing secure communication in the Internet of Things must be addressed. There are various attack in 6LoWPAN, such as Denial-of-service, wormhole and selective forwarding attack methods. And the Dos attack method is one of the major attacks in WSN and 6LoWPAN. The sensor node's energy will be exhausted by these attacks due to the battery power limitation. For this reason, security has become more important in 6LoWPAN. In this paper, we proposed a lightweight intrusion detection model based on analysis of node's consumed in 6LowPAN. The 6LoWPAN energy consumption models for mesh-under and route-over routing schemes are also concerned in this paper. The sensor nodes with irregular energy consumptions are identified as malicious attackers. Our simulation results show the proposed intrusion detection system provides the method to accurately and effectively recognize malicious attacks.

Keywords 6LoWPAN · Energy consumption · Intrusion detection

T.-H. Lee · C.-H. Wen · L.-H. Chang (✉) · H.-S. Chiang · M.-C. Hsieh
Department of Computer Science, National Taichung University of Education,
Taichung, Taiwan, Republic of China
e-mail: lchang@mail.ntcu.edu.tw

T.-H. Lee
e-mail: thlee@mail.ntcu.edu.tw

C.-H. Wen
e-mail: BCS100102@gm.ntcu.edu.tw

H.-S. Chiang
e-mail: BCS100110@gm.ntcu.edu.tw

M.-C. Hsieh
e-mail: BCS100101@gm.ntcu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_137, © Springer Science+Business Media Dordrecht 2014

Introduction

6LoWPAN is an acronym IPv6 over Low-Power wireless Area Networks. 6lowpan has defined encapsulation and header compression that allows IPv6 packets to be sent to and received from over IEEE 802.15.4 (Wireless sensor network) based networks. The packet size of IPv6 is much larger than the size of IEEE 802.15.4 MPDU. Thus, the adaptation layer is joining between the network layer and the data link layer to fragmentation and recombination the IPv6 packets into IEEE 802.15.4 radio.

There are two routing schemes in 6LoWPAN. And the characteristic of route-over is hop-by-hop. Mesh-under and route-over routing schemes can be considered as end-to-end and hop-by-hop transmission respectively. Hop-by-hop fragmentation and reassembly generates more delay but achieve better fragment arrival ratio. Whereas end-to-end scheme has less latency, but fragment loss has high probability.

In 6LoWPAN protocol stack, the last two layers are based on IEEE 802.15.4 physical and data-link layers. Thus, we investigate the energy consumption of IEEE 802.15.4 for the last two layers in the beginning, and then the 6LoWPAN performance analysis in route-over and mesh-under routing schemes are both consideration in this paper.

In order to construct the exact energy consumption model in 6LoWPAN, many study presented energy model to analysis the energy consumption on sensor node, and propose the method to improve the performance of network. We not only consider the reference of IEEE802.15.4 but also consider the reference of IEEE802.11. The work in [1], the author consider all the behavior and state of sensor to calculate the energy consumption in 802.11, and dynamic altered the routing path if the remain energy of sensor is low. The research in [2, 3] propose the energy consumption mode base on Markov chain. This model consider the multihop and error-prone channel condition to analysis the performance of 802.15.4. And the Fatma Bouabdallah [4] propose the energy consumption mode to calculate the energy consumption and transmission probability on sensor node, and use weight to altered the transmission path base on energy and quality of service.

In [5–9], authors introduced several types of attacks and the method against malicious attack. There has various attack in 6LoWPAN, such as Denial-of-service, wormhole and selective forwarding attack. Based on our research, the Dos attack is the major attack in both WSN and 6LowPAN. The Dos who launch we call the "jammer", the jammer may be is a simple device or a jamming station. And then, we will introduce four generic jammer modes:

- Constant jammer: This kind of jammer transmission large packets to occupies the channel, and make the channel busy cause the normal node can not transmission the packet.
- Deceptive jammer: Deceptive jammer transmission the packet with constant interval and without any gap.
- Random jammer: The random jammer change state between sleeping and jamming in random interval. It is good for the jammer that does not has unlimited power support. And this jammer is also difficult to be detected.

- Reactive Jammer: This jammer always sensing the channel. It always in idle mode when channel is free, and then transmission packet when it sensing there has someone transmission the packet.

The goal of above jammer is make the normal node always turn into receive mode and receive the packet, so the node will exhaust it energy.

There are few previous research works in detecting and against the malicious attack. The work in [9], the author use energy consumption mode to detect intrusion, and it use energy prediction model to identify the type of attacks. The concept of this paper is similar to our research. However, the proposed scheme in this paper based on an accurate energy consumption models in both route-over and mesh-under routing schemes to detect malicious nodes in 6LoWPAN. The goal of this research is to recognize malicious attacks by using energy consumption model compression which is not only based on the energy rising rate, but also take account the node's total energy consumption and traffic loads.

The research work [10] analysis the traffic load on sensor in network, and monitored the traffic load to detect whether is intrusion. Actually, only the sink node has the ability to monitoring all traffic loads, so the proposed detect scheme cannot put in the generic sensor node. Ponomarchuk et al. [11] analysis the energy consumption of sensor node attacked by Dos attack. From the result, it proves the energy consumption is increased when the node be attacked by Dos attack. The research work in [12], author analysis and compare the energy consumption with several Key cryptography in WSN, such as RSA-1024, SHA-1, AES. But actually in WSN and 6LoWPAN, put the key cryptography in node will reduce the data utilization. However, using the key cryptography will decrease the performance of sensor node.

The rest of the paper is organized as follows. Section Energy Prediction Model of 6LowPAN introduces energy prediction model for both route-over and mesh-under routing schemes in 6LoWPAN. Section The Lightweight Intrusion Detection Scheme Based on Energy Consumption Analysis describe the proposed lightweight energy prediction based intrusion detection scheme to recognize malicious attacks. Section Simulation Results shows the simulation results for the proposed scheme. Finally, Conclusion presents our conclusions and suggests the future work.

Energy Prediction Model of 6LowPAN

Energy Consumption Model of Node Transmission/Receive Process

Figure 1 shows the transmission process from source to destination nodes. First at all, the source node will go into a contention period T_C before the packet transmission. In the contention period, node will performed the CCA to check the channel condition. The node is going to transmission the packet when channel is

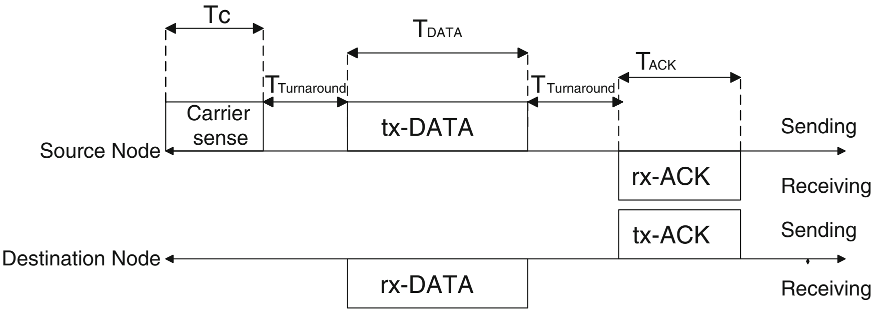


Fig. 1 Node transmission/receive process

free. Otherwise, it will turn to the backoff stage to wait for channel free. Node will wait for a turnaround time ($T_{turnaround}$) to change the mode from transmission to receive mode after the packet transmission. When the destination node receives a data frame correctly, it will send an ACK to indicate the successfly reception after $T_{turnaround}$.

The total energy consumption E_{total} is the total energy consumed in both source and destination nodes.

$$E_{total} = (1 - P_{tr}) \times E_{source} + E_{destination} \tag{1}$$

where, E_{source} is the energy consumption in source node, and the $E_{destination}$ is the energy consumption in destination node. (1 – P_{tr}) represents the packet success transmission probability. The E_{source} can be obtain from Eq. (2).

$$E_{source} = T_C \times E_{\text{sensing}} + T_{DATA} \times E_{tx} + 2 \times (T_{turnaround} \times E_{\text{sensing}}) \tag{2}$$

T_C, T_{DATA}, $T_{turnaround}$ and T_{ACK} are the contention time before node transmission, the time duration of transmission data, the time duration of node change state and the duration of node return ACK respectively. $E_{sensing}$, E_{tx} and E_{rx} are the energy consumption of sensing, the energy consumption of transmission the packets, the energy consumption of receive the packets. The $E_{destination}$ is given by:

$$E_{destination} = T_{DATA} \times E_{rx} + T_{turnaround} \times E_{\text{sensing}} + T_{ACK} \times E_{tx} \tag{3}$$

The $E_{destination}$ include the T_C, T_{DATA}, $T_{turnaround}$ and T_{ACK}. The T_{DATA} is obtained from:

$$T_{data} = \frac{Slots \times Slot_{DATA}}{250\,\text{kbps}} \tag{4}$$

Slots is the number of slots in one packet. $Slot_{DATA}$ is the number of bits in one slot. One symbol period is 16 μs, and one slot is 20 symbols. Thus, one slot is carried 80 bits of data.

$$Slot_{DATA} = 20\text{ symbol} \times 16\ \mu\text{s} \times 250\text{ kbps} = 80\text{ bits} \tag{5}$$

The $T_{turnaround}$ is given by:

$$T_{turnaround} = {}^{(12\text{ symbols} \times 4\text{ bits})}\!/_{250\text{ kbps}} = 192\ \mu\text{s} \tag{6}$$

The contention time T_C is the duration of node performs CCA, it about 8 symbol which is around 128 μs.

Energy Consumption Model for Mesh-Under Routing Scheme

In the energy prediction model for mesh-under routing scheme, the energy prediction model for mesh-under is given by:

$$E_{mesh} = [(P_{tr}^{x+h}) \times x \times E_{total} + \delta \times E_{recombination}] + \sum_{K=1}^{3} K[(1 - P_{tr}^{x+h}) \times x \times E_{total} + \delta] \tag{7}$$

Equation (7) indicates the energy consumption of node when transmit x fragment packets for mesh-under routing scheme in 6LowPAN. P_{tr}^{x+h} is the probability of all fragments successful transmission in h hops count. h, x, k and δ represent the number hops count, the number of fragments, the time of retransmission, the process delay to reassemble all fragments respectively.

Energy Prediction Model for Route-over Routing Scheme

Equation (8) represents the energy consumption of the node for route-over routing scheme in 6LowPAN. Consider the characteristic of route-over, the energy model for route-over is given by:

$$h\{[(P_{trans}^{x}) \times x \times E_{tx-rx} + \delta] + \sum_{K=1}^{3} K[(1 - P_{trans}^{x}) \times x \times h \times E_{tx-rx} + \delta] + T_{frag}\} \tag{8}$$

P_{trans}^{x} is the probability of successful transmission of a fragment in single hop. T_{frag} is the delay time to process fragmentations. The major feature of route-over is hop-by hop fragmentation and reassembly. In each hop, all fragments will recover to a completed IPv6 packet. Thus, we can consider route-over scheme as hop by hop forwarding from source to destination.

The Lightweight Intrusion Detection Scheme Based on Energy Consumption Analysis

In the proposed intrusion detection scheme, all nodes have to morning its own total energy consumption by using our energy prediction models, and the sampling rate is 0.5 s. Base on our simulation results, we assume that node is regarded as attacked if the energy consumption rate is increased over 30 % (*Energy-Rise*$_{threshold}$) from pervious energy consumption record. The intrusion detection scheme will regard the node as malicious and remove the node from the route table in 6LowPAN.

The Intrusion Detection Scheme denote by Table 1.

Simulation Results

In this section, we present the energy consumption analysis for 6LoWPAN routing schemes. Our simulation was emulated by Qualnet [13]. The simulation parameters are shown in Table 2.

Table 1 Intrusion detection scheme

```
Computing Energy Consumption //Computing Energy Consumption of nodes
alarm = 0                    //initial alerm,count the times of attacked
if(Current Time − Pervious Sampling Time ≥0.5)
//0.5 s is the energy consumption sampling rate
{
    Calculate the Current Energy Consumption of node
    Energy consumption RiseRate = (Current Energy Consumption−
    Prediction Energy Consumption)/
    Prediction Energy Consumption × 100
    if (Energy consumption RiseRate > $EnergyRise_{threshold}$)
    //$EnergyRise_{threshold}$ is 30 %,
    {
        alerm++
        if(alerm = = 2)
        //if there is two consecutive alarms, node is regarded as attacked and
        //remove the node from the route table in 6LowPAN.
    }
    else {
        alerm = 0
    }
}
```

Table 2 Simulation Parameters

IPv6 Packet Size	6,400 bytes
Number of Fragments	50
Hop Counts	2 hop
Routing	From node 1 to node 3 through node 2

The Energy Consumption of Normal Traffic Load and Dos Attack Traffic Load

The Fig. 2 is the energy consumption of node in normal and attack scenario from ideal channel. The node with normal traffic load transmits packet from 0.5 s, and the attack node attack from 2 s until normal node transmission end. The normal node and attack node transmission in interval 10 ms and 5 ms, respectively. We monitor the energy consumption each 0.5 s.

In Fig. 2, the energy consumption of node in normal traffic load is stable, and the node transmission end at 4 s. The node transmission end around at 18 s when node is under Dos attack. And the number of total received packets at receiver in normal and Dos attack scenarios are 50 and 0 respectively.

The Nodes Energy Consumption Sampling Rate in the Proposed Intrusion Detection

In this section, we detect the energy consumption every 0.5 s, and the detect threshold is 30 %. The intrusion detection scheme determined that the node has be attack when the energy increase rate over 30 %.

Figure 3 is the total energy consumption in normal and attack traffic loads. The Dos attack has the highest energy consumption 0.373 mJ per 0.5 s than normal

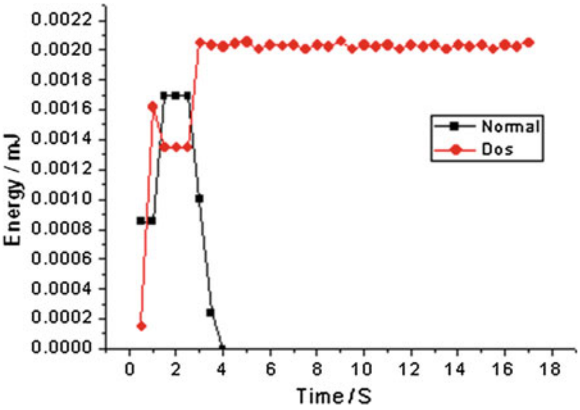


Fig. 2 Energy consumption between normal and Dos traffic loads

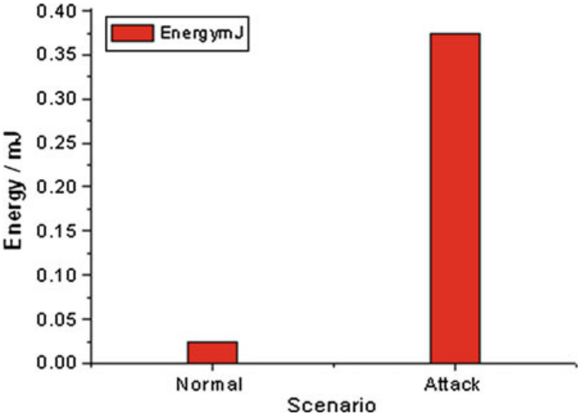


Fig. 3 The total energy consumption in normal and Dos attack traffic loads

traffic load 0.024 mJ/0.5 s due to the malicious node receives lots of attack packets. Furthermore, the probability of channel collision also increase make the retransmission times increase cause from the Dos attack in the malicious node.

Conclusion

The security issue has become more and more important in Internet of things. In this paper, a lightweight intrusion detection scheme for 6LoWPAN is proposed, which based on accurate energy consumption models to detect malicious nodes. The results show the proposed lightweight intrusion dection scheme is more efficient and accurate. From the smulation result, the energy consumption will receive large increasement from Dos attack, and the detection rate is 100 %.

Acknowledgments The authors would like to acknowledge the support from the National Science Council of Taiwan (No. 101-2119-M-142-001, 100-2221-E-142 -002) and National Taichung University regarding the MoE project (No. 1020035480A).

References

1. Lee TH, Marshall A, Zhou B (2005) A framework for cross-layer design of energy-conserving on-demand routing in multi-hop wireless networks. IEE Mobility
2. Performance Analysis of IEEE 802.15.4 MAC Protocol for WSNs in Burst Error Channels (ISCIT 2011)
3. Analytical Modeling of Multi-hop IEEE 802.15.4 Networks (2012). IEEE transactions
4. Bouabdallah F, Bouabdallah N, Boutaba R (2009) On balancing energy consumption in wireless sensor networks. IEEE Trans Veh Technol 58(6)

5. Manju VC, Kumar MS (2012) Detection of jamming style DoS attack in wireless sensor network. Parallel distributed and grid computing (PDGC). 2012 2nd IEEE international conference
6. Mpitziopoulos A et al (2009) A survey on jamming attacks and countermeasures in WSNs. Commun Surv Tutorials, IEEE 11.4(2009):42–56
7. Pelechrinis K, Iliofotou M, Krishnamurthy SV (2011) Denial of service attacks in wireless networks: the case of jammers. Commun Surv Tutorials, IEEE 13.2(2011):245–257
8. Modares H, Salleh R, Moravejosharieh A (2011) Overview of security issues in wireless sensor networks. Computational intelligence, modelling and simulation (CIMSiM), 2011 third international conference
9. Shen W et al (2012) A new energy prediction approach for intrusion detection in cluster-based wireless sensor networks. Green communications and networking. Springer, Berlin, pp 1–12
10. Ponomarchuk Y, Seo DW (2010) Intrusion detection based on traffic analysis in wireless sensor networks. 19th annual wireless and optical communications conference (WOCC)
11. Kim K, Hong J (2010) Analysis of power consumption of S-MAC protocol according to DoS attack. 4th IEEE international conference on new trends in information science and service science (NISS)
12. Wander AS et al (2005) Energy analysis of public-key cryptography for wireless sensor networks. Third IEEE International Conference on Pervasive Computing and Communications, PerCom 2005
13. QualNet simulator. http://www.qualnet.com/

A *k*-Cooperative Analysis in Game-Based WSN Environment

Hsin-Hung Cho, Fan-Hsun Tseng, Timothy K. Shih, Li-Der Chou, Han-Chieh Chao and Tin-Yu Wu

Abstract In wireless sensor networks (WSNs), the coverage problem usually accompanies the energy-saving issue, which ensures the fundamental functions are workable. Therefore, the sensor nodes must be deployed and survived in the target place for a long time, and sustain the trade-off between expected coverage and limited battery energy. The lesser sensor nodes are awake in an epoch, the longer network lifetime is achieved. It shows that the coverage problem must be solved based on the energy efficiency viewpoint. In this paper, we propose a game-based WSNs environment and solve the energy saving issue with the minimum number of competition players, which implies the most suitable duty-cycle for all sensor nodes. In the simulation results, the proposed approach achieves the lowest power consumption and longest network lifetime.

H.-H. Cho (✉) · F.-H. Tseng · T. K. Shih · L.-D. Chou
Department of Computer Science and Information Engineering, National Central University,
Taoyuan, Taiwan, Republic of China
e-mail: hsin-hung@ieee.org

F.-H. Tseng
e-mail: fanhsuntseng@ieee.org

T. K. Shih
e-mail: timothykshih@gmail.com

L.-D. Chou
e-mail: cld@csie.ncu.edu.tw

H.-C. Chao · T.-Y. Wu
Institute of Computer Science and Information Engineering, National I-Lan University,
I-Lan, Taiwan, Republic of China
e-mail: hcc@niu.edu.tw

T.-Y. Wu
e-mail: tyw@niu.edu.tw

H.-C. Chao
Department of Electrical Engineering, National Dong Hwa University,
Hualien, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_138, © Springer Science+Business Media Dordrecht 2014

Keywords Wireless sensor network • Energy saving • Integer linear programming • *k*-cooperative game

Introduction

Since the battery capacity of sensor node is limited and finite, the power consumption issue is always a hot research topic in WSNs [1, 2]. Therefore, the power consumption is the most critical design factor. The duty-cycle is a well-known scheme for saving the energy consumption in entire network, because it limits the sleeping time of sensor nodes in each epoch. Owing to fractional duty-cycle values are too large, the setting of duty-cycle usually accompanies with the coverage issue. If a sleeping time of sensor node is too long, it may make the covered area decreasing due to the insufficiently awake sensor nodes in an epoch. Some researchers have proposed some mechanisms which have the ability to self-adjust duty-cycle value, such as Distance-based Duty Cycle Assignment (TDDCA) [3]. However, the network topology of WSN is usually randomly distributed that may occur various overlapping areas. In order to avoid the misjudgment, the Dynamic Duty-cycle Dynamic Scheduling Assignment (DDDSA) [4] is our afore-proposed scheme. In this work, we consider that the duty-cycle assignment in overlapping area includes the competed phenomenon, therefore the game theory is useful for such environment. Furthermore, the proposed game model is able to apply to any game in wireless networks. For details, we defined the number of links as the players in the game. Then, utilize the number of link to express the competed phenomenon with an unprecedented attempt. In the proposed *k*-cooperative algorithm, we successfully demonstrated the competed phenomenon, decreased more redundant links, and achieved energy efficiency without losing the communicated quality.

This paper is organized as follows. In Related Works, the related literatures are introduced, such as the TDDCA scheme, DDDSA scheme, and the game-based WSNs. One thing should be noticed that the DDDSA scheme and game-based WSNs is our fore-proposed methods. After that, we analyze the game-based WSNs and define a novel *k*-cooperative problem and solve it by the proposed *k*-cooperative algorithm in Game-Based WSNs Duty Cycle Assignment. The Simulation is the simulation result, and Conclusion and Future Work is the conclusion and future work of this work.

Related Works

Traffic-Adaptive Distance-Based Duty Cycle Assignment

The researchers [3] regarded the traffic relay is relative to the distance to the sink and they analyzed the performance of receiver-based mechanism through math models [5]. Receiver-based mechanism contains several elements such as an expected traffic rate and an assignable duty-cycle value. They assume the source sensor nodes are uniformly allocated around the central sink node with the transmission range r_T. The hierarchical concentric circles are composed of n rings, and the circle is $(n-1)r$ away from the sink node with width r_T, which is represented by the nth ring. The traffic generating by the source sensor nodes in the nth ring within the ring outside of the nth ring per unit time is expressed by Γ which is given in [3] as

$$\Gamma = \lambda_g \rho_s \pi (R^2 - [(n-1)r_T]^2), \tag{1}$$

The variable λ_g represents the average traffic generation rate, ρ represents the node density of source sensor nodes, and R represents the radius of the network area. A node in the $\frac{r}{r_T}$ ring with a distance r from the sink node and the number of sensor nodes in the nth ring is defined as

$$N_n = \rho_r \pi \left\{ (n r_T)^2 - [(n-1)r_T]^2 \right\}. \tag{2}$$

The average traffic rate [6] of a node is defined as

$$\lambda_r = \frac{\lambda_g \rho_s \pi \left\{ R^2 - \left[\left(\frac{r}{r_T} - 1 \right) r_T \right]^2 \right\}}{\rho_r \pi \left\{ \left[\left(\frac{r}{r_T} \right) r_T \right]^2 - \left[\left(\frac{r}{r_T} - 1 \right) r_T \right]^2 \right\}}. \tag{3}$$

In [3], the authors consider the transmission time is relative to the duty-cycle assignment. The higher duty-cycle can make the more sensor nodes available in an epoch and avoid the probability for relays to achieve a lower latency. However, the number of awake sensor nodes is increased and the number of idle sensor nodes also increased, which consume more energy. To overcome this problem, they improve the duty-cycle that minimizes the energy consumption for a given traffic rate. Based on CSMA protocol [7], a packet should be sent after it sends a request to send (RTS) packet and waits for the clear to send (CTS) packet from destination. Sensor nodes do not transmit or receive packets which called idle sensor nodes, and the power consumption of idle sensor nodes is approximately the same with active sensor nodes in WSNs [8]. Thus, given a constant value P to represent the idle listening, transmission, and reception. Assuming the topology is a Poisson distribution, and the expected time [5] before RTS/CTS handshake is

$$t_H = \left(e^{\xi dN} - 1 \right)^{-1} \left(T_{RTS} + N_p N_r T_{CTS} \right). \tag{4}$$

The T_{RTS} and T_{CTS} are the delay time for sending the RTS packet or CTS packet respectively.

The expected total time for a complete communication is defined as

$$t_C = T_{RTS} + \chi T_{CTS} + T_{DATA} + T_{ACK}, \tag{5}$$

and the total time consumption is $t_t = t_H + t_C$. The expected energy consumption $\bar{\mathcal{P}}$ is defined as

$$\bar{\mathcal{P}} \simeq \mathcal{P}(d + \lambda_r t_H) + \left(2 - e^{-\xi dN}\right)t_t. \tag{6}$$

The TDDCA is based on above-mentioned ideas that assign the duty-cycle value dynamically. In TDDCA, each node has a counter which records the numbers of the initial and retransmitted RTS packets. If received retransmitted RTS packets in the current epoch outweighs the total number of the received initial RTS packets that expresses the neighborhood of this node in a severe situation. At this moment, the duty-cycle of this node should be increased for reducing the traffic load. Otherwise, the duty-cycle should be subtracted 1 % to eliminate the power consumption.

Dynamic Duty-Cycle and Dynamic Scheduling Assignment

The researchers proposed the DDDSA scheme [4] to allocate the duty-cycle value and reduce the power consumption in WSNs. The proposed scheme improves the uncorrected information of sensor nodes within the overlapping area in TDDCA. They utilized the idea in TDDCA which using the ring to define the distance-based network, and modified the function of counter in each sensor nodes. The original counter records the number of received retransmission RTS packets. Unlike TDDCA, the difference between the previous and current received retransmission RTS packets is calculated in DDDSA. In order to deal the sensor nodes in the overlapping area, they defined the information from each neighbor as a priority. Then, they used this information as a priority to prioritize which sensor nodes need to be served first. If a sensor node must provide service for a low priority neighbor, it must postpone to a few seconds. If there is no higher priority sensor nodes need to be served, the sensor nodes in overlapping area start to work. Although the problem of overlapping area is solved, the duty-cycle value is not the most appropriate for each sensor nodes due to 1 % per stage. In other words, the issues of power consumption and traffic load are solved imperfectly.

Game-Based WSNs

The main concept of game theory is that the two or more players compete for the reward thus achieves a fair situation in whole environment [9]. In recent years, the researchers apply game theory to wireless networks [10]. In this paper, we model all of the duty-cycle values of sensor nodes as the players of game theory. The duty-cycle value is a decisive factor for the WSNs, because the power consumption is based on the number of awake sensor nodes. The sensor nodes whether enter into the sleep state must decide by the game, and each sensor nodes have to come up with their own information. It should be mentioned that the game in this work only happens in the overlapping area, and each sensor node owns an initial duty-cycle value. The sensor nodes cannot adjust their duty-cycle value arbitrarily in order to maintain the fairness. Since the summation of changed duty-cycle value is not zero, it belongs to the non-zero-sum game in game theoretic. In brief, the game-based WSNs model is responsible for coordinating the duty-cycle vale of sensor nodes within the overlapping area.

Game-Based WSNs Duty Cycle Assignment

The non-zero-sum game represents that the profits of players are not equal to zero, which could be a win–win result. In our environment, the main object is decreasing the power consumption and extending the network lifetime by controlling duty-cycle value. The influence of duty-cycle value covers the multi-hop situation, even the entire network. It implies that a player not only considers its cost but also the cost of its neighbors. According to the duty-cycle value is a regional variable, each player just concerns the current information from itself and its neighbors. If we want to set the adjusted duty-cycle value as a reward to the winner, we must consider its fairness. Because we assure the duty-cycle value will not affect the other player's game. Therefore, we must find out the number of players for each round of the game. The follow-up two subsections, we will introduce what is the *k*-cooperative, and clearly define the proposed problem. For better readability, we have listed all the symbols in Table 1 before next subsection.

Outline of k-cooperative

We assume the duty-cycle assignment problem is a *k*-cooperative problem, and the constant *k* represents the numbers of player who joined this subset of game. They are responsible for the most suitable duty-cycle value from the competition in this game subset. The proposed *k*-cooperative problem is similar to the well-known *k*-

Table 1 Definition of the improtant notations

Variable	Definition
h	The set of the received RTS from its neighbors
L	The set of the link between two sensors
C	The set of the traffic information which through the such sensor nodes
X	The set of the same link from its neighbors
λ_{μ}	The average traffic of a node in overlapping area u
T	The threshold for the traffic

connectivity problem [11]. The concept of k-connectivity problem is simply defined as follows. Given a graph $G = (V, E)$, N is a set of minimum connected k-neighbors, and $N \subseteq V$. In short, there are least k neighbors for a node in such network topology, and this node obtains information from these k neighbors only. In game-based WSNs, only one node and its neighbors joined a round game, and the entire game model is composed of several rounds. Depend on above definition, we may encounter some problems: (1) The node obtained duty-cycle value may imbalance for the global network topology; (2) For a high-density environment, the node has a large number of neighbors so that it leads the game model operation become inefficient. In order to solve these problems, we try to limit the constant k so that the imbalance of duty-cycle value will be decreased and maintained a certain level of performance. The problem should be abided by the principle that each node has a relation with at least one node. Try to decrease the constant k from a game, thus the sum of k of all game is the minimum that achieved by greedy method is our major goal.

Definition of k-Cooperative Problem

In this subsection, the clear definition of k-cooperative problem is stated and introduced. First of all, we consider the constant k is a crucial factor for the network performance. Because there may be two or more neighbor nodes in the coverage of one sensor node, this node is able to acquire the information from its neighbors. Moreover, we can use fewer sensor nodes to complete a round of the game. However, we have to find out the most appropriate value of k such that the game will be carried out smoothly and without losing fairness.

We have to define a correlation value $X_{i,j}$ to ensure each node has least one correlation with its each neighbor. The correlation value also affects the connection relation of whole topology directly. If it is well defined, the repeated computing sensor nodes will become less. However, every sensor nodes still maintain an indirect connection relation with each node. The indirect connection relation is that a node is the one hop and also n-th hops neighbor of another node in the same time. For example, when a node A has two neighbors B and C, and the neighbor B has a neighbor D which is also a neighbor of node A. Herein, the neighbor B is the one hop and aslo two hops neighbor of node A in the same time. Wherein above situation, the value $X_{A,B}$ is equals to 2. If the $X_{A,B} \geq 1$, it represents that there is a connection relationship between these two sensor nodes. In order to decrease the operation cost, we try to let the $X_{i,j} = 1$ for every two sensor nodes. If we achieve this goal, it means that every node joins one game only. However, a node has some neighbors which correlation value $X_{i,j} > 1$, this value will be an indicator filtering out a part of the sensor nodes which have larger value, because they have more connection chance to join other game. We can use the set theory to find out the correlation value as follow.

$$X_{i,j} = |N_i \bigcap N_j| \tag{7}$$

$N_{i,1} = \{n_1, \ldots, n_k\}$ is the set of one hop neighbors of *i*st sensor nodes, and $n_j \in N_{i,1}$. The $N_{j,1}$ is the set of one hop neighbors of *j*st sensor nodes, and they are also two hops of N_i, we denoted by $N_{i,2}$.

In order to solve this problem, we formulate the *k*-cooperative problem based on the Integer Linear Programming (ILP). The ILP for the *k*-cooperative problem is defined as follows.

Minimize

$$\sum_{i=1}^{n} k_i \tag{8}$$

Subject to

$$k_i \geq 1 \tag{9}$$

$$X_{i,j} > 0 \tag{10}$$

$$\lambda_\mu \geq 0, \text{ for } \forall u \in \{0, 1\} \tag{11}$$

$$0 \leq x_i \leq 1, \text{for} \, \forall i \in V \tag{12}$$

k-Cooperative Algorithm

Before introducing the *k*-cooperative algorithm, we firstly design an initialization algorithm to create the basic network model and let each sensor nodes obtain the required information. Since we need both of the neighbor's and two hops neighbor's information, the two times broadcasts are unavoidable. Firstly, each sensor node broadcasts the own information to its neighbors, so that any sensor nodes have the required information of its neighbors in this moment. Secondly, each sensor node broadcasts again, but its contents have been substituted with the neighbors' information rather than the own information. Then, any sensor nodes not only have one hop neighbor's information but also have two hops neighbor's information via the two times broadcasts. In our proposed method, we consider that each link include a relationship with a neighbor. Hence, we use of these results to calculate the intersection of one hop neighbors and two hops neighbors. These results represent the cooperative relationship between each sensor nodes and such results can let the *k*-cooperative algorithm operate smoothly. The pseudo code is shown in Table 2.

Table 3 is the *k*-cooperative algorithm which also the main idea of our mechanism. Although our motivation and goal is the minimum cost, the maintenance of link quality is also an important point to ensure the execution of WSNs. Besides, this research is based on the distance-based WSNs and the game only happens in the

Table 2 Initialization of *k*-cooperative algorithm

k-cooperative *Initialization*

1. Broadcast RTS to its neighbors, and receive RTS from its neighbors
2. Broadcast h
3. Find out the intersection of h from each 1-hop neighbors then save them to $\mathrm{X}_{\mathrm{u,n}}$
4. Next we calculated the same neighbors for u that are based on the intersection of $\mathrm{X}_{\mathrm{u,n}}$ then save the traffic information which through the such sensor nodes to C
5. Run the *k*–cooperative algorithm

Run the following at each node u

overlapping area. We still consider the average traffic λ_{μ} of a node in overlapping area u as an initial condition for *k*-cooperative algorithm. If the λ_{μ} is higher than the threshold T, we compare the link situation $X_{\mu,n}$ by calculating intersection h. A link between two sensor nodes will be removed if these two sensor nodes have another path to communicate with each other. If a $L_{\mu,a}$ of $X_{\mu,n}$ that the sensor node a has maximal number of received RTS, the $L_{\mu,a}$ will be removed due to the enough neighbors of sensor node a. In the other case, if two arbitrary sensor nodes only have one same link, then we use of the traffic as a metric to decide which link must be removed. A simple assumption is that if the lower traffic through a sensor node, this sensor node does not need more link to relieve traffic, which shown as step 12. It represents that if more traffic through a sensor node, it needs more neighbors to play the game in order to coordinate out an optimal duty-cycle value.

Simulation

Parameters Setting

In order to check the performance of *k*-cooperative algorithm, we conduct extensive simulations in MATLAB [12]. The network size is 500 (meter) multiplies 500 (meter), which also equals to 250,000 (square meter). The number of deployed sensor nodes ranges from 100 to 1,000. Moreover, we compare the packet delivery ratio within the threshold T from Γ to $\Gamma/3$ in Fig. 3. The initial sensing radius of sensor node is 60 (meter) and initial power is 200 mAh. The energy consumption of a sensor by transmitting, receiving one byte data are 0.0144 and 0.00576 (mJ) [13].

Simulation Results

In Fig. 1, we compare the power consumption of all sensor nodes. It is clear that the higher number of sensor nodes leads to more power consumption, and the gap between original game and *k*-cooperative vice versa, because more sensor nodes

Table 3 *k*-cooperative algorithm

k-cooperative algorithm

```
If $\lambda_\mu > T$
  If $| X_{u,n} | > 1$
    for $i = 1:n$
      for $u = 1:| X_{u,i} |$
        If $|h_{index}| = Max(X_{u,i}(|h|))$
          Remove $L_{u,\ index}$ of node u
        end if
      end for
    end for
  else if $| X_{u,n} | = 1$
      for $u = 1:| X_{u,i} |$
        If $|C_{index}| = Min(X_{u,i}(|C|))$
          Remove $L_{u,\ index}$ of node u
        end if
      end for
  end if
end if
```

Run the following at each node u

bring up more links. However, we can decrease the number of links without losing the availability of the game via *k*-cooperative algorithm. It is well known that the consumption of power is based on the initialization of link establishing and the process on those links. The result shows the effectiveness of our proposed algorithm.

The Fig. 2 shows the simulation result of the network lifetime between the original game-based WSNs and *k*-cooperative WSNs. According to the result of Fig. 1, the number of links affects the length of network lifetime directly. This result represents that the decreased links achieve lower power consumption, which saves more power average on every nodes.

Fig. 1 Comparison of power consumption between original game-based WSNs and *k*-cooperative WSNs

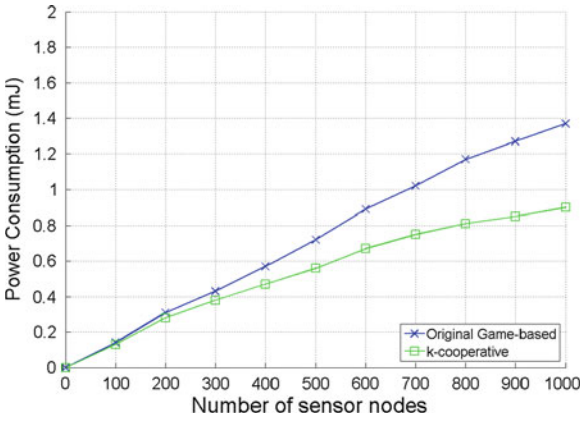


Fig. 2 Comparison of network lifetime between original game-based WSNs and *k*-cooperative WSNs

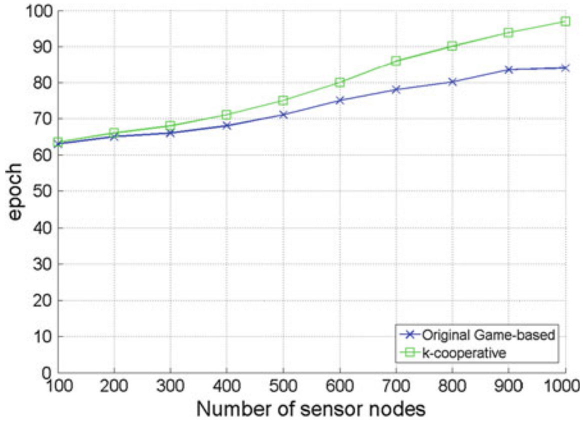
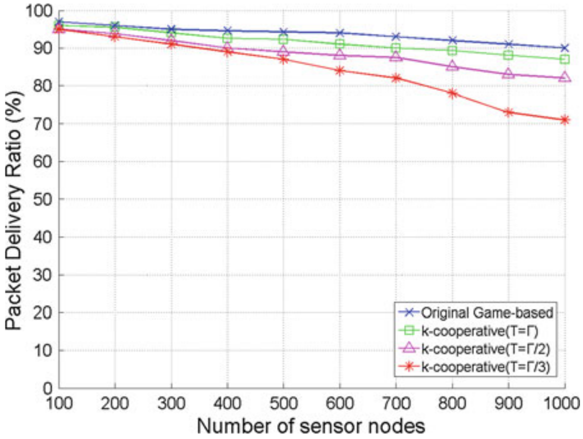


Fig. 3 Comparison of packet delivery ratio between original game-based WSNs and *k*-cooperative WSNs



The Fig. 3 reflects the packet delivery ratio between original game-based WSNs and *k*-cooperative WSNs. In this simulation setting, we investigate the threshold T, because this parameter is a key factor for our proposed algorithm. If the threshold T is well-defined, the result between original game-based WSNs and *k*-cooperative WSNs are almost the same ($T = \Gamma$). Obviously we see that the smaller the threshold ($T = \Gamma/3$), the more redundant links will be removed that there are not enough links to relieve traffic. On the other hand, if we set a bigger threshold T, some redundant links will not be removed so that the better packet delivery ratio is acquired. Thankfully the result just equal to the original game-based WSNs, because our proposed method is a subset of the original game-based WSNs.

Conclusion and Future Work

In this paper, we proposed a novel method viz *k*-cooperative algorithm to decrease the redundant links and eliminate from energy exhaustion. We not only consider to the energy efficiency issue but also ensure the communication quality which only 2 % of the gap with original game-based WSNs. The simulation results show that the proposed algorithm achieves the lower power consumption and longer network lifetime, with slightly influence on packet delivery ratio only. In the future work, the exquisite analysis of threshold *T* will be investigated and studied to achieve the better performance.

Acknowledgments This research was also partly funded by the National Science Council (NSC) of the Taiwan under grants NSC 101-2221-E-197-008-MY.

References

1. Aziz AA, Sekercioglu YA, Fitzpatrick P, Ivanovich M (2012) A survey on distributed topology control techniques for extending the lifetime of battery powered wireless sensor networks. IEEE Commun Surv Tutorials 15(1):121–144
2. Suhonen J, Kohvakka M, Kaseva V, Hämäläinen TD, Hännikäinen M (2012) Low-power WSN technology," in Low-power wireless sensor networks, chap. 1, Springer US, pp 1–6
3. Zhang Y, Feng CH, Demirkol I, Heinzelman WB (2010) Energy-efficient duty cycle assignment for receiver-based convergecast in wireless sensor networks. In: Proceeding of the IEEE global telecommunications conference (GLOBECOM), Miamim, Florida, USA, pp 1–5
4. Cho H-H, Chang J-M, Chen C-Y, Huang S-Y, Chao H-C, Chen J-L (2011) An energy-efficient dynamic duty-cycle and dynamic schedule assignment scheme for WSNs. In: Proceeding of the IEEE Asia-Pacific services computing conference (APSCC), Jeju, Korea, pp 384–388
5. Zorzi M, Rao RR (2003) Geographic random forwarding (GeRaF) for Ad Hoc and sensor networks: multihop performance. IEEE Trans Mob Comput 2(4):337–348
6. Merlin CJ, Heinzelman WB (2010) Duty cycle control for low-power-listening MAC protocols. IEEE Trans Mob Comput 9(11):1508–1521
7. Ziouva E, Antonakopoulos T (2002) CSMA/CA performance under high traffic conditions: throughput and delay analysis. Comput Commun 25(3):313–321
8. Crossbow Technology. Available at http://www.xbow.com/. Accessed 2 July
9. Politis C, Dixit S, Lach HY, Uskela S (2004) Cooperative networks for the future wireless world. IEEE Commun Mag 42(9):70–79
10. Brown DR, Fazel F (2011) A game theoretic study of energy efficient cooperative wireless networks. J Commun Networks 13(3):266–276
11. Penrose MD (1999) On k-connectivity for a geometric random graph. Wiley Random Struct Algorithms 15(2):145–164
12. MathWorks. Available athttp://www.mathworks.com/. Accessed 2 July
13. Liang W, Chen B, Yu XJ (2008) Response time constrained Top-k query evaluation in sensor networks. In: Proceeding of 14th IEEE international conference on parallel and distributed systems (ICPADS), Melbourne, Australia, pp 575–582

An Enhanced Resource-Aware Query Based on RLS-based SIP Presence Information Service

Jenq-Muh Hsu and Yi-Han Lin

Abstract Many innovative services over the standardized SIP-based IMS signaling infrastructure have been widely deployed in the next-generation converged networks. The presence is a key feature for designing context-ware network service. It can efficiently reflect resource availability through subscription and notification mechanisms. Thus, it can apply the presence service to design a resource list server (RLS) to collect a set of resource information, such as resource availability and allocation in order to provide an efficient resource reporting for resource subscribers. In this paper, an improved RLS-based Resource-Aware Query Routing scheme is proposed, in which each server can obtain its neighbors' utilization states notified from RLS. Besides, the proposed scheme is capable of minimizing message traffic between local presence server and its neighboring presence servers while maintaining full responsiveness for the resource subscribing management.

Keywords Resource list server · Resource-aware query · Presence service

Introduction

With the rapid development of mobile services, mobile service providers will shortly deploy the SIP-based [1] mobile services based on IP Multimedia Subsystem (IMS) [2] framework for next-generation services. Such services can

J.-M. Hsu (✉)
Department of Computer Science and Information Engineering,
National Chiayi University, Chiayi, Taiwan, Republic of China
e-mail: hsujm@mail.ncyu.edu.tw

Y.-H. Lin
Department of Computer Science and Information Engineering,
National Chung Cheng University, Chiayi, Taiwan, Republic of China
e-mail: lyha94@cs.ccu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_139, © Springer Science+Business Media Dordrecht 2014

exploit the joint presence information between a consumer and a vendor, such as a subscriber (i.e. SIP entity requesting a service) and a notifier (i.e. SIP resource providing a specific service). Presence information is a collection of contextual attributes within a discovery of available and suitable services. Service discovery has been a topic of research and standardization activities for quite many years. It also allows to providing a service by appropriately composing the service functionality based on the other available service functionality. Hence, service discovery is an essential piece within a service provisioning architecture.

Presence is a key feature of the context-aware applications. Presence-based applications typically leverage upon contextual information such as location, availability, schedule, and local information, such as "Gas Station Finder" or "Restaurant Finder" services, for providing the nearest available resource information to the requesters based on exploiting the presence attributes of requesters or of known end-points.

Presence information actually is stored across a distributed set of presence servers due to the issues of information locality. That is, the queries satisfying the requesting predicates on presence attributes are often routed to multiple servers finding suitable resources for requesters. Thus, an efficient context-aware query mechanism is needed to effectively find the resource information for satisfying the resource requirement from requesters. This paper adopts the features of Resource-Aware Query Routing (RAQR) [3] to proposed a LRS-based RAQR for efficiently query the resource information among distributed SIP presence services.

The rest of the paper is organized as follows. Related Work briefly introduces the related work. The proposed RLS-based resource-aware query routing is presented in The RLS-Based RAQR Scheme for LRF Applications. Experimental Result shows the experimental results and performance evaluation. Finally, a conclusion is made in Conclusions.

Related Work

A minimization of unnecessary notification traffic in IMS presence system is presented in [4] for reducing the SIP message traffic and enhancing the SIP-based service efficiency. In which, an optional resource list server (RLS) can manage all subscriptions of presentities on a resource list to package and forward the presence information of all presentities in bundles to all authorized watchers according to their subscription preferences. Therefore, it can efficiently reduce the SIP messages among user agents and application services through the packed bundle of the presence information. In addition, SIP event throttling [5] is also efficiently reduced the SIP traffic messages. It uses the RLS connecting a set of event resources and applies a throttle mechanism to limit the ratio of SIP event notification among the RLS and watcher applications. Therefore, event throttling mechanism can reduce the number of event notification messages without increasing the message sizes. However, a water application initiates the event throttling mechanism. It will lose the frequently-updated information due to the message throttling. A proper throttling rate of SIP message events will be acted as a trade-off between message reduction and information loss.

Presence-based application typically leverage upon context-aware attributes, such as location, availability, schedule, and local information. It can provide the nearest and available resources or services to the mobile users according to the contextual information of requesting users and responding context-aware information from presence application. They are generally called Live Resource Finder (LRF) applications [3]. The innovation of LRF application can increase the matching degree of service requests according to the location of resources and other contextual attributes.

Resource-Aware Query Routing scheme for LRF-based applications is proposed in [3]. It uses the spatial-temporal distribution and consumption of resources and forwards the LRF queries to alternate presence servers. It can efficiently avoid the flooding problem of query and presence information and uses the Quality-of-Response (QoR) metrics to choose the proper presence server to match the LRF query for responding the available resource according to the current contextual information. For absorbing the advantages of RAQR scheme, this paper adopts the features of RAQR scheme to propose an enhanced resource-aware query based on RLS-based SIP presence service for efficiently improve the query qualities of resource requesting for mobile users.

The RLS-Based RAQR Scheme for LRF Applications

In this section, the proposed RLS-based RAQR scheme is illustrated by using SIP events for LRF applications. The main feature of RLS-based RAQR scheme is the utilization of state notification on a Presence Server (PS) list that points to the presence servers having a surplus of available resources. Resource utilization state is indicated that it facts a crunch state in its local resources. A presence server also maintains additional information, such as information expiration for up-to-dating the newest presence information. In addition, each RLS is associated with the other presence server on a PS list and tracks the resource utilization states of presence servers located in the same domain for evaluating to select the proper available resources.

For obtaining the PS list, the available supplier spool of each presence server should be large enough. Thus, the proposed RLS-based RAQR employs a simple SIP event notification where the originating presence server (the one facing the crunch) delivers a SUBSCRIBE message (informing its need for a certain resource) with expiration time with zero value, which is sent to the local RLS (managing all subscriptions to presence servers on a PS list). The local RLS also subscribes to the state of the other non-local presence server, and thus SUBSCRIBE message is forwarded to the remote RLSs in different domains. On receiving the SUBSCIBE responses, the local RLS will uses the QoR values of the presence servers to determine the proper presence servers that can satisfy the resource load of its request. The local RLS finally responses to the originating server with a NOTIFY message, offering a list of presence servers containing the information of available resources.

Figure 1 shows the SIP message flow used in RLS-based RAQR scheme. The RLS relieves the watcher from subscribing to and managing notifications for all addresses in its contact list. Upon the PS list stored on local RLS, the RESERVE message is issued to PS1 and the utilization state of PS2 is set to SURPLUS in the same domain. The local RLS also subscribes the states of the other non-local presence servers. And then SUBSCRIBE message with an asking rate denoting the additional resources required by the user is forwarded to the remote RLS in different domains. Since PS1 and PS2 fall into the surplus state with available resources, they respond with NOTIFY messages which responding rates offered by r1 and r2 respectively. Remote RLS provides the PS list in which the available supplier pool of each server is large and enough used by another NOTIFY message. In the local RLS, while receiving the SUBSCIBE responses, it exploits the QoR values of presence servers to determine the proper set of presence servers satisfying the load of resource requests. Finally, the subscriber will receive a NOTIFY message with a list of presence servers indicating the available resources.

Experimental Result

In this section, the experiments are performed to evaluate the performance of the proposed RLS-based RAQR scheme [6]. For evaluate the performance of RLS-based RAQR relative to alternative available resource discovery manners, it performs extensive simulation-based studies.

In the experiment, an overlay network of 25 presence server were connected another ones generated by BRITE (http://www.cs.bu.edu/brite/). Each RLS manages all subscription to five presence servers. The coverage area of the entire service provider is represented by a (10,000 × 10,000) grid and the presence server manages an identically-sized partition of the grid. If a particular service is matched to a customer's query, its state will be changed from "available" to "unavailable" during the service time. At the end of the duration, the service' state will be resumed to available at the presence server which originated the query.

Parameter settings in the experiment are depicted in Fig. 1. For the resources, it assumes that upper threshold T_H is $20T$, lower threshold T_L is $10T$ and resource distribution follows an exponential distribution with a mean of 25. All periods are expressed as a factor of T, where T corresponds to the length of a clock tick in the simulation.

Table 2 shows the experimental result of the simulation setting illustrated Table 1. Total number of query divided by number of presence servers equals to the average arrival query per presence server. Thus, average arrival query per

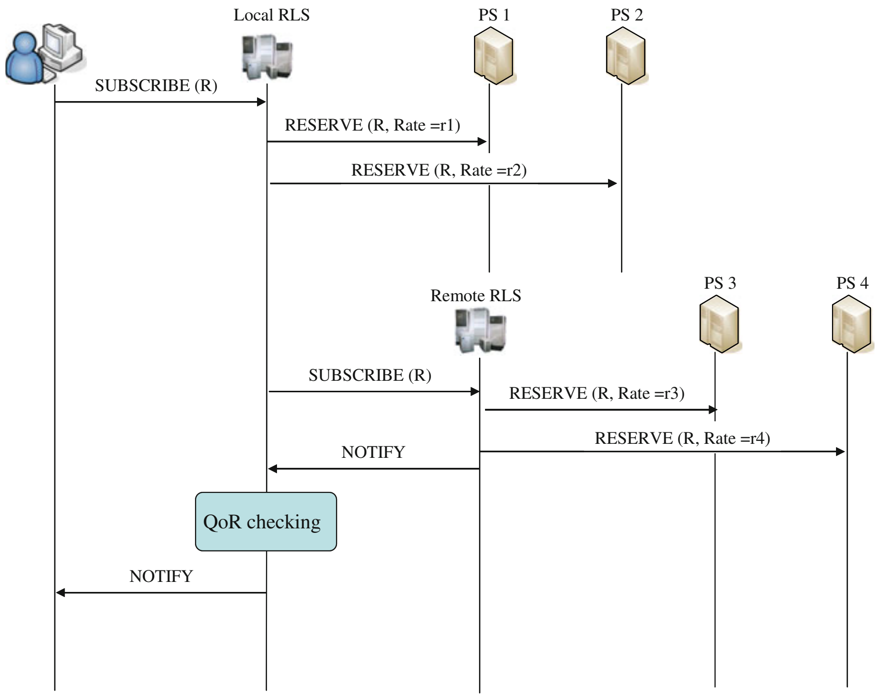


Fig. 1 SIP message flow used in RLS-based RAQR scheme

Table 1 Parameters setting of simulation

Parameter	Value
Number of presence server	25
Number of domain	5
Total number of resources	641
Mean of resource distribution	25
T_L	10
T_h	20
Mean of inter arrival time per query (λ_{qat})	1
Mean of service time per resource consumption (λ_{rs})	0.5
End time	60
Mean of the number of resource consumption per query (λ_{noorr})	10

presence server is 61.12. That is, arrival query per presence server is also equal to the value that λ_{qat} multiplied by end time is 60.

In the experiment, two discovery manners querying available resource are used to compare the difference of evaluating performance: RLS-based RAQR which LRF queries are forwarded from RLS to alternate presence server and RAQR without RLS which LRF queries are routed from originating PS to alternate presence server. The result indicates that RLS can be used for reducing signaling load since the

Table 2 Experimental result

Result	RAQR without RLS	RLS-based RAQR
Message number	1,731	636
Total number of query	1,528	1,528
Number of query miss	266	167
Query miss rate	0.17	0.11
Delay time	145.76	56.94

Presence User Agent (PUA) subscribes to the resource list on the RLS. Instead of subscribing to all members in a list, the PUA can subscribe to the PS list that conveys utilization states of individual PS. Hence, the reducing number of SIP messages in RLS-based RAQR will be better than which in RAQR without RLS. A missed query is occurred while there are no available resources locally and with a resource crunch. The result indicates that RLS-based RAQR leads to the more query matches better than RAQR without RLS. As the same result, the delay time of resource requesting in RLS-based RAQR is also much few better than RAQR without RLS.

Conclusions

In this paper, an enhanced resource-aware query based on RLS-based SIP presence service is proposed. The main feature of the proposed scheme utilizes the state notification in a PS list reflecting the resource state of PS server whether it has sufficient available resources or not.

The main advantage of adopting RLS is to package the presence information of all presentities in bundles and forward it to all authorized watchers. Thus, the RLS-based RAQR scheme is capable of minimizing message traffic between local presence server and its neighboring presence servers while maintaining full responsiveness for the resource subscription management. For roaming users, the number of messages between visited and home network will be effectively reduced for packing the various individual notification messages into a packaged notification message.

In the future, additional contextual attributes will be considered, such as localized information, resource popularity and user preferences, in our proposed scheme to enhance the computing of resource selection accurately.

References

1. Rosenberg J, Schulzrinne H, Camarillo G, Johnston A, Peterson J, Sparks R, Handley M, Schooler E (2002) SIP: session initiation protocol. IETF RFC 3261. Available at http://www.ietf.org/rfc/rfc3261.txt
2. Zhuang W, Gan YS, Loh KJ, Chua KC (2002) Multi-domain policy architecture for IP multimedia subsystem in UMTS. In: The IFIP TC6/WG6.2 and WG6.7 conference on network control and engineering for QoS, security and mobility, vol 235. pp 27–38

3. Chakraborty D, Dasgupta K, Misra A (2006) Efficient querying and resource management using distributed presence information in converged networks. In: The 7th international conference on mobile data management (MDM'06)
4. Wegscheider F (2006) Minimizing unnecessary notification traffic in the IMS presence system. In: The 1st international symposium on wireless pervasive computing
5. Niemi A (2007) Session initiation protocol (SIP) event notification extension for notification throttling, IETF Draft. Available at http://tools.ietf.org/id/draft-niemi-sipping-event-throttle-05.txt
6. Lin YH (2007) Efficient resource subscribing management using distributed presence information in converged networks. Mater Thesis, National Chung Cheng University

Part XIV
Networking and Applications

Authentication of Real-Time Communication System Using KIS Scheme

Binayak Kar and Eric Hsiao-kuang Wu

Abstract In global communication environment, signature computation will be frequently performed on a relatively insecure device that cannot be trusted all times to maintain the secrecy of the private key. To deal with this, Dodis et al. [1] proposed a strong key-insulated signature schemes whose goal is to minimize the damage caused by secret-key exposures. This environment will become more important when we focus on real time communication like telephony, TV shopping, electronic voting etc. Any flaws in the authentication system cause a critical damage to the real time environment. Considering this scenario we proposed a KIS scheme based on elliptic curve cryptography, which minimizes the damage of key exposer. Its security is based on elliptic curve discrete logarithm problem (ECDLP) assumption, and efficient in terms of computational cost and signature size.

Introduction

Real-time systems (RTS) [2] are systems, intended for the interaction (in the first turn, for control) with real physical objects and this interaction must run at a real physical time, in those temporary scales, in which live these objects. There are distinguished Soft-RTS (violation of timing restrictions in some range does not lead to the system failure), Hard-RTS (violation of timing restrictions leads to the system failure) and Firm-RTS (violation of timing restrictions in some range does not lead to the system failure with certain probability).But, communication is a

B. Kar (✉) · E. H. Wu
Department of Computer Science and Information Engineering, National Central University, Jhongli, Taiwan
e-mail: binayakar@wmlab.csie.ncu.edu.tw

E. H. Wu
e-mail: hsiao@csie.ncu.edu.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_140, © Springer Science+Business Media Dordrecht 2014

very important issue for RTS. Consider any of the real time communication service like telephony, online shopping, TV shopping, electronic voting or any broadcasting services. We are handling all these communication by some insecure devices. Several Signature scheme has been proposed for the secure authenticated communication by using several cryptographic mechanisms. But these insecure devices cannot be trusted all times to maintain the secrecy of the private key of the scheme. Dodis et al. [1] proposed a Key-Insulated signature scheme whose main goal is to minimize the damage caused by the secret key exposes.

Considering various real-time communication services [3], we have designed a key-insulated signature scheme that can provide a secure authentication service. In which if the signing key of the signer is leaked still the adversary cannot generate the original signature as the signing key is valid for particular time period and after that it is get updated. The verifier can check the validity of the signature using the same verification key. The verification key is not necessary to update regularly as the signing key; this reduces the communication overhead. Because when the signer update the signing key, if it will update the verification key then it has to transmit the verification key to the receiver followed by the certificate revocation list. Then the receiver will access the certificate authority to check the validity of the verification key. Strong key-insulated signature scheme facilitate the mechanism to update the signing key without updating the verification key. We have used elliptic curve cryptosystem to design our algorithm; as an elliptic curve group could provide the same level of security afforded by an RSA-based system with a large modulus and correspondingly larger key; e.g. a 256-bit ECC public key should provide comparable security to a 3,072-bit RSA public key and the size of a DSA public key is at least 1,024 bits, whereas the size of an ECDSA public key would be 160 bits [4].

The remainder of this paper is organized as follows. Related works are presented in Related Works. Mathematical Primitives describes mathematical primitives for building our protocol. Basic Model of KIS Scheme presents the basic model of KIS scheme. We will describe the new KIS scheme in KIS Scheme for Real-Time System. We give the correctness proof of our proposed scheme and analyze the performance and security in Analysis of the New KIS Scheme. The paper is concluded in Conclusions.

Related Works

Key-insulated cryptosystems [5] focused on the case of public-key encryption. The goal of key-insulated security is to minimize the damage caused by secret-key exposures. Now in this electronic era digital signing is at the heart of Internet based transactions and e-commerce. In this global communication environment, signature computation will be frequently performed on a relatively insecure device (e.g., a mobile phone) that cannot be trusted to completely (and at all times) maintain the secrecy of the private key. In [1] Dodis et al. focused on key-leakage of digital signatures. They proposed a strong key-insulated signature scheme. In which they construct strong $(N-1; N)$-key-insulated schemes based on any trapdoor signature scheme; that leads to very efficient solutions based on, e.g., the RSA assumption in the random oracle model. In [6] Deleito et al. proposed a new strong and perfectly key-insulated signature scheme, more efficient than previous proposals and whose key length is constant and independent of the number of insulated time periods. It is a forward-secure scheme in which an adversary needs to compromise an user at a second time period before being able to compute future secret keys.

Broadcasting and communications networks are used together to offer hybrid broadcasting services which incorporate a variety of personalized information from communications networks in TV programs. These services have many different applications that run on user terminals. Malicious service providers might distribute applications which may cause user terminals to take undesirable actions. An environment is necessary where any service provider can create applications and distribute them to users. In [7] proposed a protocol in which, a broadcaster distributes a signing key to a service provider that the broadcaster trusts. As a result, users can verify that an application is reliable. Another application of the KIS scheme is bidirectional broadcasting service [8, 9] in which a signer communicates with a huge number of receivers. In which, to renew the verification key, the signer has to send his new verification key to all receivers in an authentic manner. An efficient strong key-insulated signature scheme is proposed in [10]. The scheme is more efficient and is secure under the discrete logarithm assumption in the random oracle model. Traditional identity-based signature schemes typically rely on the assumption that secret keys are kept perfectly secure. But more and more cryptographic primitives are deployed on insecure devices such as mobile devices, key-exposure seems inevitable. It does not matter how strong the scheme is, once the secret key is exposed, it will be broken completely. To deal with this problem in [11, 12] use the key-insulation mechanism to minimize the damage of key-exposure in IBS schemes. Certificate-based key-insulated signature scheme [13] is proven to be existentially unforgeable against adaptive chosen message attacks in the random oracle model. Identity (ID)-based key-insulated cryptography has received much attention from cryptographic researchers. Reference [14] is a new and efficient ID-based key-insulated signature scheme with batch verifications.

Mathematical Primitives

In this section, we discuss the elliptic curve and its properties. We then discuss the rules for adding points on elliptic curve and the elliptic curve discrete logarithm problem.

Elliptic Curve over Finite Field

Let a and $b \in Z_p$, where $Z_p = \{0, 1, \ldots, p - 1\}$ and $p > 3$ be a prime, such that $4a^3 + 27b^2 \neq 0(\text{mod} p)$. A non-singular elliptic curve $y^2 = x^3 + ax + b$ over the finite field $GF(p)$ is the set $E_p(a, b)$ of solutions $(x, y) \in Z_p \times Z_p$ to the congruence $y^2 = x^3 + ax + b(\text{mod} p)$, where a and $b \in Z_p$ are constants such that $4a^3 + 27b^2$ / $= 0(\text{mod} p)$, together with a special point O called the point at infinity or zero point.

The condition $4a^3 + 27b^2 \neq 0(\text{mod} p)$ is the necessary and sufficient to ensure that the equation $x^3 + ax + b = 0$ has a non-singular solution [15]. If $4a^3 + 27b^2 = 0(\text{mod} p)$, then the corresponding elliptic curve is called a singular elliptic curve. If $P = (x_P, y_P)$ and $Q = (x_Q, y_Q)$ be points in $E_p(a, b)$, then $P + Q = \text{O}$ implies that $x_Q = x_P$ and $y_Q = -y_P$. Also, $P + \text{O} = \text{O} + P = P$, for all $P \in E_p(a, b)$. Moreover, an elliptic curve $E_p(a, b)$ over Z_p has roughly p points on it. More precisely, a well-known theorem due to Hasse asserts that the number of points on $E_p(a, b)$, which is denoted by $\#E$, satisfies the following inequality [16]: $p + 1 - 2\sqrt{p} \leq \#E \leq p + 1 + 2\sqrt{p}$. In addition, $E_p(a, b)$ forms an abelian group or commutative group under modulo p operation.

Addition of Points on Elliptic Curve over Finite Field

The following parameters about the proposed proxy signature scheme over the elliptic curve domain are required. We take an elliptic curve over a finite filed $GF(p)$ as $E_P(a, b) : y^2 = x^3 + ax + b(\text{mod} p)$, where a and $b \in GF(p)$. The field size p is considered as a large prime. We take G as the base point on $E_p(a, b)$ whose order is n, that is, $nG = G + G + \cdots + G(ntimes) = \text{O}(\text{mod} p)$.

The elliptic curve addition differs from the general addition [17]. Let $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ be two points on elliptic curve $y^2 = x^3 + ax + b(\text{mod} p)$, with $P \neq -Q$, then $R = (x_3, y_3) = P + Q$ is computed as follows: $x_3 = (\lambda^2 - x_1 - x_2)(\text{mod} p)$,

$$y_3 = (\lambda(x_1 - x_3) - y_1)(\text{mod} p);\ \text{where}\ \lambda = \begin{cases} \dfrac{y_2 - y_1}{x_2 - x1}(\text{mod} p),\ \textit{if}\ P \neq Q \\ \dfrac{3x_1^2 + a}{2y_1}(\text{mod} p),\ \textit{if}\ P = Q \end{cases}$$

In elliptic curve cryptography, multiplication is defined as repeated additions. i.e., if $P \in E_p(a, b)$, then $4P$ is computed as: $4P = P + P + P + P(\text{mod} p)$

Discrete Logarithm Problem

The discrete logarithm problem (DLP) is as follows: given an element g in a finite group G whose order is n, that is, $n = |G|$ and another element $h \in G$, find an integer x such that $g^x = h(\text{mod} n)$. It is relatively easy to calculate discrete exponentiation $g^x(\text{mod} n)$ given g, x and n, but it is computationally infeasible to determine x given h, g and n, when n is large.

Elliptic Curve Discrete Logarithm Problem

Let $E_p(a, b)$ be an elliptic curve modulo a prime p. Given two points $P \in E_p(a, b)$ and $Q = k \cdot P \in E_p(a, b)$, for some positive integer k. $Q = k \cdot P$ represents the point P on elliptic curve $E_p(a, b)$ is added to itself k times. The elliptic curve discrete logarithm problem (ECDLP) is to determine k given P and Q. It is relatively easy to calculate Q given k and P, but it is computationally infeasible to determine k given Q and P, when the prime p is very large.

Basic Model of KIS Scheme

In this section, we will discuss the basic model of KIS scheme [1]. This KIS scheme consists of five polynomial time algorithms $(Gen, Upd^*, Upd, Sign, Vrfy)$.

Gen: Key generation algorithm is a probabilistic algorithm that takes as input a security parameter 1^k and the total number of time periods N. It returns a master key SK^*, a verification key VK, and an initial signing key SK_0.

Upd^* :Partial key generation algorithm is a probabilistic algorithm that takes as input indices i and j for time periods and a master key SK^*. It returns a partial key $SK'_{i,j}$

Udp :Key update algorithm is a deterministic algorithm that takes as input indices i, j, a signing key SK_i, and a partial key $SK'_{i,j}$. It returns a signing key SK_j for time period j

$Sign$:Signing algorithm is a probabilistic algorithm that takes as input an index i for a time period, a message M, and a signing key SK_i. It returns a pair $\langle i, s \rangle$ consisting of a time period i and a signature s

$Vrfy$:Verification algorithm is a deterministic algorithm that takes as input a verification key VK, a message M, and a pair $\langle i, s \rangle$. It returns a bit b. If b is $TRUE$, then accept the signature, otherwise reject

KIS Scheme for Real-Time System

In this section, we describe the different phases of our proposed Key-Insulated Signature Scheme.

Setup Phase

In this section, we generate a large prime p. We take an elliptic curve over the finite field $GF(p)$ as $E_P(a, b) : y^2 = x^3 + ax + b(\text{mod}p)$ such that the elliptic curve discrete logarithm problem (ECDLP) becomes intractable. We take G as the base point on $E_p(a, b)$ whose order is q. Then randomly generate a private key d in the range $0 < d < q$ and computes the public key Q, as $Q = d \cdot G$. $h(\cdot)$ represents a secure one-way hash function (for example, SHA-1). In summary, we list the generated domain parameters in Table 1.

Key Generation Algorithm

Randomly select d' in the range $0 < d' < q$ which is kept by the signer. Compute the master key d_0 as $d_0 = d - d'$, stored in the secure device. Calculate $Q_0 = d_0 \cdot G$ and $Q' = d' \cdot G$.

Set the verification key as $VK : < q, G, Q, Q_0, Q', h_1, h_2 >$

Partial Key Generation Algorithm

From the secure device randomly select k_A, $0 < k_A < q$ and compute $R_A = k_A \cdot G$. Set $(x_A, y_A) = R_A$ and compute $r_1 = x_A(\text{mod}q)$. Calculate $v_1 = h_1(r_1, T)$ using time period T. Compute the partial key $s_1 = v_1 \cdot d + r_1 \cdot d_0(\text{mod}q)$.

Send (s_1, r_1, T) to the signer.

Table 1 Parameters generated in setup phase

m	Message to be signed
p	A large prime number
$E_p(a, b)$	Elliptic curve over $GF(p)$
G	A base point on $E_p(a, b)$
q	Order of G
$h_1(\cdot, \cdot)$, $h_2(\cdot, \cdot, \cdot, \cdot)$	Secure one-way hash functions
d	A private key, $0 < d < q$
Q	A public key, $Q = d \cdot G(\text{mod}q)$

Key Update Algorithm

The signer calculates $v_1 = h_1(r_1, T)$ and verify the partial key as follows:

$$s_1 \cdot G = v_1 \cdot G + r_1 \cdot Q_0$$

if it is successful, then the signer compute the signing key for the time period T as:

$$sk_T = (s_1 + d')(\mathrm{mod}q).$$

Signing Algorithm

Now the signer randomly select k_S, $0 < k_S < q$ and compute $R_S = k_S \cdot G$. Set $(x_S, y_S) = R_S$ and compute $r_S = x_S(\mathrm{mod}q)$. Calculate $v_S = h_2(r_1, r_S, T, m)$ by using the input massage m and time period T. Then compute $\delta_S = (v_S \cdot k_S - sk_T)(\mathrm{mod}q)$. Finally the signer transmits the signature tuple with the message $\langle m, (\delta_S, v_S, r_1), T \rangle$ to the verifier for verification.

Verification Algorithm

Using $Q, Q_0, Q', m, (\delta_S, v_S, r_1)$ and T the verifier compute $v_1 = h_1(r_1, T)$ and $W = v_S^{-1}(\mathrm{mod}q)$. Compute $X = (x'_S, y'_S) = (G \cdot \delta_S + v_1 \cdot Q + r_1 \cdot Q_0 + Q')W$ if $X == 0$, then reject the signature. Otherwise compute $r'_S = x'_S(\mathrm{mod}q)$.

Accept the signature if and only if $r'_S == r_S$.

Analysis of the New KIS Scheme

Correctness Proof

Proof 1 We show that the equation $s_1 \cdot G = v_1 \cdot G + r_1 \cdot Q_0$ holds good in order to check the validity of the partial key as follows:

$$L.H.S = s_1 \cdot G = (v_1 \cdot d + r_1 \cdot d_0)G$$
$$= v_1 \cdot d \cdot G + r_1 \cdot d_0 \cdot G = v_1 \cdot Q + r_1 \cdot Q_0 = R.H.S$$

Proof 2 We have the equation $\delta_S = v_S \cdot k_S - sk_T$ to check whether the signature is valid or not. Multiplying G on both side of the equation, we will get:

$$\begin{aligned} G \cdot \delta_S = G \cdot (v_S \cdot k_S - sk_T) = v_S \cdot k_S \cdot G - sk_T \cdot G \\ = v_S \cdot k_S \cdot G - (s_1 + d') \cdot G = v_S \cdot k_S \cdot G - (s_1 \cdot G + d' \cdot G) \\ = v_S \cdot k_S \cdot G - (v_1 \cdot Q + r_1 \cdot Q_0 + Q') \end{aligned}$$

Rearranging the equation we will gate:

$$k_S \cdot G = (G \cdot \delta_S + v_1 \cdot Q + r_1 \cdot Q_0 + Q') \cdot v_S^{-1}$$

i.e. $(x'_S, y'_S) = (G \cdot \delta_S + v_1 \cdot Q + r_1 \cdot Q_0 + Q')W$

As a result, we have $x_S == x'_S$, that is, $x_S(\text{mod}q) == x'_S(\text{mod}q)$ and hence, $r_S == r'_S$.

Performance Analysis

To achieve reasonable security, RSA and DSA should employ 1024-bit moduli, while a 160-bit modulus should be sufficient for ECC [4]. Moreover, the security gap between the systems increases dramatically as the moduli sizes increases. For example, 300-bit ECC is dramatically more secure than 2,048-bit RSA or DSA [18]. Since we constructed our scheme over elliptic curves, so we can get better security with small key size. Due to small signing and verification key, the signature size is reduced and the computational cost for signing and verification is efficient.

For analysis of computational costs required for different phases in our scheme, we use the following notations shown in Table 2.

It is clear to note that the computational costs of different phases of KIS Scheme are as follows: Key Generation Algorithm: $T_{OA} + 2T_{ECM}$;

Partial Key Generation Algorithm: $T_{OA} + 2T_{OM} + T_{ECM} + 2T_{\text{mod}} + T_H$;

Key Update Algorithm: $2T_{OA} + 2T_{ECM} + T_H + T_{\text{mod}}$;

Signing Algorithm: $T_{OA} + T_{ECM} + T_{OM} + T_H + 2T_{\text{mod}}$;

Verification Algorithm: $3T_{OA} + 4T_{ECM} + T_{INV} + T_H + T_{\text{mod}}$

Table 2 Notations used in computation of computational cost

T_E	Time taken for an exponential operation
T_H	Time taken for a one-way hash function $h(\cdot)$
T_{INV}	Time taken for a modular inverse operation
T_{EMC}	Time taken for a scalar (elliptic curve) multiplication
T_{OM}	Time taken for an ordinary multiplication
T_{mod}	Time taken for a modular operation
T_{OA}	Time taken for an ordinary addition

Security Analysis

In the standard framework of identification protocols the signer sends an initial message, the verifier sends a random challenge and the signer respond with some answer. Here we have a master key which is stored on a physically-secure device. At the beginning of time period T, the signer interacts with the secure device in order to obtain a key sk_T valid for current time period. The signer (who may be operating on an insecure device) proves his identity to a verifier in period T using key sk_T. Let us consider an adversary, who interacts with the signer in an execution of the identification protocol during various time periods, and may additionally compromise the insecure device and obtain the temporary keys. As the validity of the key is for limited time period, so the adversary will not be able to successfully impersonate the signer during any time period other than those in which it compromised.

The security of a KIS scheme means a KIS scheme is secure against a signing key leakage. i.e. the signing key sk_T was leaked by a signer and used by a third party(adversary). Still the KIS scheme is considered secure as the adversary cannot generate valid signature using the leaked signing key sk_T. Since signing key sk_T is valid for that T period of time in which the compromise occurred.

The security of a Strong KIS scheme means a KIS scheme is secure against a signing key or master key leakage. If signing key or a master key was leaked by a signer and used by a third party, still the adversary cannot generate valid signature, because the master key d_0 and d' are managed by two different servers. Even if the master key d_0 is leaked, the signing key sk_T cannot be updated without d'.

Conclusions

In this paper, we have considered different aspect of real-time communication system and proposed a secure and reliable authentication mechanism. We have designed a key-Insulated Signature scheme that minimized the damage caused by key-exposure. The main feature of this algorithm is it is based on elliptic curve cryptography. This reduces the key size as well as the signature size and increase the security compared to DLP and RSA as the security is based on elliptic curve discrete logarithmic problem. We have computed the computational cost; in addition performance of the scheme is analyzed.

Acknowledgments The paper is supported partially by the "Cross-platform smart phone-enable stroke prevention and rehabilitation health system" project and partially by the NSC-101-2811-E-008-013 project.

References

1. Dodis Y, Katz J, Xu S, Yung M (2003) Strong key-insulated signature scheme. In: Public Key Cryptography PCK, pp 130–144
2. Kopetz H (2011) Real-time system: design principles for distributed embedded applications, vol 25. Springer
3. Lin L, Lin P (2007) Orchestration in web services and real-time communications. Commun Mag, IEEE 45(7):44–50
4. Barker E, Barker W, Burr W, Polk W, Smid M (2011) Recommendation for key-management-part 1: general (revision 3). NIST Spec Publ 800:57
5. Dodis Y, Katz J, Xu S, Yung M (2002) Key-insulated public key cryptosystems. In: Advances in cryptology Eurocrypt, Springer pp 65–82
6. Gonzalez-Deleito N, Markoitch O, DallOlio E (2004) A new key-insulated signature scheme. In: Information and communication security, pp 9–12
7. Ohtake G, Ogawa K (2012) Application authentication for hybrid services of broadcasting and communication networks. In: Information security applications, pp 171–186
8. Ohtake G, Hanaoka G, Ogawa K (2010) Efficient provider authentication for bidirectional broadcasting services. IEICE Trans Fundam Electron, Commun Comput Sci 93(6):1039–1051
9. Matsuda T, Hanaoka G, Ogawa K (2007) A practical provider authentication system for bidirectional broadcast service. In: Knowledge based intelligent information and engineering systems. Springer, pp 967–974
10. Ohtake G, Hanaoka G, Ogawa K (2008) An efficient strong key-insulated signature scheme and its application. In: Public key infrastructure, PKI, pp 150–165
11. Weng J, Liu S, Chen K, Li X (2006) Identity based key-insulated signature with secure key-updates. In: Information security and cryptology. Springer, pp 13–26
12. Weng J, Liu S, Chen K, Ma C (2007) Identity based key-insulated signature without random oracles. In: Computational intelligence and security, pp 470–480
13. Du H, Li J, Zhang Y, Li T, Zhang Y (2012) Certificate based key-insulated signature. In: Data and knowledge engineering. Springer, pp 206–220
14. Wu TY, Tseng YM, Yu CW (2012) Id based key-insulated signature scheme with batch verifications and its novel applications. In: Int J Innovative Comput, Inf Control 8(7(A)):4797–4810
15. Nickalls R (1993) A new approach to solving the cubic: Cardan's solution revealed. In: The mathematical Gazette. pp 354–359
16. Stallings W (2003) Cryptography and network security, principles and practices. Practice Hall
17. Koblitz N (1987) Elliptic curve cryptosystem. Math Comput 48(177):203–209
18. Remarks on the security of elliptic curve cryptosystem: a Certicom white paper, CTEC Cryptosystem, updated July (2000). http://www.certicom.com/

Innovative Wireless Dedicated Network for e-Bus

Eric Hsiao-Kuang Wu, Chung-Yu Chen, Ming-Hui Jin and Shu-Hui Lin

Abstract In recent years, several kinds of wireless network have been widely deployed. In this work, we focus on the e-bus-system for movements of buses. Passengers can get the bus information through the intelligent bus stop when they are waiting for the bus. In today's environment of Taiwan, the transportation information is provided through some specific Mobile network operators. In this case, it costs a huge expenditure every month for the charge of data transmission. Therefore, we propose a new network system with Digital Mobile Radio (DMR) and Super Wi-Fi to replace the mobile network operator. Furthermore, when the catastrophe happened, the proposed system can also change to be used for the rescues. The implementation of the system architecture, the application of DMR and super WiFi in the proposed system, and the operation scenario are described in this paper.

Keywords Digital mobile radio · Intelligent transportation systems

E. H.-K. Wu (✉) · C.-Y. Chen
Department of Computer Science and Information Engineering, National Central University,
Jhongli, Taiwan
e-mail: hsiao@csie.ncu.edu.tw

C.-Y. Chen
e-mail: giawoba@wmlab.csie.ncu.edu.tw

M.-H. Jin · S.-H. Lin
Department of Smart Network System Institute, SNSI Institute for Information Industry,
No. 133, Minsheng East Road, Taipei 10574, Taiwan
e-mail: Jinmh@iii.org.tw

S.-H. Lin
e-mail: vickylin@iii.org.tw

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_141, © Springer Science+Business Media Dordrecht 2014

Introduction

In the recent years, many public and private organizations, such as police forces, fire brigades and health emergency associations [1], have built their own trunking or simulcast digital radio solutions to replace the old analog radio networks. In this paper we focus on the application of communication techniques in e-bus-system. If the bus stop provides the bus information, passengers can grasp it immediately, and can appropriately adjust the schedule. In Taiwan, in order to enhance the service performance of the bus transport system, Ministry of transportation and communications (MOTC) has developed Intelligent Transportation Systems (ITS) [2] plan. The combination of electronics, communications, information and the management of transportation technology have strengthened the supervision and management of the bus. Thereby, the proposed plan is to attract people to take the bus and, therefore, to increase the usage of public transportation. Its main objectives are to save the energy consumed by the transportation and to avoid the traffic congestion. The e-bus-system in Taiwan is used to transmitting the information through some specific Mobile networks. In this case, it needs to pay a lot of monthly expenditure for data transmission. And it always introduces burden to the government and the bus companies. In paper [3], it compared some common professional mobile radio (PMR) standards. The result shows that Digital Mobile Radio (DMR) [4] has been identified as the best solution, which grants cost saving, high coverage, spectral efficiency and simplicity in network configuration and is well suitable in wide area with a low/medium density of traffic. And this characteristic is suitable to be applied for the e-bus-system.

In this paper, we propose a new network system with DMR and Super Wi-Fi to replace the current system, which is supported by the mobile network operator. The proposed system will use for the e-bus-system as usual, however, it can also be used for emergency communication in rescue mission when the catastrophe happens. Thus the proposed system provides a compensative solution for the public network that is usually unable to communicate when the occurrence of disaster.

The organization of this paper is as follows. The architecture of the proposed system is provided in the following section. Experiments presents the implementation and experiments of the proposed system. The experiment is given in Discussions. And, finally, we provide some conclusions in the last section.

Innovative e-Bus System

System Architecture

Our e-bus-system is divided into three parts, as shown in Fig. 1. First, we use the DMR technique to implement the communication links between BUS and the central control server. We will set the DMR based GPS device in the bus

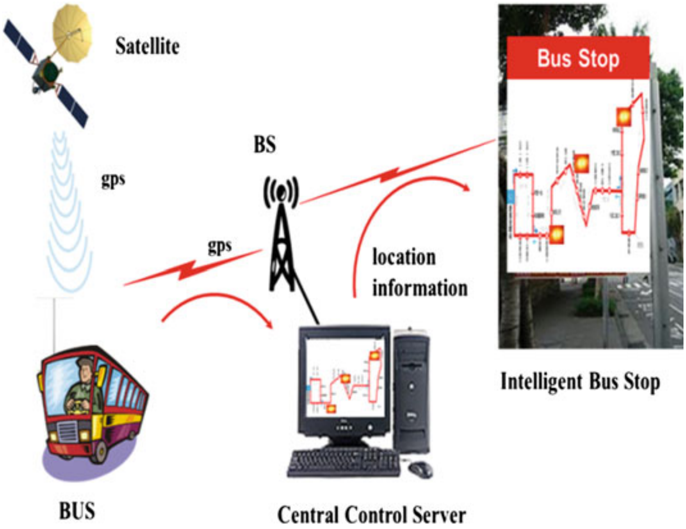


Fig. 1 System architecture

afterwards. It will transmit the GPS information back to the server side with the fixed time intervals, and the format is following the [5]. Second, we build the central control server to record and process the GPS information. We transformed the GPS information into simply bit string, and then transmit to the bus stop. The bus stop will show up the LED to let public know where is the bus. The bus stop we designed is the dynamo bus stop, so that it can't handle too complexity information. Third, we use super Wi-Fi to send data to the bus stop. The reason why we use super Wi-Fi will be explain later. Finally, our system replace the network system which provided by the network operator.

Digital Mobile Radio

DMR is one of Digital PMR solution proposed by the European Telecommunications Standards Institute (ETSI) and it is diffused digital technologies in Europe. This technology was developed by ETSI to grant gradual migration from the analogical conventional system to the digital mode without new licenses and without changing the existing network architecture. DMR uses 2 time slots on a 12.5 kHz bandwidth carrier and adopts time division multiplexed accessing (TDMA) with 4-FSK modulation. Since the modulated signal has constant envelope, a transmitter can work in saturation (clipping) mode (C class or superior) with very low consumption. It is noted that DMR has the maximum bit rate of 9.6 kbps and can works in simulcast mode to provide a wider coverage area (until 80 km), using a frequency pair only. Network and terminals can be dual mode,

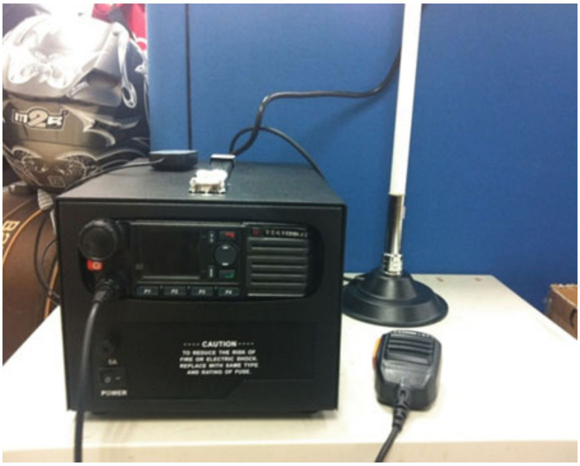


Fig. 2 DMR communication device

thus granting the coexistence of analog and digital devices. It provides for voice and data, in which tier 1 is direct mode, tier 2 is conventional mode, and tier 3 is trunking mode. Figure 2 shows the example of DMR communication equipment. It will be set in the bus for experiment, and the work is to send the GPS information back to the central control side periodically. Every device has the unique ID so that we can identify the bus. Furthermore, each device is equipped with a digital two-way radio. Both bus driver and dispatcher can communication with each other to deal with emergencies or temporary dispatching immediately. It enhances the ability of being adaptable.

Super Wi-Fi

The proposed system adopts the super Wi-Fi as the communication technique. The super Wi-Fi, or IEEE 802.22 [6] and IEEE 802.11af as it is technically known, is a term coined by the United States Federal Communications Commission (FCC) to describe a wireless networking proposal. FCC plans to apply it for the creation of longer-distance wireless internet. The use of the trademark "Wi-Fi" in the name has been criticized by people because it is not based on common known traditional Wi-Fi technology or endorsed by the Wi-Fi Alliance. Instead of using the 2.4 GHz radio frequency of Wi-Fi, the 'Super Wi-Fi' proposal uses the lower-frequency white spaces between television channel frequencies. These lower frequencies allow the signal to travel further and penetrate walls better than the higher frequencies previously used. The FCC's plan is to allow those white space frequencies to be used for free, as happens with Wi-Fi and Bluetooth. Here we have noticed that the data is send from the base station to all the bus stops. It can be expected that the data flow will be much bigger than the traffic between bus and central control side, and DMR may not deal with it. In our design, we use super Wi-Fi to handle this part to send the corresponding bus position to the bus stop. Compared to DMR, super Wi-Fi has farther transmission range and higher transmission rates, and the transmission rate is the main reason why we use the super Wi-Fi.

Central Control Server

In the central control server, its main functions are to process the received location information from buses by referring to the geographical information and bus routes information base. Here we need to associate the received GPS information to the position of the bus stop, and expressed as a bit string of the bus routes. The software process of the proposed system in the central control server is illustrated in Fig. 3. The bus will send its bus ID with location information, which includes longitude and latitude, to the central control server periodically. As shown in Fig. 3, when received the location information sent by the bus, the central control server associates the longitude and latitude with the geographic information base to determine its location in the map. Then the server determines the route of the reporting bus by referring to its bus ID. It is noted that the bus route could be helpful for the location identification of the bus. Finally, the central control server sends the bus location to the bus stops, which are associated with this bus route, and transformed it into web page for query.

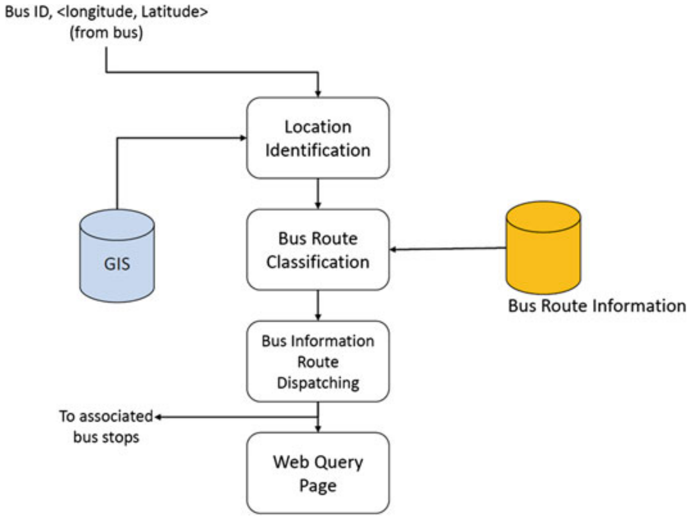


Fig. 3 Information process of central control server

One bus stop is corresponding to one bit, in which the bit set to one represents the bus stop that the bus is going to arrive. When the bus stop receives the bit string, it will light up the LED to show the bus current position. Here we have set up the web site to show the records and bus information, furthermore we also build our own GIS (Geographic Information System) [7] to show the bus currently position on the map. The record and the map will be dynamic update periodically, so that public can also grasp the latest BUS information on the web site. Combine these three parts; our e-bus-system has been completed.

Experiments

In this section, an experiment is performed by using our system to test in the real environment. We performed our experiment in the Taipei City. Here we set the DMR equipment in our personal vehicle, and then we follow the bus to imitate the real situation. And we took the travel of Bus 518 as the experimental example. That is Bus 518 equipped with the DMR communication device to periodically report its location information during the experiment.

In Fig. 4, the lines represent the road, and triangle represents the current location of the bus. It is noted that the map used in Fig. 4 can be constructed in layer approach. Thus we can selectively define which kind of information to be visible on the map. This function is very useful for user interface. For example, as there are many bus routes in the city, we can choose to show some specific routes only. It is noted that the relationship between the bus stops and the current bus location or the bus arrival time prediction [8] is more meaningful for the passengers. Thus the determined location information of the bus shall be mapped to longitude and latitude of the nearest bus stop. Figure 5 shows the bus position being related to the bus stops associated with the route of this bus, and arrived time during the experiment is also provided. It is noted that, in addition to the route selection, the system also provides the capabilities of car tracing, map view, history information, and the transformation of GPS. Thus users can access versatile information from the system.

Discussions

In the past, traditional analog radio system offers some great benefits, such as lower cost of ownership, to meet the requirements of users of radio communication range and voice communications, simple and reliable operation capability. However, analog radio technology has reached its cap because the radio spectrum interval is the main limitation to achieve the accurate and reliable communications. Thus spectral efficiency and channels cannot meet the requirements of digital communications due to the problems of congestion and interference.

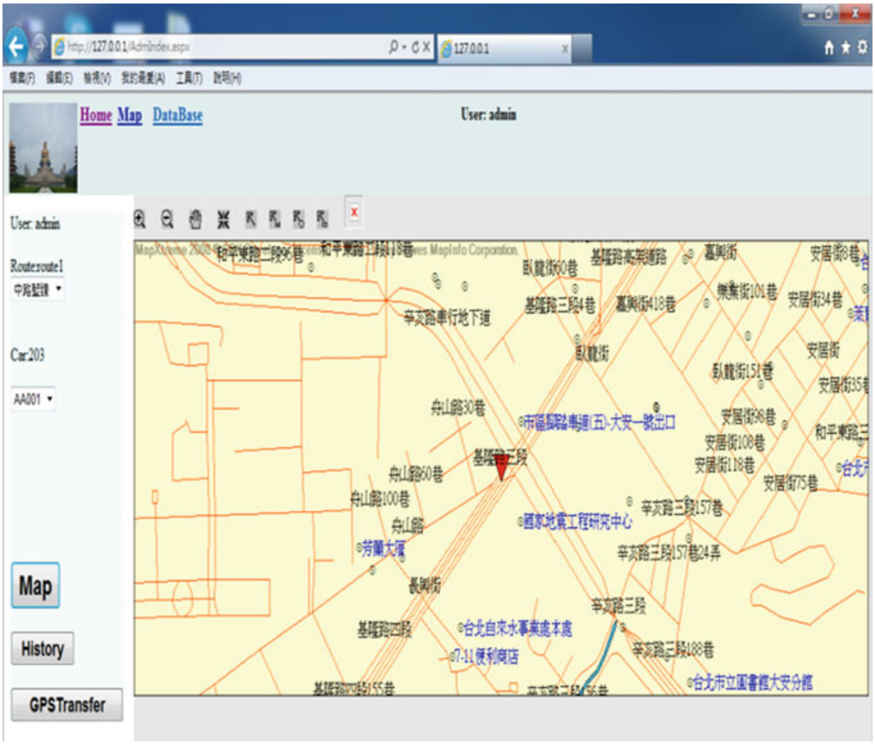
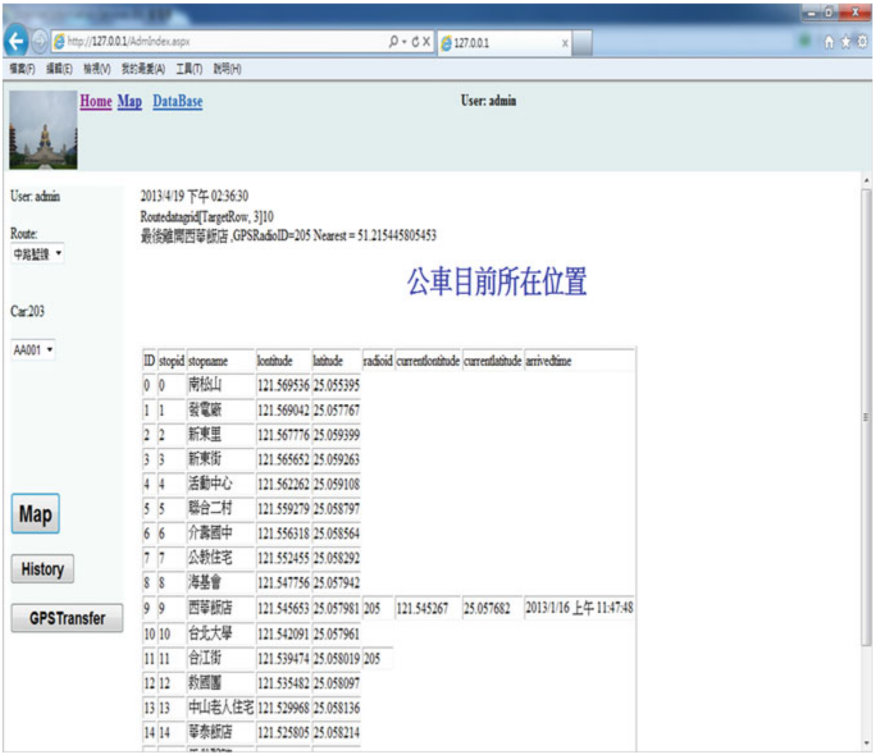


Fig. 4 Geographic information of the experiment (*Filled triangle* the bus location)

Our system can divide into three parts. First part is the connection between bus and the central control server. In this part the DMR communication device is use for sending the GPS information back to the central control side. However, our system has only experiments with one bus currently, so when the bus increased, system capacity is need to be concerned. Second part is the central control server. This part processes the GPS data. It correspond GPS information to the position of the bus stop. When the bus stop receives the information, it will lights up the LED to show the bus current position. Final, the third part is the connection between central control server and the bus stop. Our system use Super Wi-Fi to send the data. Because we have noticed that the data is send from the base station to all the bus stop, Can be expected that the data flow will much bigger than the traffic between bus and central control side, and DMR may not deal with it.

In our experiment, the proposed e-bus-system demonstrates that it can provide accurate information for query. However, after analysis the GPS records we have noticed that the GPS single is unstable in real environment. Because the bus may pass through the tunnel or the viaduct, GPS device may be unable to receive the signal from satellites. In such locations, the system will not receive the information sent by the bus. We also notice the GPS accuracy has a little deviation, these



2013/4/19 下午 02:36:30
Routedatagrid[TargetRow, 3]10
最後離開西華飯店 ,GPSRadioID=205 Nearest = 51.215445805453

公車目前所在位置

ID	stopid	stopname	lontitude	latitude	radioid	currentlontitude	currentlatitude	arrivedtime
0	0	南松山	121.569536	25.055395				
1	1	發電廠	121.569042	25.057767				
2	2	新東里	121.567776	25.059399				
3	3	新東街	121.565652	25.059263				
4	4	活動中心	121.562262	25.059108				
5	5	聯合二村	121.559279	25.058797				
6	6	介壽國中	121.556318	25.058564				
7	7	公教住宅	121.552455	25.058292				
8	8	海基會	121.547756	25.057942				
9	9	西華飯店	121.545653	25.057981	205	121.545267	25.057682	2013/1/16 上午 11:47:48
10	10	台北大學	121.542091	25.057961				
11	11	合江街	121.539474	25.058019	205			
12	12	救國團	121.535482	25.058097				
13	13	中山老人住宅	121.529968	25.058136				
14	14	華泰飯店	121.525805	25.058214				

Fig. 5 Bus position related to the bus stops and times during the experiment

problems are necessary to be concerned. Furthermore, in our experiments, the GPS information is submit every twenty seconds. However, we noticed that during these twenty seconds, the bus may drive through two or more bus stop. If we increase the frequency of the bus information, it will be more accurate. But the server will face the big data issues. How to balance the tradeoff is also a big challenge.

In the geographic information system, we have built our own map. The reason why we didn't use the Google map is that we have to get the fully control to add additional layer into the map, such as the bus stop information, bus route information, etc. This is not possible to be implemented with Google map. In addition, as we can fully obtain and control the geographic information of the map and assign the bus stops at specific locations, it provides the capability to tune accuracy of the location identification especially when the bus information is missed. Furthermore, based on the current architecture, we can develop more attractive applications in the future.

Our system will use for the e-bus-system as usual, however, when the catastrophe happened, DMR can also use for the communication in rescue mission, seize the prime time of rescue. It solve the problem that public network is usually unable to communicate when disaster happened. Some similar system such as P25 [9], TETRA [10] are already used widely in police forces, fire brigades in other countries. In this paper, we demonstrate the feasibility of the proposed system for e-bus service from the real experiment. And we need to develop some rescue-based scenarios to verify the usability and to measure the system performance.

Conclusions

Due to the environment in Taiwan, most of the network resources (i.e. spectrum) have been allocated for some specific usages such as private mobile networks, digital TV, radio broadcasting, etc. It is necessary to carefully reuse the existing spectrum for the development of some valuable information distribution systems. In this paper, we implement a new network system with DMR and Super Wi-Fi to be used as the communication in e-bus-system to replace the current system that utilizes the spectrum band of mobile networks. Currently, the DMR is used the transmission from bus to the control server and the super Wi-Fi is adopted for transmission from the control server to the bus stops. In order to simplify the communication architecture, the simpler architecture may be possible after further examination on the performance and interference of both communication spectrums.

In the proposed system, passengers can easily get the bus information through the intelligent bus stop or on our web site immediately. However, as mentioned in previous section, we have learned that the GPS single is not stable in real environment. Fortunately, we can access the map information directly. Therefore, we plan to develop some location refinement techniques according to the speed of the bus and geographic information to solve this problem in the near future. Our work is just at the beginning toward the wide deployment. In addition to the location refinement, we are expecting to work on the digital surveillance system with the intersection monitors. There still a lot of applications worth to be developed.

Acknowledgments The paper is supported partially by the “Cross-platform smart phone-enable stroke prevention and rehabilitation health system” project and partially by the NSC-101-2811-E-008-013 project.

References

1. Pawelczak P, Prasad RV, Xia L, Niemegeers IG (2005) Cognitive radio emergency networks—requirements and design. In: Proceedings of IEEE international symposium dynamic spectrum access networks (DySPAN), pp 601–606
2. Weiland RJ, Purser LB (2000) Intelligent transportation systems. Transportation in the new millennium
3. Onali T, Sole M, Giusto DD (2011) DMR networks for health emergency management: a case study. In: 7th international wireless communications and mobile computing conference (IWCMC)

4. ETSI, TS-102-361, electromagnetic compatibility and, radio spectrum matters (ERM): digital mobile radio, (DMR) systems (2007)
5. Langley RB (1995) NMEA 0183: a GPS receiver interface standard. GPS world 6.7, pp 54–57
6. Stevenson C, Chouinard G, Lei Z, Hu W, Shellhammer S, Caldwell W (2009) IEEE 802.22: the first cognitive radio wireless regional area network standard. IEEE Commun Mag 47:130–138
7. Burrough PA et al (1998) Principles of geographical information systems, vol 333. Oxford university press, Oxford
8. Lin W-H, Zeng J (1999) Experimental study of real-time bus arrival time prediction with GPS data. Trans Res Rec 1666:101–109
9. TIA, TIA-102: project P.25 (2004)
10. ETSI, EN-300-392, terrestrial trunked radio (TETRA) (2009)

Cross-Platform Mobile Personal Health Assistant APP Development for Health Check

Eric Hsiao-Kuang Wu, S. S. Yen, W. T. Hsiao, C. H. Tsai, Y. J. Chen, W. C. Lee and Yu-Wei Chen

Abstract Our team proposes a concept allowing patients taking health check anytime everywhere, which can increase patients' attention to their own health condition. To improve the user experience and convenience, the system must be designed to simply operate and easily connect with the medical devices. Moreover, the system must have the ability to communicate between the patients and doctor or medical personnel. In this paper, we illustrate our system, such as user interface, storage, display and cloud system. The user interface is designed with standards-based Web technologies. We use PhoneGap to build cross-platform mobile apps with HTML, JavaScript, and CSS. Because patients need to keep their record of health check, we use SQLite database for storage. Moreover, for the patients'

E. H.-K. Wu (✉) · S. S. Yen · W. T. Hsiao · C. H. Tsai · Y. J. Chen · W. C. Lee
Department of Computer Science and Information Engineering, National Central University,
Jhongli, Taiwan
e-mail: hsiao@csie.ncu.edu.tw

S. S. Yen
e-mail: freehourse2sh@gmail.com

W. T. Hsiao
e-mail: wayne12345.168@gmail.com

C. H. Tsai
e-mail: a968574123@hotmail.com

Y. J. Chen
e-mail: denny923@livemail.tw

W. C. Lee
e-mail: win60615@gmail.com

Y.-W. Chen
Department of Neurology, Landseed Hospital, No. 77, Kwang-Tai Road,
Ping-Jen City 32449 Tao-Yuan County, Taiwan
e-mail: yuwchen@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_142, © Springer Science+Business Media Dordrecht 2014

health check report which shall be easily understood, we design line charts to display the data. In this paper, we implement the Personal Health Assistant system.

Keywords Health check · PhoneGap · Cross-platform

Introduction

Nowadays, smartphones are in widespread use and enough to replace the feature phones [1]. With the smartphones growth, the applications are explosive increasing. The field of smartphones applications appeals lots of developers, because the operating systems of smartphones allow users to install unofficial applications on the OS, moreover providing the application market for developers to sell their own applications. There are several operating systems, Google Android, Apple iOS, Windows Phone which have its own application store, Google Play Store, Apple App Store, Windows Phone Apps store respectively.

The Population aged 65 and above has increased notably over time. According to Taiwan Ministry of the Interior, in the end of 2012 the Population aged 65 and above has occupied 11.15 % of Taiwan population [2]. Taiwan has become aging society. Furthermore, the elder need more care, especially in health care. Therefore health check becomes increasingly important. On the other hand, people in Taiwan are suffering from chronic disease for a long time. Hypertension, hyperlipidemia, diabetes, cancer, for example, are some of the main chronic diseases. In the 2011 the Statistics of Causes of Death show that chronic diseases occupy up to 75 % of total causes of death [3]. Besides, the percent of elder having chronic diseases is higher than younger. It infers health care for elder is considerably important. To prevent chronic diseases or as therapy, regular health check is a useful method. Moreover, regular health check is not only for this way. There are lots of benefits of regular health check, such as detection of disease symptom, earlier treatment, and so on.

Android is the smartphone market leader [4]. To integrate our medical devices with smartphone applications. We have developed the health check application for Android. Nonetheless, it is not enough, there are other smartphone operating systems, iOS, Windows Phone for example. In order to fulfill the usage form other operating systems, we have to develop other versions for them. Therefore we use PhoneGap as the tool to develop the cross-platform application with Web Language, HTML, JavaScript, and CSS. We do not have to know the programming language for android, iOS, and Windows Phone, yet we can fulfill the application for all operation systems. The advantage is decreasing the development cost, and simpler management.

In order to storing the data of our application, we use database to achieve our goal. PhoneGap storage API is based on the W3C Web SQL Database Specification and W3C Web Storage API Specification. Some devices already provide an implementation of this specification. This is why we using SQLite to implement our application [5].

We can view the health check records in the table by loading data from SQLite database, but sometimes the data is considerable that are impossible to view specifically. Instead, line chart is better to show the overview. The line chart in our app is written in JavaScript and uses the new canvas element to load graph data from SQLite database.

The mobile devices and the networks are inseparable. We cannot leave our application working on its own, because we should make differences between other applications. What we should do is that the devices can always exchange data with a remote server where can store data permanently. In this way, we integrate distributed data of the same person in a remote server, and then the users do not need to rely on a single device and they could access their data anytime everywhere.

Facebook, for example, provides users to share their statuses, photos, and videos on their post wall. Their friends can also like or reply the statuses. Therefore we integrate the feature with ours application. Patients can upload their photos, videos to cloud servers, sharing with their friends. Moreover adding patients' family or doctors as friend and uploading patients' health check report can let patients' family or doctors know their health condition, instead of sending email. Personal Health Assistant becomes not only health check recorder but also a social platform sharing photos, video, and even their health check report.

Ultimately, we fulfill our proposal. Patients could take health check anytime everywhere by using medical device and view the result in their smartphones, including body analysis, body temperature, blood glucose, and blood pressure, even sharing them with their doctors, family or friends to let them know patients' condition. We think the application can provide integration of the four medical detection devices, which are body analysis, body temperature, blood glucose, and blood pressure respectively. Also, it has the advantages of providing portable, convenient. Moreover, chart view is a feature providing patients or doctor to view obviously. The personal health assistant aims to play a necessary role in elders' and patients' daily life.

Related Work

A healthcare application should apply to everyone including patients and common people. In other words, we believe that monitoring health status by self in the daily life is the responsibility for everyone, so our application design to make more motive power for users to take care of their health through social networks. Some market-available healthcare products or applications did not offer to common users. For instance, there is an application [6] of observing and analyzing ECG (Electrocardiography) waveforms on Android devices. The application tends to give medical staff a convenient and portable platform to see the heart status of patients since traditional devices such as PCs which are too heavy to carry. ECG waveforms analyzing is unusual to common people, therefore, they could not recognize what happened on his status. Our application does not aim to specific people but common people who wonder their health status.

Another paper [7] introduces their product how to measure elderly's blood pressure (BP) and pulse rate. Here we have more measurement devices such as blood pressure (BP), Blood Glucose (BG), Body Temperature (BT) and user can record data about the Body Analysis Scale (BAS) by inputting BMI, weight and etcetera. Most important of all, the Daily Check App can send the application data via any useable network (3G, Wi-Fi) to the server, and the collected statics in server can represent to others via "Social Network". "Social Network" is a private platform allows you to post your healthy situation or chat with friend online, sharing every special moment to your old friends or families.

Ultimately, we want to view our health check on the go with diversity data of our health, so our team provides an application to collect and analyze the data and convert it into line chart. Above of functions available on the market in some application, but we want to create more interaction during the people and be attentive to others. For this purpose that everyone can share their health check and some photos or videos to their friend, nurses, and doctors to tell the user's condition and health recent, so we combine the social network.

Personal Health Assistant System

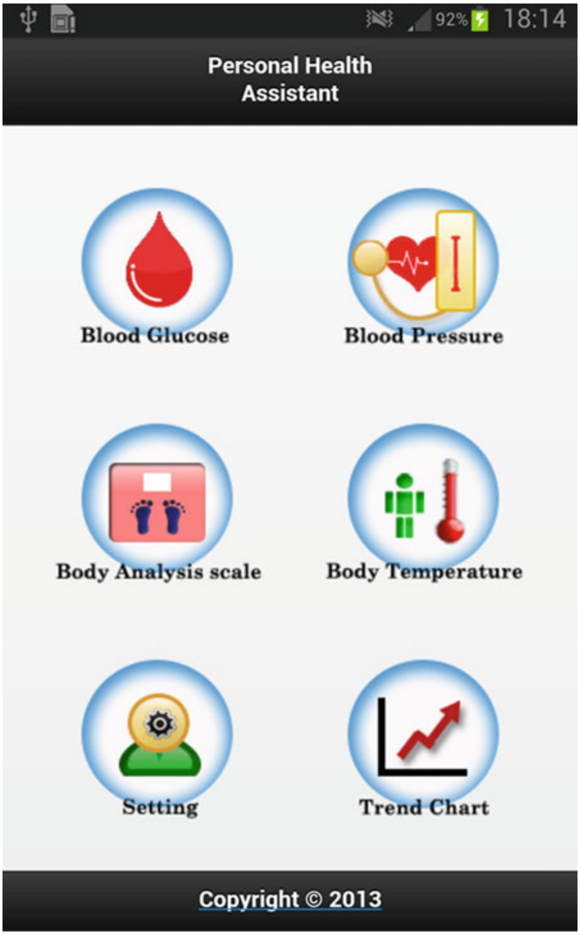
System Architecture

The entire system is based on PhoneGap which is a solution of the development of cross-platform mobile applications. PhoneGap allows us build user interface with web technologies such as HTML5, CSS3, and JavaScript. But there is a little difference between mobile applications design and website design. For example, mobile applications have to consider its screen size and its work flow which is different from web design. Consequently, our team imported a JavaScript framework—jQuery Mobile which is a unified, HTML5-based user interface system. JQuery Mobile solves the problems of responsive pages [8] to apply many distinct kinds of screen sizes. Besides, jQuery Mobile used AJAX technology on switching between pages. With combination of PhoneGap and jQuery Mobile, we could briefly build the entire architecture and the user interface. Moreover, it also keeps a good user experience [9].

Another feature of jQuery Mobile is that it provides many essential components of user interface which has been styled, and then we can simply design buttons, header, and footer and so on.

In the start page of the application, we directly place six large buttons lead to each measure pages of health status and setting page (Fig. 1). At each measure

Fig. 1 Start page



page, we divide it into four parts: Input, Data, Chart and Email, and these parts separately provide distinct functions. For this reason, we design four tabs on the top to switching these functions (Figs. 2 and 3). At the last, we also have a setting page for configurations, and a Trend Chart page to integrate all charts of measures in one page to see an entire situation of health about self.

For the input, we provide three ways to obtain measured value:

1. Users manually key in data.
2. Receive the data via Bluetooth protocol from the measure device
3. Receive the data via NFC protocol from the measure device

See Fig. 3, users can simply tap one of three buttons to obtain input and store into a database. For data display, the Data function will list all records from the database. For a run chart, the Chart functions provide a clear chart composed of all

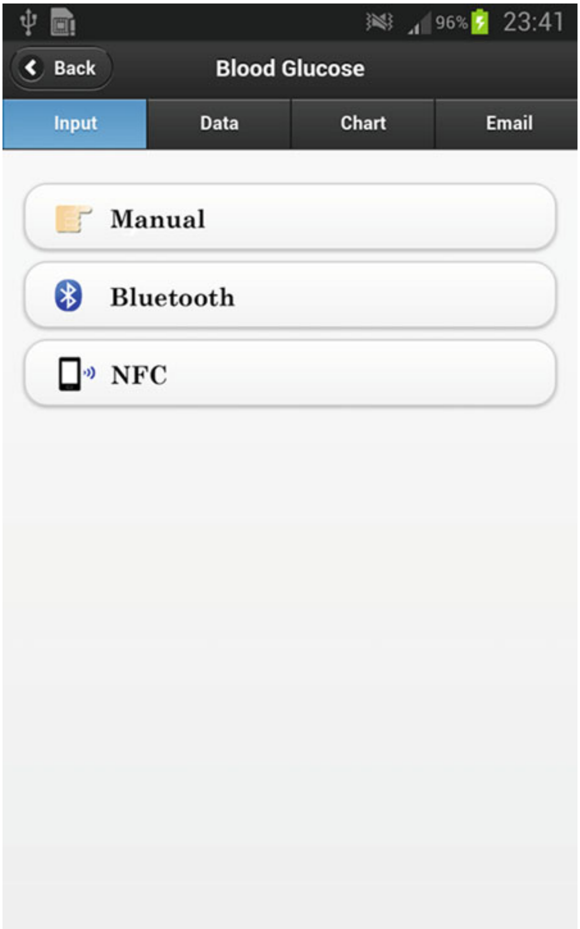


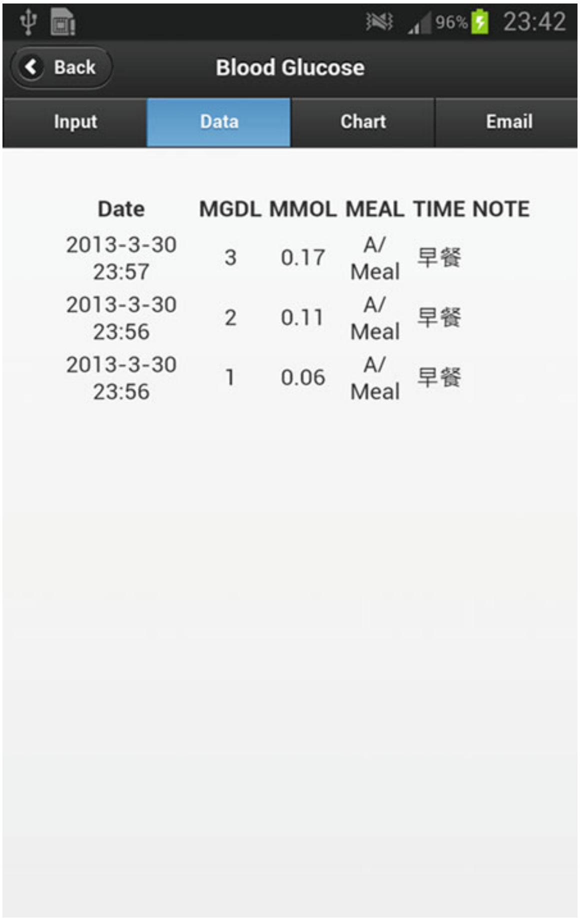
Fig. 2 Blood glucose page

records. In the end, the Email function could send an E-mail contains weekly or monthly records to a doctor or someone by keying in an E-mail address.

Consequently the Fig. 4 show the hierarchy chart of our user interface design.

Storage on Smart Phones

SQLite is a relational database management system contained in a small C programming library. In contrast with other database management systems, SQLite is not a separate process that is accessed from the client application, but an integral part of it.



Date	MGDL	MMOL	MEAL	TIME	NOTE
2013-3-30 23:57	3	0.17	A/ Meal	早餐	
2013-3-30 23:56	2	0.11	A/ Meal	早餐	
2013-3-30 23:56	1	0.06	A/ Meal	早餐	

Fig. 3 Data tab-page

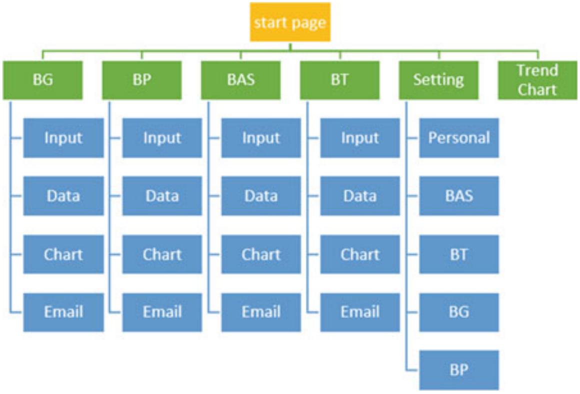


Fig. 4 Hierarchy chart of UI design

SQLite is ACID-compliant and implements most of the SQL standard, using a dynamically and weakly typed SQL syntax that does not guarantee the domain integrity.

SQLite is a popular choice as embedded database for local/client storage in application software such as web browsers. It is arguably the most widely deployed database engine, as it is used today by several widespread browsers, operating systems, and embedded systems, among others. SQLite has many bindings to programming languages. The source code for SQLite is in the public domain.

Display Chart on Smart Phones

For the purpose of data visualization on different platforms, we use up-to-date JavaScript and HTML5 technologies to draw the graph. JavaScript and now HTML5 canvas allow for quick and easy 2D drawing and are built into all modern browsers. HTML5 Canvas is also supported by all modern browsers and mobile devices meaning our charts and graphs will be seen by the widest possible number of users. Let’s briefly introduce HTML5 canvas and JavaScript.

HTML5 Canvas

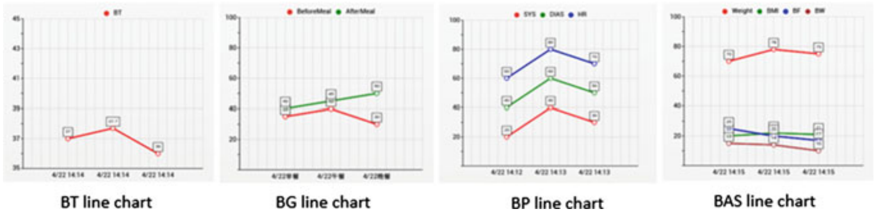
HTML5 canvas is a new HTML tag. It allows bitmap drawing and is controlled by JavaScript. In other word, it likes a piece of paper which is a part of page and we can draw on it by using JavaScript [10, 11].

Canvas uses a “fire and forget” drawing methodology to renders its graphics directly. If we want to change something, usually we must redraw the entire canvas and this is important when providing animated or interactive charts to users. To conquer this problem, canvas need fast to draw on and very responsive.

When we build a canvas, we have to define a drawable region in HTML code with height and width attributes. JavaScript code will access the area through a set of drawing functions like other common 2D APIs, thus allowing for dynamically generated graphics. Now HTML5 canvas usually is used to build animations, graphs, image composition, and games.

JavaScript

JavaScript is one of interpreted computer programming language. It is also the world’s most popular programming language and implemented as part of web browsers so that client-side script could interact with the user, control the browser, communicate asynchronously and alter the document content that was displayed. It



BT line chart

BG line chart

BP line chart

BAS line chart

Fig. 5 Line charts of BT, BG, BP, and BAS

is a multi-paradigm language, supporting object-oriented, imperative, and functional programming styles [12].

By using JavaScript and HTML5 technologies, our application can be cross-platform between iOS and Android and even other mobile operating systems.

In our application, Fig. 5 respectively shows distribution of BT, BG, BP and BAS data over time.

Cloud and Sociality

What does connecting with could actually do? When the data was sent to a remote server, the users can watch their records and charts directly. On the other hand, users do not even have anything but a browser; they could access their records anywhere. The result is that promotes the users' data portability.

On the implementation of the client side, we used PhoneGap Connection API to detect whether the Wi-Fi or 3G/4G enabled. If there are data which are not been uploaded, the client will send a HTTP request to the web server and upload remain parts of data in the SQLite database.

On the server side, we built the environment by LAMP [13] which is a web service bundle to process all requests from clients. We write a processor in PHP to receive requests and send back responses. Later, the processor will store the uploaded data to the MySQL database. Finally, the data will be presented with records or charts on the web pages with the help of a JavaScript library of charts. With charts, users could easily understand their health condition that is consistent with the mobile application.

IMS [14], IP Multimedia Subsystem, is an architectural framework for delivering IP multimedia services. The users can connect to IMS in various ways, such as via WLAN or 3G/4G, so does the IMS terminals may be mobile phones, PDAs, or computers. Therefore, we establish an IMS server in a Linux system. The users can have voice communications or video communications with their friends through our IMS server, following the SIP (Session Initiation Protocol) [15]. It provides the function like Skype allowing patients to chat together or share the experience to each other and help patients feel less lonely.

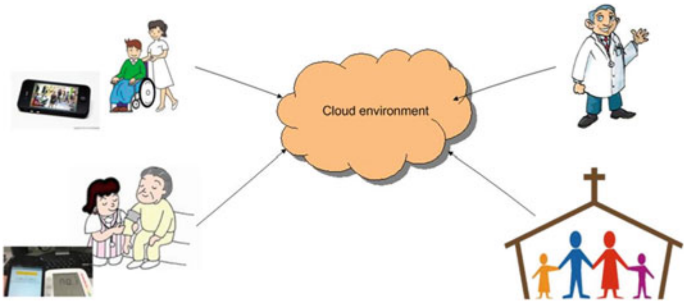


Fig. 6 Scenario of system

Scenario of System

Due to the implementation mentioned above, we can have a scenario for our system (Fig. 6). As far as the patients who need health care concerned, our system provide that not only patients can take health checks at home and uploading the result to the cloud server, but also the doctors can review the patients' reports. Furthermore, patients can use the social part of our application to know the situation of their family or friends. Considering the patients who are elder are not capable to read amount of words on the screen of smart phones, our social application are designed to only allow users sharing the photos and the videos which was taken recently. Even if the patients' bad health condition, the patients' do not lose the chance to know the situation of their family or friends.

Discussion

As far as elder is concerned, they were suffer from disease of every variety. Health condition needs to be considered anytime. Body analysis, body temperature, blood glucose, and blood pressure are the common health condition. By observing the value of them, we can discover the difference from previous data, even detecting the chronic diseases, to achieve the effect of prevention and treatment.

Although, elder can take health check at home, it is difficult to operate the complicated medical device. What they need is the simple and user-friendly way to use. Therefore, using Personal Health Assistant can simplify the complicated operation.

However, there are some difficulties to develop cross-platform application. First, PhoneGap is framework like wrapper that wraps the website with PhoneGap and deploy to the mobile platforms. It means the JavaScript functions are mapped to native function with intermediary. Actually PhoneGap is a web view to view the website which we write with HTML, JavaScript, and CSS. Therefore, comparing with the native application, PhoneGap application has a fatal disadvantage that it is inefficiency and slow. Second, there are some function are not be supported, because website app is not the same as mobile app. Take android for example; Android API use java as standard programing language, but web app is made with HTML not suitable for Android API. The solution is to write plugins that let JavaScript function call native code function. Nonetheless, it is not a good solution, writing plugins means you have to write with native code, against the original intention of cross-platform.

Conclusion

Integrating the smartphone application technology with medical service, and the universality of smartphone make taking health check more convenient and simpler. Therefore, elder or patients are willing to take health check more frequently. It can effectively make people pay close attention to own health.

Acknowledgments The paper is supported partially by the "Cross-platform smart phone-enable stroke prevention and rehabilitation health system" project and partially by the NSC-101-2811-E-008-013 project.

References

1. Zheng P, Ni LM (2006) Spotlight: the rise of the smart phone. Distributed systems online, IEEE 7.3
2. 2013 4th week 內政統計通報(101年底人口統計分析), 2013.01.26, department od stastics, ministry od the interior, R.O.C. (TAIWAN)
3. Stastics of Causes of Death (2011) Dec 2012, Department of health, executive Yuan, R.O.C. (TAIWAN)
4. Butler M (2011) Android: changing the mobile landscape. Pervasive Comput, IEEE 10(1):4–7
5. Junyan LV, Xu S, Li Y (2009) Application research of embedded database SQLite. In: IEEE international forum on information technology and applications, IFITA'09, vol 2
6. Lorenz A, Oppermann R (2009) Mobile health monitoring for the elderly: designing for diversity. Pervasive Mob Comput 5(5):478–495
7. Hii P-C, Chung W-Y (2011) A comprehensive ubiquitous healthcare solution on an Android™ mobile device. Sensors 11(7):6799–6815
8. Darie C (2006) AJAX and PHP: building responsive web applications. Packt Publishing Ltd
9. Hassenzahl M, Tractinsky N (2006) User experience-a research agenda. Behav Inf Technol 25(2):91–97
10. Allen S et al (2010) Pro Smartphone cross-platform development: iPhone, Blackberry, windows mobile and android development and distribution. Apress
11. Kessin Z (2011) Programming HTML5 applications: building powerful cross-platform environments in JavaScript. O'Reilly Media, ISBN: 978-1-449-39908-5
12. Negrino T (1998) JavaScript for the World Wide Web. Peachpit Press, San Francisco
13. Rosebrock E, Filson E (2006) Setting up LAMP: getting Linux, Apache, MySQL, and PHP working together. Sybex

14. Camarillo G, Garcia-Martin M-A (2007) The 3G IP multimedia subsystem (IMS): merging the Internet and the cellular worlds. Wiley
15. Rosenberg J, Schulzrinne H, Camarillo G, Johnston AR, Peterson J, Sparks R, Handley M, Schooler E (2002) SIP: session initiation protocol, RFC 3261, Internet engineering task force, June 2002

Cross-Platform and Light-Weight Stroke Rehabilitation System for New Generation Pervasive Healthcare

Eric Hsiao-Kuang Wu, C. C. Tseng, Y. Y. Yang, P. Y. Cai, S. S. Yen and Yu-Wei Chen

Abstract We propose a concept allowing disabled patient taking their rehabilitation treatment without going to the hospital. We design a portable device of rehabilitation and it can be lightweight in an advanced version. Particularly, we select PhoneGap to be a mobile development framework and to build the specific application for mobile devices using JavaScript, HTML5, and CSS3 instead of using device-specific language in our system. The great advantage is cross-platform that represent write once and run anywhere. In our software, we lead the patient to do the exercise of rehabilitation step-by-step and then store recovery status in database. Finally, the system will play restful music when patient do exercise. According to the mentioned above, we developed the Mobility Rehabilitation that combines mobility, entertainment, and storing ability.

Keywords Rehabilitation · PhoneGap · Lightweight · Cross-platform

E. H.-K. Wu (✉) · C. C. Tseng · Y. Y. Yang · P. Y. Cai · S. S. Yen
Computer Science and Information Engineering, National Central University,
Jhongli, Taiwan
e-mail: hsiao@csie.ncu.edu.tw

C. C. Tseng
e-mail: ben482922@gmail.com

Y. Y. Yang
e-mail: q0955895260@yahoo.com.tw

P. Y. Cai
e-mail: tigebrp6@hotmail.com

S. S. Yen
e-mail: freehourse2sh@gmail.com

Y.-W. Chen
Department of Neurology, Landseed Hospital, No. 77, Kwang-Tai Road, Ping-Jen 32449,
Tao-Yuan County, Taiwan
e-mail: yuwchen@gmail.com

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_143, © Springer Science+Business Media Dordrecht 2014

Introduction

Recently, Smart Phone has being widely used and the number of applications also increases. It is not only because of the prevalence of Smart Phone, but also the strong operating systems that provide the users to install software and games easily. Compared with all the existing operating systems, Android seems to be more developer-friendly since it releases it's source code and the kernel and file system that allow programmers to modify and fix the bugs. For example, other systems don't release the open source including Windows Mobile and Apple iOS. Due to the convenience for both users and developers, Android-based Smart Phone is now most popular in the market.

The percentage of the population of elders is increasing in Taiwan for decades. Stroke and dementia are key problems of the functional impairment among the old people, lead great impact to the whole society. Stroke is one of the major causes of dementia. Several risk factors might cause this cerebrovascular attack including diabetes, hypertension, hyperlipidemia and smoking. Besides the acute onset of neurological deficit, the pathological changes of small cerebral vessels have been associated chronic deterioration of mentality. Modified Framingham Stroke Risk Profile (FSRP) is proposed to predict the risk of stroke in 10 years and had been proven to be an independent predictor for the declining cognition over time. To prevent and cure the cerebrovascular and degenerative disorders, we need to develop novel strategies for their prevention, treatment and rehabilitation.

Android market, an online store could download the Android applications, provides variety of applications; nonetheless, the rehabilitation applications haven't caught people's attention. In the past, physical therapists have been gradually decreased, but patients who suffer from work injuries, sports injuries and stroke increased need the system. Moreover, in our current rehabilitation system, it is inconvenient for individuals with mobility problems to receive the treatment. Without the extra help, patients are difficult going to hospitals or rehabilitation clinics. These statements address the importance of rehabilitation applications, providing some simple methods for patients to take treatment at home. This may decrease the frequency for patients to go to the hospitals, make the best use of hospitals' facilities and handle issue of therapists' shortage. Thus, our goal is to design and implement a pervasive mobility rehabilitation system [1].

First, we use standardized web APIs to tap into device's motion sensor called accelerometer. The accelerometer can capture device motion in the x, y, and z direction. Then we can make the design of the corresponding sensing to meet the demand for rehabilitation motion, according to the different requirements of each patient.

Second, we choose the PhoneGap to develop the program that can run on the heterogeneous platform. Third, when the sensing procedures are completed and detected correctly, we need a mechanism to record the achievement of rehabilitation. We use a database to record date, time, frequency, and other related information of the rehabilitation by using the standardized web APIs.

Until now, almost functions have been completed, but we think the design of the system is not perfect enough to make the mobile rehabilitation device light and handy. In the later section, we will make a lighter device connect to our Smart Phone. This improvement makes our system be much closer to the "mobile" characteristic.

Related Work

In recently years, there are more and more papers explore in improving the rehabilitation training [2] uses android-based language a specified language to achieve a novel ubiquitous rehabilitation system (Ubi-REHAB) based on Augmented Reality (AR) technology, which is designed to enhance the recovery of upper extremity functions in patients with stroke. Although their system seems good, their development tool limits the usage of the program. On the contrary, we use PhoneGap to be the development platform in our rehabilitation system. This makes our system more attractive because of the cross-platform characteristic of Phonegap. It reduce the burden on the developer because they don't need to worry about the problem brought by the diversity of the platform. Most importantly, it reduces the threshold while the application is applied. It benefits patients. Another Ref. [2] shows a prototype for a new generation of game therapies based on Smartphone to achieve the purpose of rehabilitation of articulations. For some patients with serious condition, they can't exert their power. According to this situation, we develop another motion sensor which is much lighter than Smartphone to help more patients. This can address another advantage of our system. Patients not only can play game on Smartphone but also can do exercise on peripheral device in our system. Finally, "cross-platform" and "lighter" differentiate the difference from other paper.

Mobility Rehabilitation System

System Architecture

We proposed four mainly characteristics of our system development. First, we developed a mobile phone app for rehabilitation training. Patients can do the rehabilitation through the smart phone anytime at any place. Second, we use PhoneGap, a mobile development framework package Web application, as a native application to develop our application. We take advantage of its cross-platform characteristics to solve problems due to the variety of the platform. Third, we use a database to store patient's information. And the last one point is we develop another external embedded mobile device to reduce the full application weight comparing with using Smartphone to taking the rehabilitation training. According to the mentioned above, our system make patients to do rehabilitation simply, conveniently, easily, and efficiently.

Rehabilitation on Smartphone

With an increasing of mobile device usage in our society and the worldwide deployment of mobile application, the vision of Pervasive Healthcare or healthcare to anyone, anytime, and anywhere could be possible. In our stroke rehabilitation application, we design a suit of rehabilitation procedures especially for patients with stroke. In order to make sure that users could enjoy the rehabilitation without complicated instructions and misunderstanding, we make our system to be more user friendly for patients and we also make our user interface present with facilitation and the brisk style.

This application chiefly uses two sensors. One is originally built-in a smart phone, called accelerometer. The other sensor is Gyro-Sensor. Both of them capture device motion in the x, y, and z direction. Furthermore, our application uses database to preserve the data. So that doctors can know about patients' condition. We also play some music while patients take the rehabilitation training by our application.

In addition to achieve the exquisite design in our application, we also cooperate with the local hospital for further studying in the procedure of rehabilitation and the related movements. After we visit the rehabilitation center, we realize how the patients feel when they take the rehabilitation training. And due to this precious experience, we adjust the degree of difficulty of the program to an appropriate level. The hospital staff gave us many diverse professional suggestions. Through bilateral communication and coordination, we improve our application with proficiency as well as completeness. And following is the operative instructions to the application (Fig. 1).

When patients entering the application, patients would see two big buttons. One of them records the latest progress, the other lets user enter the main menu. In the primary stage, there are four types of the rehabilitation activities (Comb, Arms raised, Arms declined, Wrist flip) available for the patients. After choosing one of the buttons, there is a demonstrative picture to guide user how to do the motion. And just press the "start" button, set the game level and then start enjoying it. If the mission is finished, they could get reward as well. Besides, even if the game is forced to interrupt, the program could record the latest progress, so the previous score and the reward are still recorded in the database.

Rehabilitation on Bracelet

In the beginning, we use two embedded boards to develop a lighter sensor roughly (Fig. 2). One is Bluetooth module playing role as a transmission tool transmits

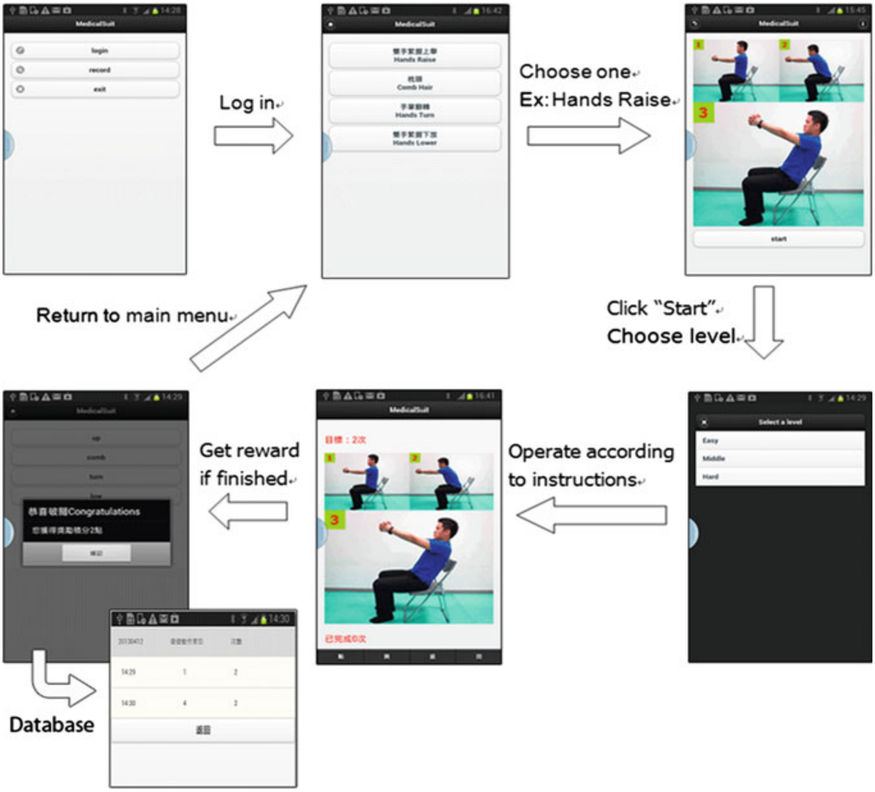


Fig. 1 Flowchart of smart phone application

statics about the angular velocity of the three-axis every moment from the other circuit board called Gyro-Sensor to smart phone. And the work for the Gyro-Sensor is motion sensing precisely. When phone receives data, it will determine whether the motion is correct or not. And this sensing motion will not be done until patients make the right motion. In the end, when all of the components are completed, we combine these two circuit board to a much perfectly lighter sensor we called bracelet. With this bracelet, rehabilitation process is much easier [3].

PhoneGap

We can say that there is a trend that more and more programmers or teams join the mobile developments of application. PhoneGap is just the platform that can produce cross-platform mobile application based on HTML, CSS, and JAVASCRIPT (Fig. 3). It has some great features worth mentioning. The first one is compatibility that make application write once and run anywhere (Fig. 4). Second, PhoneGap

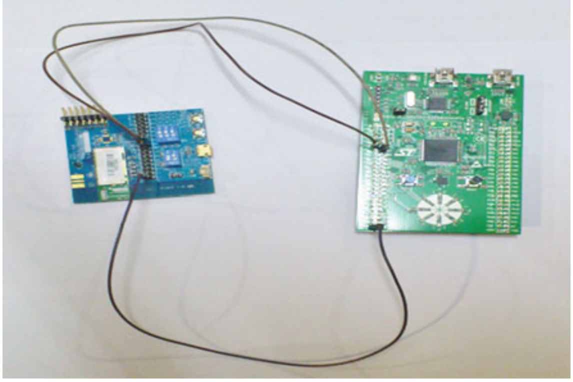


Fig. 2 Bluetooth and Gyro-Sensor

use W3C standard and combine with jQuery Mobile make development more powerful. But it has the problem of lower performance compared to native code. So according to different request, choose suitable platform to develop applications to meet their needs. For example, high-level games need high-speed requirement, so developer use native code to develop. But in our system, we pay most of our attention on the utilization of the rehabilitation application, so we use PhoneGap to be our development platform [4]. In conclusion, the characteristic of "cross-platform" will benefit more patients [5].

G-Sensor and Gyro-Sensor

In our system, we use two kinds of sensor. And the difference between them is their sensing way. One is G-Sensor called accelerometer responses for detecting "triaxial acceleration". This sensor is also in the general smart phone. The other is Gyro-sensor which detects "angular velocity" in the x, y, and z direction. Giving a more precise description, these "3-axis" gyroscopes we used have a single sensing structure for motion measurement along all three orthogonal axes different from other gyroscopes that use two or more structures. This sensing way can eliminates

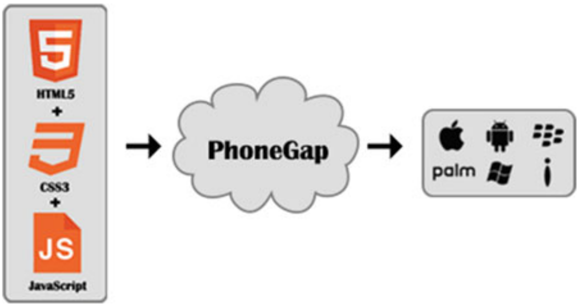


Fig. 3 Schematic diagram for PhoneGap

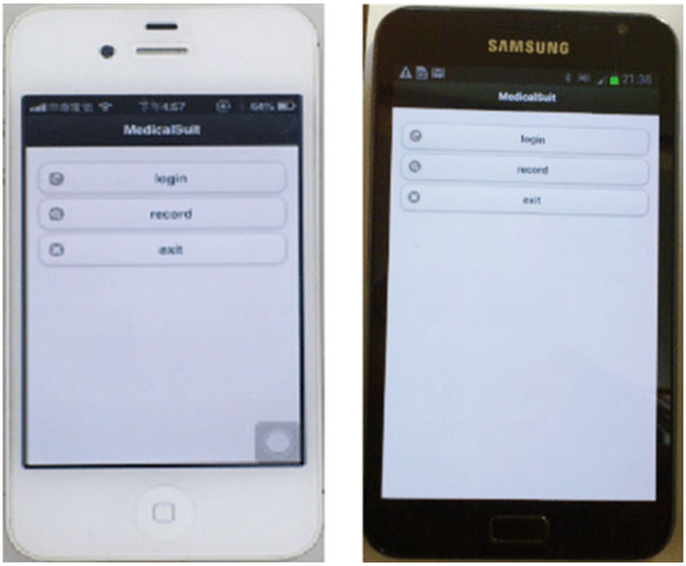


Fig. 4 Cross-platform characteristics of PhoneGap

any interference between the axes that inherently degrades the output signal, increasing accuracy and reliability of motion-controlled functionalities. From the above mentioned, we also can use gyroscope to be our powerful motion sensor [6].

Scenario of System

Due to the implementation mentioned above, we can have a scenario for our system (Fig. 5). For a patient who needs to do rehabilitation training, it is difficult to take a long trek to rehabilitation clinic. And even the patient arrives at the hospital, he can only do the boring training as usual. At this moment, our rehabilitation APP has an opportunity to show. No matter patient is traveling or going to some place lacking of rehabilitation clinics, he or she can still do rehabilitation.

After the treatment of our system to the patient, our system will record the patient's rehabilitation result and upload this record to database. Thus, doctor can check the database to learn about the condition of rehabilitation for patients.

Discussions

Most of physical disable person need to take some training for recovery. Rehabilitation can make significant improvement on the recovery of patients' body.

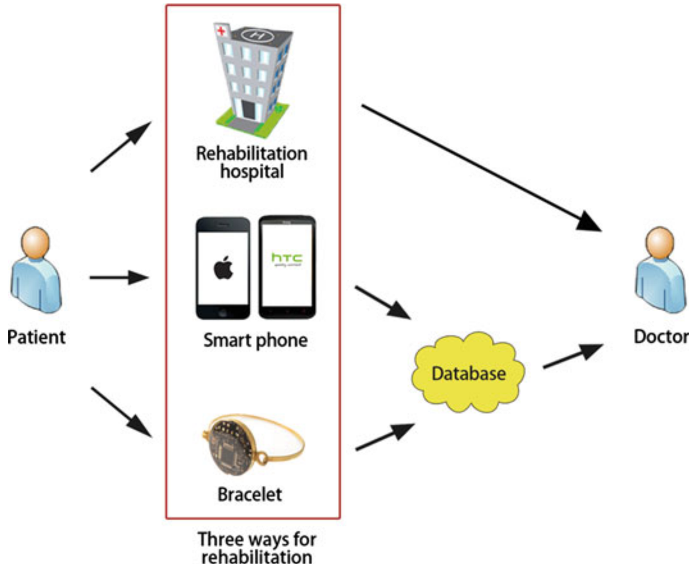


Fig. 5 Scenario of system

However, the patients need long-term rehabilitation to recovery after the operation. This makes tremendous changes in their life. Not only the patient's daily-life, but also their families'. Recently, the demand of rehabilitation is dramatically increasing in medical resource. Also, as the aging society approaching, the long-term care for the elder is becoming more and more important.

In the past, patients need to go to the hospital in person to use the rehabilitation instruments which are large and expensive. After getting home, they might not continue their rehabilitation because of the lack of the instruments. Our project is aim to dig out a way to let the patient of stroke to rehabilitate continually at their home in an easier and more comfortable way.

Conclusion

Due to the popularity of smart phones, there is a new developing area in medical care. Convenience and efficiency in the further medical care is going to be more important in the future. Therefore, we implement a cross-platform rehabilitation application on the smart phone., and we also realize an embedded bracelet for our mobile application.

In the future, we will pay attention to improve our accuracy of the motion detection by sensors. We can also implement a feedback mechanism in our mobile application, so we can provide a different rehabilitation training course by considering the different recovery situation of patients. Due to pursuing the passive health care, there are some undeveloped medical area we can explore and work on. And patients who suffer from stroke will be more and more cheerful when they take rehabilitation training day by day.

Acknowledgments The paper is supported partially by the "Cross-platform smart phone-enable stroke prevention and rehabilitation health system" project and partially by the NSC-101-2811-E-008-013 project.

References

1. Varshney U (2003) Pervasive healthcare. IEEE Comput Mag 36(12):138–140
2. Choi Y (2011) Ubi-REHAB: an android-based portable augmented reality stroke rehabilitation system using the eGlove for multiple participants. In: International conference on virtual rehabilitation (ICVR), 27–29 June 2011, pp 1–2
3. Deponti D, Maggiorini D, Palazzi CE (2011) Smartphone's physiatric serious game. In: IEEE 1st international conference on serious games and applications for health (SeGAH), 16–18 Nov 2011, pp 1–8
4. Motoi K (2005) Development of a wearable device capable of monitoring human activity for use in rehabilitation and certification of eligibility for long-term care. In: 27th annual international conference of engineering in medicine and biology society. IEEE-EMBS 2005, 17–18 Jan 2006, pp 1004–1007
5. Sin D (2012) Mobile web apps—the non-programmer's alternative to native applications. In: 5th international conference on human system interactions (HSI), 6–8 June 2012, pp 8–15
6. Smutny P (2012) Mobile development tools and cross-platform solutions. In: 13th international carpathian control conference (ICCC), 28–31 May 2012, pp 653–656
7. Barthold C (2011) Evaluation of gyroscope-embedded mobile phones. In: IEEE international conference on systems, man, and cybernetics (SMC), 9–12 Oct 2011, pp 1632–1638

Off-line Automatic Virtual Director for Lecture Video

Di-Wei Huang, Yu-Tzu Lin and Greg C. Lee

Abstract This research proposed an automatic mechanism to refine the lecture video by composing meaningful video clips from multiple cameras. In order to maximize the captured video information and produce a suitable lecture video for learners, video content should be analysed by considering both visual and audio information firstly. Meaningful events were then detected by extracting lecturer's and learners' behaviours according to teaching and learning principles in class. An event-driven camera switching strategy was derived to change the camera view to a meaningful one based on the finite state machine. The final lecture video was then produced by composing all meaningful video clips. The experiment results show that learners felt interested and comfortable while watching the lecture video, and also agreed with the meaningfulness of the selected video clips.

Keywords Video content analysis · Video composition · Virtual director · Lecture video analysis

D.-W. Huang · G. C. Lee
Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan, Republic of China

Y.-T. Lin (✉)
Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taipei, Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_144, © Springer Science+Business Media Dordrecht 2014

Cycle Embedding in Alternating Group Graphs with Faulty Elements

Ping-Ying Tsai and Yu-Tzu Lin

Abstract The alternating group graph, which belongs to the class of Cayley graphs, is one of the most versatile interconnection networks for parallel and distributed computing. Cycle embedding is an important issue in evaluating the efficiency of interconnection networks. In this paper, we show that an *n*-dimensional alternating group graph AG_n has the following results, where F is the set of faulty vertices and/or faulty edges in AG_n: (1) For $n \geq 4$, AG_n -F is edge 4-pancyclic if $|F| \leq n - 4$; and (2) For $n \geq 3$, AG_n-F is vertex-pancyclic if $|F| \leq n - 3$. All the results are optimal with respect to the number of faulty elements tolerated, and they are improvements over the cycle embedding properties of alternating group graphs proposed previously in several articles.

Keywords Alternating group graph · Pancyclicity · Fault-tolerant · Cayley graph · Cycle embedding · Interconnection network

P.-Y. Tsai (✉)
Taiwan Geographic Information System Center, Taipei, Taiwan, Republic of China
e-mail: bytsai0808@gmail.com

Y.-T. Lin
Graduate Institute of Information and Computer Education,
National Taiwan Normal University, Taipei,
Taiwan, Republic of China

Y.-M. Huang et al. (eds.), *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, Lecture Notes in Electrical Engineering 260,
DOI: 10.1007/978-94-007-7262-5_145, © Springer Science+Business Media Dordrecht 2014