Vietnamese handwriting recognition for automatic data entry in enrollment forms

Article · May 2015

DOI: 10.1109/ICITEC.2014.7105589

3 authors, including:

Nguyen Hoach
Hanoi Achitecture Univeristy
8 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Shinq-Jen Wu
Da-Yeh University
47 PUBLICATIONS **456** CITATIONS
SEE PROFILE

Vietnamese Handwriting Recognition for Automatic Data Entry in Enrollment Forms

Hung Pham-Van,
Vietnam Academy of Science and Technology,
hungpv@ioit.ac.vn

Hoach The Nguyen,
Hanoi Architecture University,
nthoach@gmail.com

Shinq-Jen Wu
Dayeh University,
jen@mail.dyu.edu.tw

***Abstract*—This paper presents an efficient method of feature extraction for Vietnamese handwriting and an approach to group Vietnamese characters into similar sub-classes to support recognition process. Furthermore, it's the algorithm to look up two dictionaries simultaneously with least database storage in the post-processing in order to increase the recognition accuracy. The results of our approach on recognizing Vietnamese handwriting such as name, birthplace and birthday are also presented. By using Vietnamese Personal Name Dictionary for post-processing, the recognition accuracy for Vietnamese characters, words, person name is 91.23%, 85.41% and 96.03% in comparison with 75%, 79.67% and 77.81% from common methods correlatively. The statistics are based on 2384, 2248 and 302 testing samples of Vietnamese character, words and personal name respectively. The recognition accuracy for province/city name is 99.8% among 64 provinces of Vietnam. Our approach is embedded into the automatic data entry system for enrollment forms, VnHandwritten1.0, the first commercial application in recognizing Vietnamese handwriting in Vietnam. On this application, our method proves itself an efficient method for commercial purpose.**

Keywords—Automatic data entry; Vietnamese handwriting recognition; post processing.

I. INTRODUCTION

Vietnamese handwriting (VH) recognition problem still challenges applications. From the practices of data digitalizing with Vietnamese handwriting, we propose a number of methods and techniques to efficiently recognize VH. From our applications and results, we summarize our achieved progress on VH recognition with three innovations as following:

- Feature extraction of 5-sub images.
- Sub-classification for recognition
- Dynamic Programming based algorithm for simultaneous checking two dictionaries.

The achieved results on recognizing VH fields in forms such as character, words, personal name, birthplace and birthday are presented. The application is embedded in automatic data entry system for enrollment forms using a scanner to release much effort and time of office staff.

Logically, our paper is presented into 4 parts with 3 parts for 3 innovations, 1 for results hereinafter.

II. FEATURE EXTRACTION

A recognition system typically consists of three main parts: feature extraction, classification and post-processing. Feature extraction is a crucial step because it contributes to the accuracy of the whole recognition system. As a first part, its error will affect seriously the following parts. Among a variety of feature extraction techniques, the decision on a specific technique depends on recognized objects. Vietnamese writing has its own characteristics to be recognized. Therefore, in this work, we propose an efficient method of feature extraction.

The feature extraction is performed by extracting the original image into three parts: the upper mark (`,',?,~,^,v), the under mark (.) and the body of character. The system recognizes the body and the marks individually before recombining the recognized body and marks.

Each original image is separated into 5 sub-images as follows:

- Mark separated by space

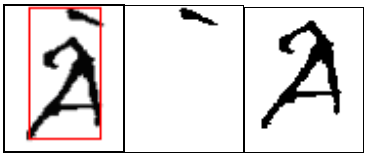


Fig. 1. Separated by connecting

- Mark separated by region

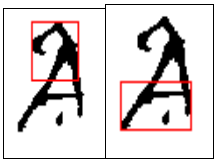


Fig. 2. Separated by regions

The 5 sub-images are normalized as follows:

- Determining the boundary of each sub-image
- Converting into uniformed 16x16 matrixes

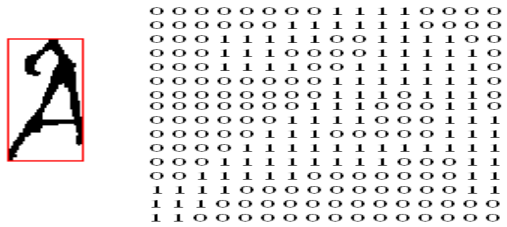


Fig. 3. Boundary and uniformed 16x16 matrix

The five equivalent 16x16 matrixes are used to recognize VH characters in following steps.

III. SUB-CLASSIFICATION

Vietnamese writing system has totally 89 characters. We categorized these characters into similar sub-classes to increase the accuracy of system. For each sub-class, we used its own specific features to recognize instead of common ones. For example, the feature of marks should be extracted in the accented character subclasses. The way to categorize characters into subclasses and how to perform recognition (as shown in flow chart) is detailed hereafter:

Subclasses:

23 classes without marks:
{[A],B,C,D,Đ,[E],G,H,[I],K,L,M,N,[O],P,Q,R,S,T,[U],V,X,[Y]}

12 classes of E: { E,È,É,Ẻ,Ẽ,Ê,Ề,Ế,Ể,Ễ,Ệ,Ê}
18 classes of A:
{A,À,Á,Ả,Ã,Â,Ầ,Ấ,Ẩ,Ẫ,Ă,Ằ,Ắ,Ẳ,Ẵ,Ạ,Ậ,Ặ}
6 classes of Y: {Y,Ý,Ỷ,Ỹ,Ỵ}
6 classes of I: {I,Ì,Í,Ỉ,Ĩ,Ị}
12 classes of U: {U,Ù,Ú,Ủ,Ũ,Ư,Ừ,Ứ,Ử,Ữ,Ụ,Ự}
18 classes of O:
{O,Ò,Ó,Ỏ,Õ,Ô,Ồ,Ố,Ổ,Ỗ,Ơ,Ờ,Ớ,Ở,Ỡ,Ọ,Ộ,Ợ}
Classes with drop marks:
{EÈÉẺẼÊỀẾỂỄ, ỆỆ}
{AÀÁẢÃÂẦẤẨẪĂẰẮẲẴ, ẠẬẶ}
{YÝỶỸ, Ỵ}
{I Ì Í Ỉ Ĩ, Ị}
{U Ù Ú Ủ Ũ Ư Ừ Ứ Ử Ữ, Ụ Ự}
{O Ò Ó Ỏ Õ Ô Ồ Ố Ổ Ỗ Ơ Ờ Ớ Ở Ỡ, Ọ Ộ Ợ}
Classes with upper mark:
{EẸ, ÈÉẺẼÊỀẾỂỄỆ}
{AẠ, ÀÁẢÃÂẦẤẨẪĂẰẮẲẴẠẬẶ}
{YỴ, ÝỶỸỴ}
{IỊ, Ì Í Ỉ Ĩ,}
{UỤ, Ù Ú Ủ Ũ Ư Ừ Ứ Ử ỮỰ}
{OỌ, Ò Ó Ỏ Õ Ô Ồ Ố Ổ Ỗ Ơ Ờ Ớ Ở Ỡ, Ộ Ợ}
4 classes of mark of YUI: {`,',?~}
8 classes of mark of EO {`,',?,~,^`,^',^?,^~}
12 classes of mark of A {`,',?,~,^`,^',^?,^~,v`,v',v?,v~}

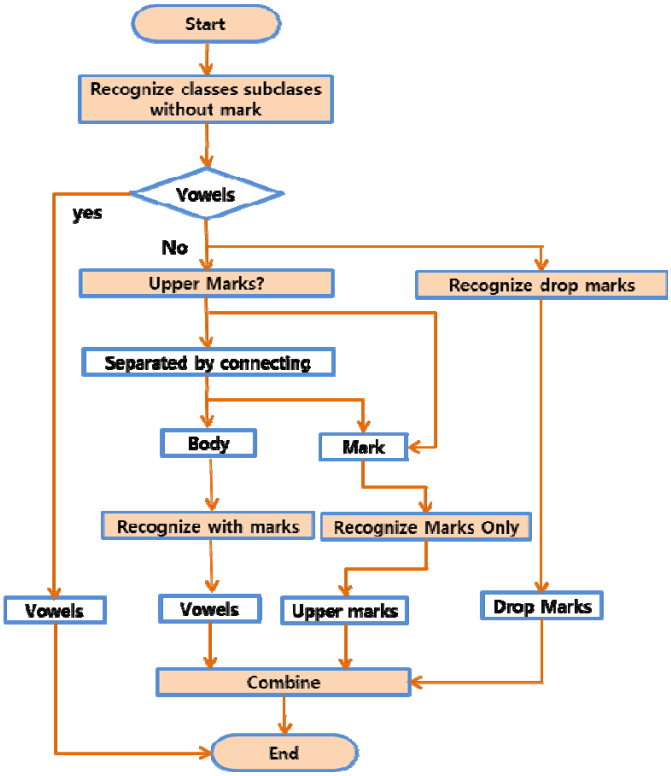


Fig. 4. Recognition flow chart

IV. POST PROCESSING WITH TWO DICTIONARIES AND DYNAMIC PROGRAMING

In the processing stage, each word is checked in a specific dictionary. If the recognized word is not included in any dictionary, it will be probably the combined phrase from two dictionaries. For example, Vietnamese personal names consist of two parts: Family name and Given name. Therefore, we have family name dictionary and given name dictionary. To reduce the number of combined names in one dictionary, we use the two separated dictionaries simultaneously to conduct the post processing with dynamic programing algorithm as follows.

Definition:

Let's consider A, B as two dictionaries. If word x belongs to dictionary A then the notation is: $x \in A$.

If the word x is combined from word x1 and x_2 then the notation is: $x=x_1|x_2$.

The combination of two dictionaries is defined as:

$C=A+B=\{x|\ x \in A \text{ or } x \in B\}$

The multiplication of two dictionaries is also defined as:

$C=A*B=\{x=x_1|x_2,\ x_1 \in A \text{ and } x_2 \in B\}$

$\underline{x}$ is called the invert of word x, means:

if x="abc" then $\underline{x}$="cba".

$\underline{A}$ is called the invert dictionary of A, it means:

$\underline{A}=\{\underline{x} \mid x \in A\}$

Number of words in dictionary A is defined as: card(A). Therefore:

$$card(A+B)=card(A)+card(B)$$

$$card(A*B)=card(A)*card(B)$$

Define matching array of sub-words

In dynamic programming, we use matching array D[k][h] of two words (x and y). Note that: $D_{xy}[k][h]$. k=len(x) and h=len(y). So, $D_{xy}[i][j]$ is matching array of two sub-words of x,y: "$n_1n_2…n_i$" và "$m_1m_2…m_j$".

Given a word x and a dictionary A, We define two array K[] and T[] as following:

$K_{xA}[i]=Min\{D_{xy}[i][len(y)] \mid y \in A\}$ where i=1..len(x)

$T_{xA}[i]=argmin_y\{D_{xy}[i][len(y)] \mid y \in A\}$ where i=1..len(x)

Problem Statement:

Given two dictionaries A and B and a word x, let's find the word in A*B which is best similar to the word x. Provided that complexity of this algorithm is the same as finding in dictionary A+B.

Algorithm:

- Using dynamic programming to correct the $\underline{x}$ (invert of x word) with $\underline{B}$ dictionary to get $K_{x\underline{B}}[]$ and $T_{x\underline{B}}[]$
- Using dynamic programming to correct the x with A dictionary to get $K_{xA}[]$ and $T_{xA}[]$
- Finding the best position (pos_{cut}) to split x into x_1 and x_2 such as: x_1 should be corrected with A dictionary and x_2 should corrected with B dictionary.

$$x=x_1|x_2$$

pos_{cut} is computed as following:

$$pos_{cut}=argmin_j\{K_{xA}[j]+K_{\underline{x}B}[len(x)-j] \mid j=1..len(x)\}$$

So, the x word is corrected to:

$$T_{xA}[pos_{cut}] \mid \{T_{\underline{xB}}[len(x)- pos_{cut}]\}^{-1}$$

Complexity of algorithm:

Because the post processing only works on A dictionary and $\underline{B}$ dictionary. So, the complexity of the algorithm is same working on A+B dictionary.

A. Distance Function "d"

In dynamic programming, we use "d" distance between two characters. The problem is how we choose the distance and which one to choose? The edit distance D(x,y) is commonly used. D(x,y) is the number of operators to edit "x" string to "y" string (delete, replace, insert)

The edit distance has some drawbacks as following:

"ccn" → "con" or "cen" ?

Using edit distance, we can't determine "con" or "cen" because these two words also need one operator to edit from "ccn".

In edit distance, distance between two characters is as following:

$$d(x_i,y_j)=0 \text{ if } x_i=y_j$$

$$d(x_i,y_j)=1 \text{ if } x_i \neq y_j$$

Therefore, the edit distance doesn't consider the similar characters. We propose an improvement on this distance which considers the similar characters by the results of recognition engine.

We get N best results in order of descending accuracy instead of one result. (N best) (N=89 Vietnamese characters).

Let's assume N best results in order of descending accuracy as following:

$$x_0,x_1,…,x_{n-1}$$

The improved distance is defined as d_0 provided that:

$$0 \leq d_0 \leq 1$$

$$d_0(x_0,y_j)=0 \text{ if } x_0=y_j$$

$$d_0(x_0,y_j)=\alpha+(1-\alpha)*(j-1)/(n-2) \text{ if } x_0 \neq y_j$$

α depends on quality of classify engine. For example, with results from SVM technique in *table 2*, we choose α=0.5.

"MIOH->MÍCH"

TABLE I. POST-PROCESSING USING DYNAMIC PROGRAMMING AND NBEST (D_0 DISTANCE)

M(0)	I(0)	O(0)	**H(0)**
N(0.5)	**Í(0.5)**	Ó(0.5)	N(0.5)
	Ì(0.51)	Ó(0.51)	
	L(0.52)	**C(0.52)**	

B. Post-processing with Some Types of Data

- *Post-processing for Vietnamese words*:

We used the word dictionary (6742 words) and dynamic programming to correct the output of recognition system (minimized d_0 distance) with example:

"MIOH->MÍCH"

- *Post-processing for Vietnamese birthplace*:

We also used the province name dictionary (64 provinces) and dynamic programming to correct the output of recognition system (minimized d_0 distance) with example:

"BẠC LIÌU → BẠC LIÊU"

- *Post-processing for Vietnamese personal name:*

Vietnamese personal names are decomposed into 2 parts: "surname" and "name". We use two dictionaries: the Vietnamese surname dictionary with about 200 surnames and the Vietnamese name dictionary with about 10.000 names. So, we have totally about 2 millions Vietnamese personal names by combination surname and name.

By using both dictionaries (surname and name) and dynamic programming, we enhanced the recognition quality (minimized d_0 distance) as this example:

IRẦN VĂN HIẾU->TRẦN VĂN HIẾU

V. EXPERIMENTS AND SOFTWARE APPLICATION

A. Experiment 1

Our data have 100.000 Vietnamese character samples (Black&White, 300dpi) from 611 writers. All samples are divided into 10 parts: 7 parts for SVM training and 3 parts for testing (about 70.000 character samples for training and 26064 character samples for testing).

- With 26064 testing character samples, the accuracy of Vietnamese character is about 84.66% (22065/26064) (Without dictionary).

- With 26064 testing character samples and using Vietnamese word dictionary (6742 words), we generated 2248 words for testing. The accuracy of Vietnamese character recognition is about 89.95% (5951/6616). The accuracy of Vietnamese words recognition is about 85.41% (1920/2248).

- With above 26064 testing character samples and province name dictionary (64 provinces), we generated 511 province name samples for testing. The accuracy of Vietnamese province name recognition is about 99.8% (510/511).

- With the same 26064 testing character samples and person name dictionary, we generated 302 personal name samples for testing. The accuracy of Vietnamese person name recognition is about 96.03% (290/302) with 2 million person name dictionary.

TABLE II. IMPROVEMENT OF PROPOSED SUB-CLASSIFICATION

Number of Vietnamese character for testing (89 classes)	**Accuracy**	
	Proposed sub-classification	**None sub-classification**
26064 (uniform distribution)	22065 **84.66%**	15153 58.14%
2384 (dictionary distribution)	2175 **91.23%**	1788 75.00%

(Using LIBSVM, kernel function is RBF, gamma=0.01)

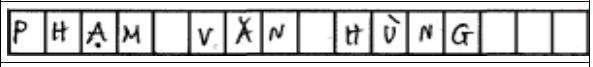
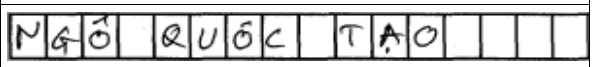
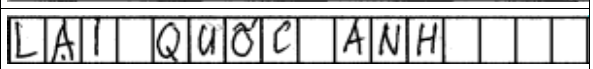
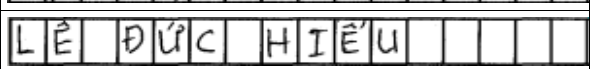
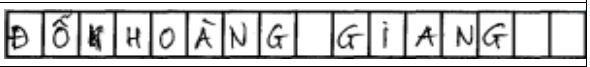
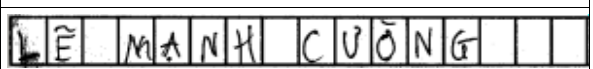
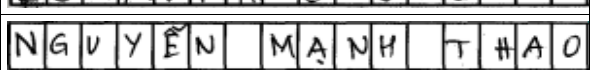
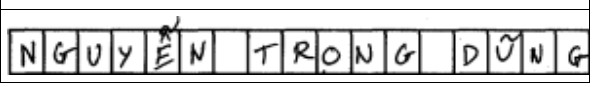
TABLE III. IMPROVED ACCURACY WITH PROPOSED POST-PROCESSING

Type of field recognition	**Number of testing**	**Accuracy**	
		Using person name dictionary with d_0 (proposed)	**Using word dictionary with d (common)**
Vietnamese word (Dictionary has 6742 Vietnamese words)	2248	1920 **85.41%**	1791 79.67%
Vietnamese person name (2 million person name)	302	290 **96.03%**	235 77.81%

B. Random testing

Randomly asking eight people to write his (her) name, there is only one mistake (5th person) because of erasing.

TABLE IV. SOME VIETNAMESE HANDWRITING IMAGES

TT	**Input images of 8 persons for testing**
1	
2	
3	
4	
5	
6	
7	
8	

Result of name recognition

1	PHẠM VĂN HÙNG
2	NGÔ QUỐC TẠO
3	LẠI QUỐC ANH
4	LÊ ĐỨC HIẾU
5	**ĐÔN** HOÀNG GIANG
6	LÊ MẠNH CƯỜNG
7	NGUYỄN MẠNH THAO
8	NGUYỄN TRỌNG DŨNG

C. Software Application

These proposed techniques have been embedded into automatic data entry system for enrollment forms, VnHandwritten1.0. It is the first commercial application of recognizing Vietnamese handwriting technology in Vietnam.

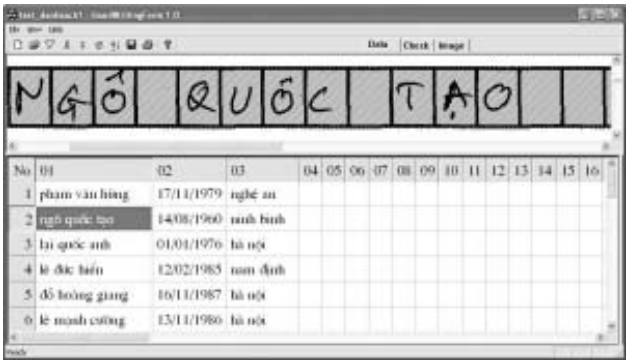


Fig. 5. Automatic data entry system for enrollment forms

VI. CONCLUSION

In this paper, we present three improved techniques and results on VH recognition with three keys: feature extraction with 5 sub-images; categorizing Vietnamese character classes

into similar subclasses, post processing using Vietnamese personal name dictionary and dynamic programming (2 million person names, 2 dictionaries: name and surname).

This result has been applied to automatic data entry system for enrollment forms, VnHandwritten1.0. It is the first commercial application of recognizing Vietnamese handwriting technology in Vietnam. However, our proposed methods are only applied to some handwriting forms such as enrollment forms with fields of names, addresses, places, and words...The VH recognition tasks are conducted on scanned images from fixed scanner. For further works on improving VH recognition quality, our groups will conduct online VH recognizing in associating with mobile scanner, camera or Kinect Sensor…

ACKNOWLEDGMENT

This research is supported by the National Key-Project Program, MOST, Vietnam.

REFERENCES

[1] Ngo Quoc Tao, Pham Van Hung, “Online Continues Vietnamese Handwritten Character Recognition Based on Microsoft Handwritten Character Recognition Library” Circuits and Systems, 2006. APCCAS 2006. IEEE Asia Pacific Conference on. p2024 – 2026.

[2] Pham Anh Phuong, Ngo Quoc Tao, Luong Chi Mai, "An Efficient Model for Isolated Vietnamese Handwritten Recognition," iih-msp, pp.358-361, 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008

[3] Chong Long, Xiaoyan Zhu; Kaizhu Huang; Jun Sun ; Hotta, Y.; Naoi, S. "An efficient post-processing approach for off-line handwritten chinese address recognition", Signal Processing, 2006 8th International Conference on, IEEE,2006

[4] G. Vamvakas, B. Gatos, I. Pratikakis, N. Stamatopoulos, A. Roniotis and S.J. Perantonis, "Hybrid Off-Line OCR for Isolated Handwritten Greek Characters", The Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA 2007), ISBN: 978-0-88986-646-1, Innsbruck, Austria, February 2007, pp. 197-202.

[5] Qiang Fu, X.Q. Ding, C.S. Liu, Yan Jiang, and Zheng Ren. “A hidden Markov model based segmentation and recognition algorithm for Chinese handwritten address character strings”. In ICDAR ’05: Proceedings of the Ninth International Conference on Document Analysis and Recognition, volume 2, pages 590–594, Seoul, Korea, 2005. IEEE Computer Society.

[6] Yong Ge and Qiang Huo. “A comparative study of several modeling approaches for large vocabulary offline recognition of handwritten Chinese characters”. In Pattern Recognition, 2002. Proceedings. 16th International Conference on, volume 3, pages 85–88, 2002.

[7] Zhi Han, Chang-Ping Liu, and Xu-Cheng Yin. “A two-stage handwritten character segmentation approach in mail address recognition”. In ICDAR’05: Proceedings of the Ninth International Conference on Document Analysis and Recognition, volume 1, pages 111–115, Seoul, Korea, 2005. IEEE Computer Society.

[8] Do Nang Toan, Nghiem Anh Tuan, “Improving efficiency of printed Vietnamese character recognition systems”, Special issue Research and Development on Telecommunications and Information Technology, DGPT-POST and Telecommunication Journal, No. 9, pp 82-87. 2003

[9] Vu Hai Quan, Pham Nam Trung, Nguyen Duc Hoang Ha “A robust method for the Vietnamese handwritten and speech recognition” Pattern Recognition, 2002. Proceedings. 16th International Conference on (Volume:3). p732-735, 2002.

[10] Nguyen, D.K., Bui, T.D. “On the problem of classifying Vietnamese online handwritten characters” Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on. p803-808, 2008

[11] Hendrik Pesch, Mahdi hamdani, Hens Forster and Hermann Ney. “Analysis of Processing Techniques for Latin Handwriting Recognition”. Proceedings of 2012 International Conference on Frontiers in Handwriting Recognition. (ICFHR), Page(s): 280 – 284, 2012.

[12] Prasad, J.R. ; Kulkarni, U.V. “Trends in Handwriting Recognition” International Conference on Emerging Trends in Engineering and Technology (ICETET), 2010,Page(s): 491-495.