VSUM: Summarizing from Videos

Yu-Chyeh Wu
National Central University, Jhongli City, Taoyuan, Taiwan
bcbb@db.csie.ncu.edu.tw

Yue-Shi Lee
Ming Chuan University, Gwei Shan District, Taoyuan, Taiwan
leeys@mcu.edu.tw

Chia-Hui Chang
National Central University, Jhongli City, Taoyuan, Taiwan
chia@csie.ncu.edu.tw

Abstract

Summarization on produced type of video data (like news or movies) is to find important segments that contain rich information. Users could obtain the important messages by reading summaries rather than full documents. The researches in this area could be divided into two parts: (1) Image Processing (IP) perspective, and (2) NLP (Nature Language Processing) perspective. The former put emphasis on the detection of key frames, while the later focused on the extraction of important concepts. This paper proposes a video summarization system, VSUM. VSUM first identifies all caption words, and then adopts a technique to find the important segments. An external thesaurus is also used in VSUM to enhance the summary extraction process. The experimental results show that VSUM could perform well even if the accuracy of OCR (Optical Character Recognition) is not sophisticating.

1. Introduction

Large video data is growing up in quick with the rising of Internet. Oh and Bandi [15] defined three kinds of video data:

(1) Produced Type: like news, movies.
(2) Raw Type: like the surveillance, and traffic video.
(3) Medical Type: like the Echocardiogram.

The produced type of data often contains rich and diverse information. For example, the Discovery movie, Napoleon, described several important events that occurred in Napoleon's life. They included the experience of his growing, his wife Josephine, and some important fights when he became the emperor of France.

The second type is to monitor the abnormal situations among all frame sequence. Most of image processing researchers addressed large work on motion detection and pattern recognition. Their target goal is to detect the exceptional behaviors in a given frame sequence, which was called key frame detection. For example, [5] [9] [15] developed systems, which could online cluster similar frames and find the key frames. The medical type of video data is usually used to identify organs or objects in animals.

As mentioned above, the produced type of video data is more popular than others because of its variety and richness. Unfortunately, this kind of data has grown up in quick. Thus, it is very difficult for people to view all of the videos in details. A summary (often ten-to-thirty percent of the original document) gives users an overview description rather than a full document. This is obviously when summarizing long papers. Sentences in the summary are often short but important. Users could find their interesting parts in a short summary, and only chose these parts to view. This paper presents a system, VSUM, for automatic summarization from produced type of video data.

Previous researches for summarization could be categorized into two classes:

(1) Image Processing (IP) perspective.
(2) Nature Language Processing (NLP) perspective.

The former addressed on pattern recognition and object motion detection. They have brought a lot of successful on the raw type of video data. However videos in produced type have its variety and specific content, e.g., the much of transitions on scenario and important descriptions in caption words. Traditional image processing was focused on the abnormal behaviors detection in frame sequence. It may not be suitable for the produced type of video data, since the rich backgrounds and roles.

On the other hand, researchers who work in NLP area also performed well in the news-like documents, but it may not be suitable for produced type of video data. Most of the news articles described specific

Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04)
0-7695-2217-3/04 $20.00 © 2004 **IEEE**

events, and almost all of the content does not bias to the main topic very much. Usually a video data is long and has several unexpected segments that may irrelevant to the main topic. Thus we should improve the original techniques to suit for this kind of videos.

The remainder of this paper is organized as follows: Section 2 describes the overview of our system which contains two parts: OCR module and Summarization module. Section 3 introduces OCR module for identifying words from images. Section 4 describes the paragraph segmentation while Section 5 talks about the summarization techniques based on previous results. Section 6 discusses the tested data and comparing the system performance. Finally concluding remarks and future work are given in Section 7.

2. Overview of VSUM

The video summarization diagram of VSUM is given in Figure 1.

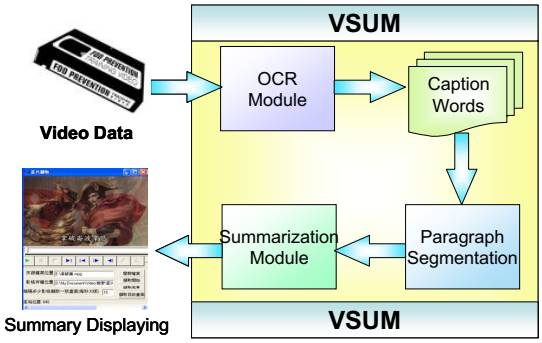

Figure 1: Overview of VSUM

While videos (like movies or news) input to the VSUM, the OCR module is trigger to recognize all of the caption words. The words in the same frame were regarded as a sentence. The paragraph segmentation then group several sentences into passages. Finally, our summarization module ranks both for sentences and passages. For example, when we want to summarize a discovery film, i.e., Napoleon, VSUM would process the video data as input. The OCR module identifies caption transitions and recognizes all caption words. Figure 2 is one of the images from Napoleon. We treat all of the words which appear in the same frame as a sentence. For instance, all of the words "拿破崙波拿巴" ("The Napoleon Bonaparte") in Figure 2 would be regarded as a sentence. After words and sentences identification, the summarizer selects the most informative sentences or paragraphs as summary, and the video player would display each important shots.

Figure 2: An image extracted from Napoleon

3. OCR Module

OCR Module processes all of the input frame sequence and identifies all of the caption words. Figure 3 shows the key processing flow of this module.

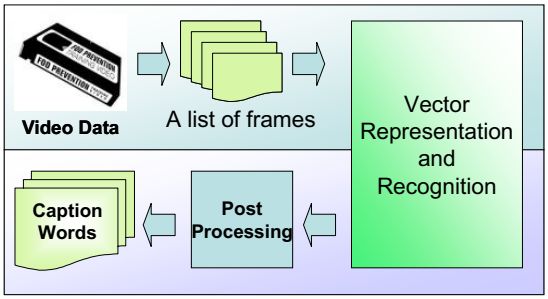

Figure 3: The framework of OCR module

Similar to other video processing researches [1] [11], the first step of OCR module is to decompose input video into list of frames. Then the kernel module starts the following steps: filtering, representing, and character recognizing. These three steps are described below.

3.1 Filtering

Filtering is an important processing before identify characters. Briefly speaking, this stage could remove the noisy blocks. We employ some of the well-known techniques, like: large non-relevant areas removing, extraordinary line deletion and multi-frames cleaning, details can be seen in [4] [11] [21] [22].

For example, Figure 5 (a) shows the result of binarizing and removing large black areas from Figure 2. Figure 5 (b) is the result after extraordinary line deletion. Images which pass through filtering techniques can make the text area clearly.

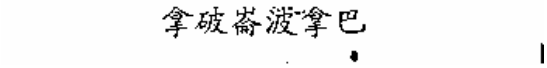

Figure 5(a): Filtering by removing large black areas

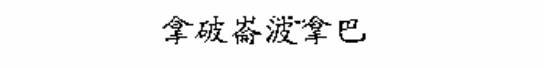

Figure 5(b): Filtering by removing extraordinary lines

3.2 Representing

It should be noted that the OCR task was built on single character recognition. In order to separate each image character, we adopt some heuristics. The image character identification can be considered as a classification task. Each character can be represented as a vector space, and this vector is used to find the word category.

In VSUM, we combined several feature selection algorithms to represent the character vector. Table 1 shows these approaches of VSUM.

Table 1: List of feature selection methodologies

Feature Name	Number of Dimensions
Peripheral feature from the four corners [12]	48
Black jump distribution for 0, 45, 90 degrees [18]	40
Projection value to the bisecting perpendicular line [10]	48
Peripheral feature from the bisecting perpendicular line [18]	48
Local Stroke Distribution [4]	256
Total dimensions	440

3.3 Character recognizing

Each single character can be viewed as a vector, and this step is used to find the most relevant characters in training set. As mentioned in 3.2, OCR task is a kind of classification. To solve the classification problem, an algorithm is used to find the most relevant category which suits for a testing vector. There are many satisfied classification algorithms, like Support Vector Machines (SVM), Neural Network [18], k-Nearest Neighbor (kNN) [4], and Decision Trees. In this paper we used a kNN to categorize each character.

4. Paragraph segmentation

A video data is often long and includes several sub-topics. Traditional summarization approaches could perform well in news-like articles. But it may not be suitable for a long document. Summarizing a long document is difficult than a short one since its branches of sub-topics. Generally speaking, there are more than 4710 words in a video firm, but only 300-1000 words in each news article. By the way, news articles often short, but important. The best way for processing long document is to group several neighbor sentences into a passage. The passage can be seen as a meaningful group. Then, we can derive the most important passage as summary. This is different from other research, because the long document (video) could not be treat as common news article.

4.1 Segmentation

We assume that if a sentence occurs after its previous sentence tightly, it should be viewed as the same passage. Different from flat-document, events in a video have their time occurrence, i.e. we could consider the time-stamp of each sentence. For example, Figure 1 occurred in frame 325, and its previous sentence occurred in frame 270. The frame number is the time occurrence of the sentence.

Figure 6 shows the initial statistical histogram of caption words distribution. Gap means the time difference between two near sentences, and the vertical axis is the number of words in a sentence. After calculating the mean, median, min/max gap, we set the cutting threshold[1], and all of caption words are going to aggregate to passages by gap difference estimation. For example: frame #5235, frame #5520, and frame #5850 would be separate into two paragraphs:

$passage_a$: frame #5235, and frame #5520
$passage_b$: frame #5850

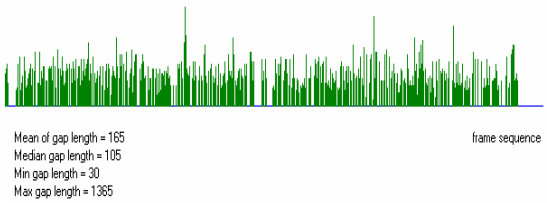

Figure 6: Distribution of words with time sequence

It is trivial that the time difference between frame #5235 and frame #5520 is less than the cutting threshold, and frame #5520 and frame #5850 is greater than threshold. Therefore the two frames would be clustered in the same passage, and the last one is the starting of another passage.

[1] The cutting threshold in this paper is 300.

5. Video Summarization

Summarization from videos is a novel research, as mentioned above, previous image processing researchers devoted the pattern recognition approaches to extract the behavior of some specific objects. This paper focused on the content-based summarization system, which extracts important passages or sentences from produced-type of video data.

Figure 7 is the summarization module of VSUM. Since the previous modules, OCR and paragraph segmentation produced the passages and sentences from input video data. The pre-processor uses the Chinese word segmentation system to find the word boundaries. Finally, the sentence scorer would rank each sentences or passages based on some criteria, and the summary of the video would be displayed.

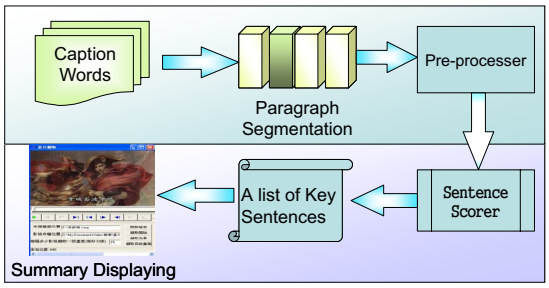

Figure 7: The framework of summarization module

5.1 Pre-processing

The extraction of English words is much easier than (from a space symbol) Chinese. The famous word segmentation system of Chinese is the AutoTag 1.0[2] which was developed by Academia Sinica had great success on this task. In addition, the system also provide a tagging service for terms, by this way, we could use the extra information to improve the quality of summary. For example, noun and verb tags often bring much more information value than other tags, like preposition, and conjunction.

5.2 Sentence scorer

NTCIR[3], DUC[4] defined two summary types:

(1) Abstraction type
(2) Extraction type

Terms in abstraction type of summaries would be different from the original document, while the later exploit the extraction of important sentences from the original document. It should be noted that video summarization is away from traditional text summarization, due to its visual aids. If the summary tries to broke the sentence or insert some words from other resources, it could not display the summary as a movie player. Luckily, the extraction type of summary could be used to be displayed since it estimates the importance of each sentence. In order to make it flexible, the summary would be generated in this type.

The sentence scorer is referred from three criteria:

(1) Density of content words:
It is clearly that a sentence with high ratio of content word is more important. Chinese content words have two characteristics: Bigram, and Verb/Noun tag. Chinese content words should satisfy with the above constraints. We can calculate the density score by count the number of content words per sentence, and the following formula is the evaluation function.

$$\text{Density } S_i = \frac{\#of\_content\_words\_in\_S_i}{length\_of\_S_i}$$

(2) TFISF:
TFISF (Term Frequency*Inverse Sentence Frequency) is a variant measurement of TFIDF (Term Frequency*Inverse Document Frequency) for content words. We ignore the TF value, instead of ISF, since there has low proportions that a term appears twice in the same sentence (there is not a term which occurs more than twice in the same sentence in our testing video corpus). The TFISF estimation can be shown as follows:

$$\text{TFISF } S_i = \sum_{|S_i|} ISF(t_j)$$

(3) Thesaurus knowledge:
We adopt a simple thesaurus mapping criteria in this paper, which expand the title words from input video. When a sentence contains title words, it gets w_1 scores; similarly the w_2 means that the sentence contains thesaurus words. Table 2 shows the result of title words expansion from title words in WordNet thesaurus[5].

[2] http://godel.iis.sinica.edu.tw/CKIP/ws/
[3] http://research.nii.ac.jp/ntcir/
[4] http://duc.nist.gov/
[5] http://www.cogsci.princeton.edu/~wn/

Due to the lake of Chinese external thesaurus, we use three steps title words expansion in this paper:

(1) Chinese-English translation
(2) WordNet reference
(3) English-Chinese translation according to WordNet results

Table 2: Title words expansion from WordNet

Query: Napoleon	
1	Napoleon, Napoleon I, Napoleon Bonaparte, Bonaparte, the Little Corporal -- (French general who became emperor of the French (1769-1821))
2	napoleon -- (a rectangular piece of pastry with thin flaky layers and filled with custard cream)
3	Napoleon, nap -- (a card game similar to whist; usually played for stakes)

The translation task is done by another famous bilingual translation system[6]. We select all of the synonyms and content words with the first gloss from WordNet and use the following criterion:

$$T(S_i) = \sum_{|Si|} (M_{ij} \times W_1 + N_{ij} \times W_2)$$

M_{ij} is the flag of term j in a sentence i, if term j match with the title words. N_{ij} is another flag which enables if term j match with the thesaurus words.

There are still some problems for this approach, like the Word Sense Disambiguation (WSD) and Machine Translation (MT). Here, we do not take action to solving the ambiguity senses instead we select the top 1 word meaning from WordNet, because the higher ranker the meaning, the more probabilistic the word sense. Further, the MT result is the other problem for this approach, and it can be done by a famous translation system.

5.3 Output selection

In this paper we present two kinds of summary output: passage level and sentence-level. Summary in type of Passage-level would display the most meaningful passage among all others as the summary, and sentence-level plays the important sentences according to their time-occurrence. In order to make consistent between these two summary types, the number of sentences in the sentence-level should be equivalent with the number of sentences in the passage -level. By this way, the number of words in a summary can be determined automatically without parameter tuning. Comparing with other researches, this is the first trial in automatic summary length control in long document.

6. Data Sets and Experiments

6.1 Dataset

Video films used for testing our approach are the descriptive video films from Discovery. All the features of our testing corpus were listed in Table 3. At the beginning, all films were converted into MPEG-1 format, with resolution 352x240 pixels per frame. Most of the qualities of the video frames are satisfied with filtering (as mentioned in 3.1).

Table 3: Features in video corpus

Total number of video films	30
Total number of sentences	16126
Total number of characters	141441
Average number of sentences per video film	537.53
Average number of words per video film	4714.7

The short description of discovery movie could be viewed on the web site (http://www.discovery.com),

6.2 Results on OCR

Table 4 lists the OCR performance on inside and outside testing results, and each character had been identified by human recognition. The OCR testing data: Napoleon which describes important events of Napoleon and some great wars that lead by him.

Table 4: OCR result on Napoleon

	Correct	Incorrect	Unknown
Inside test	92.1%	7.6%	0.0%
Outside test	81.5%	16.3%	2.0%

For outside test, the OCR module identifies each character with trained and non-feedback character database, while inside test makes use of all the character database contains original and feedback ones.

Comparing with advanced researches [12], the OCR results in lower performance, since the character set and the images are quite in difference, they testing set of them are text-images, without any complex and variety backgrounds; while compare with [11], the

[6] http://www.dreye.com/

Figure 8: Summary results in Passage-Level
Video source name: Napoleon

OCR accuracy was outperformed. It is worth to note that OCR error does not affect the quality of summary too much, since our summarizing technique aims to extract the importance keywords in the document, when the keywords are identified successfully, we can ignore the OCR error of other non-relevant terms. However, the limits caused by OCR error still controls the results of summary. We will improve the accuracy of OCR module in the future research to make a better quality of summary.

6.3 Results on summarization

Figure 8 shows the partial summary result in passage-level which derived from the video Napoleon. The order of displaying sequence is left-to-right and top-to-down. The numbers of sentences in summary of Napoleon are five, so does the other proposed summary type: sentence-level. The side of right-bottom is the English summary after translating the original summary. Figure 9 is the sentence-level of the Napoleon, and all of the descriptions are the same as Figure 8. Due to the paper length of limitation, the remaining summarization result (in JPEG format) was also available on this web site (http://140.115.51.16/VSUM/index.html).

6.4 Discussion

For the performance of video OCR, the result shows that our training corpus might effectively recognize the caption words in inside test, which gets 92% percent of accuracy, but did not perform well in outside test with 81.5%. Besides the problems of the simpler feature selection and filtering techniques, in addition, we also find three problems in the OCR module:

(1) the selection of classifiers
(2) the processing of image
(3) the different font type between training and testing data

Hong et al.[4] reported that the nearest classifier still could identify words effectively, but it still a slightly weaker than some famous classification algorithms, like Probabilistic Neural Nets, Learning Vector Quantization, Support Vector Machines,…etc. If the OCR module was trained on these advanced learning algorithms, we trust that the performance of the character recognition would be raised up further. Secondly, the unsatisfying results of image processing (especially character segmentation) are another limitation of the OCR result. Almost the cutting error would cause the error in the identification stage. When the character segmentation could be improved, the OCR result would be also improved again. Finally, due to the different font type, the performance is also

1. His was still expected by his parents.
2. He planed to chase French away from his hometown-Corsica.
3. They were in an inferior position with fewer armies and poorer equipment.
4. Napoleon hold the chance of promotion, and he swear to be faithful to the government.
5. Napoleon made use of the trend of the times of the revolution, and reached the top of success.

Figure 9: Summary results in Sentence-Level
Video source name: Napoleon

limited in the lake of similar training character sets. Moreover, the passage-level type of summary shows its better readability than the sentence-level one, this is because the passage often contains more context information than a single sentence. However this phenomenon is more frequent and notable in long-documents. For example, the descriptions of the first frame in Figure 9, is mentioned that the parents of Napoleon had much highly expectations to him, but the second and third frames are related to the army and fight, surprisingly, the remaining two frames talk about the reason why Napoleon could succeed. There should be a main topic in a tight summary which conveys the important subjects from the original documents.

7. Conclusions and future work

In this paper passage segmentation and cross-language knowledge integration are presented. The simple but reliable segmentation technique aims to split a long document into several passages, and these will be used for further generating two types of summarizations. Different from the news video, the segmentation of the sub-topics can be identified by the speaker change, background change, and audio type concisely, but it is difficult to construct a complexity model to identify the shot for each scene. The speaker of each discovery video film is always the same man, and the background is always changing. Therefore a low-cost segmentation technique could help us to group sentences into passages.

On the other hand, in order to extend the vocabulary, we combine an extra-knowledge base and use a famous English-Chinese-Japanese multi-lingual translation system to help us to add up more lexical words across two languages. This is because the subjects, topics of Discovery videos are from the western stories, and the lake of reliable Chinese knowledge base. In addition, traditional summarization techniques may be suit for news-like articles, but it would cause the summary difficult to read. We improved and presented the passage-level type of summary based on the traditional summarization criterions and by integration of external knowledge base. The result shows the high readability and informative. In the future, we would try to further combine more internal resources, like speech, and external resources, like Internet to make the model more robust.

References

[1] M. Brunn, Y. Chali and B. Dufour, "The University of Lethbridge Text Summarizer at DUC 2002", In *Proceedings of the Document Understanding Conference*, 2002, pp. 39-44.

[2] W. H. Cheung, K. F. Pang, M. R. Lyu, K. W. Ng, and I. King, "Chinese Optical Character Recognition for Information Extraction from Video Images", *In Proceedings of International Conference on Imaging Science Systems and Technology (CISST)*, 2001, Vol. 1, pp. 269-275.

[3] J. Goldstein, M. Kantrowitz, V. Mittal, J. Carbonell, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics", *In Proceedings of the 22th ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 121-128.

[4] T. Hong, S.W. Lam, J.J. Hull, S.N. Srihari, "The Design of a Nearest-Neighbor Classifier and Its Use for Japanese Character Recognition", *In Proceedings of Third International Conference on Document Analysis and Recognition*, 1995, Vol. 1, pp. 270-291.

[5] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, S. Harp, "Urban Surveillance Systems: from the Laboratory to the Commercial World", *In Proceedings of the IEEE*, 2001, Vol. 89, No. 10, pp. 1478-1497.

[6] K. Ishikawa, S. Ando and A. Okumura, "Hybrid Text Summarization Method based on the TF Method and the Lead Method", *In Proceedings of the 2nd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop*, 2001.

[7] J. Kupiec, J. O. Pedersen, and F. Chen, "A Trainable Document Summarizer", *In Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 68-73.

[8] H. Jing, "Sentence Simplification in Automatic Text Summarization", *In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, 2000.

[9] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic Monitoring and Accident Detection at Intersections", *IEEE Transactions on Intelligent Transport System*, 2000, Vol. 1, No. 2, pp. 108-118.

[10] S. H. Kim, "Recognition of Handwritten Korean Address Strings by Tight-Coupling of Minimum Distance Classification and Dictionary-Based Post-Processing", *Journal of Computer Processing of Oriental Languages*, 1998, Vol. 12, No. 2, pp. 207-221.

[11] C. J. Lin, C. C. Liu, H. H. Chen, "A Simple Method for Chinese Video OCR and Its Application to Question Answering", *Computational Linguistics and Chinese Language Processing*, 2001, Vol. 6, No. 2, pp. 11-30.

[12] C. L. Liu, I. J. Kim, and J. H. Kim, "Model Based Stroke Extraction and Matching for Handwritten Chinese Character Recognition", *Pattern Recognition*, 2001, Vol. 34, Issue 12, pp. 2339-2352.

[13] Y. Nakao, "How Small a Distinction Among Summaries Can The Evaluation Method Identify?", *In Proceedings of the 2nd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop*, 2001.

[14] C. Nobata, S. Sekine, M. Murata, K. Uchimoto, M. Utiyama and H. Isahara, "Sentence Extraction System Assembling Multiple Evidence", *In Proceedings of the 2nd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop*, 2001.

[15] J. H. Oh, and B. Bandi, "Multimedia Data Mining Framework for Raw Sequences", *In Proceedings of Multimedia Data Mining of Knowledge Discover in Database (MDM/KDD) Workshop*, 2002, pp. 1-10.

[16] K. Ohtake, D. Okamoto, M. Kodama, and S. Masuyama, "Yet another Summarization System with Two Modules Using Empirical Knowledge", *In Proceedings of the 2nd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop*, 2001.

[17] D. Radev, "Text summarization tutorial", *In Proceedings of the 23th ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.

[18] R. Romero, D. Touretzky, and R. Thibadeau, "Optical Chinese Character Recognition Using Probabilistic Neural Network", *Pattern Recognition*, 1997, Vol. 8, No. 30, pp.1279-1292.

[19] Y. Seki, "Sentence Extraction by Tf/idf and Position Weighting from Newspaper Articles", *In Proceedings of the 3rd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop*, 2002.

[20] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, Vol. 22, No. 8, pp. 888-905.

[21] J. C. Shim, C. Dorai and R. Bolle, "Automatic Text Extraction from Images and Video for Content-Based Annotation", *In Proceedings of International Conference on Pattern Recognition*, 1998, pp. 618-620.

[22] V. Wu, R. Manmatha, and E. M. Riseman, "TextFinder: An Automatic System to Detect and Recognize Text in Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, Vol. 21, No. 11, pp.1224-1229.

[23] H. Yang, L. Chaison, Y. Zhao, S. Y. Neo, and T. S. Chua, "VideoQA: Question Answering on News Video", *In Proceedings of the 11th Annual ACM International Conference on Multimedia (ACMM)*, 2003, pp. 632-641.

[24] D. Zhang, and W. S. Lee, "Question Classification using Support Vector Machines", *In Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2003, pp. 26-32.