

RECOGNITION OF MYANMAR HANDWRITTEN COMPOUND WORDS BASED ON MICR

Ei Ei Phyu, Zar Chi Aye, Ei Phyu Khaing, Yadana Thein and Myint Myint Sein
University of Computer Studies, Yangon, Myanmar
E-mail: eieiphyu.ucsy@gmail.com, zarchiaye.ucsy@gmail.com, chuchu0218@gmail.com

KEYWORDS: MICR (Myanmar Intelligent Character Recognition), Statistical and Semantic approach, On-line recognition, voting system, Unicode

ABSTRACT: This paper proposes On-line Handwritten Myanmar Compound Words Recognition System. Under this system, MICR (Myanmar Intelligent Character Recognition) engine is used. It is one kind of ICR (Intelligent Character Recognition) and based on Statistical and Semantic approach and voting system. Statistical and semantic information includes horizontal black stroke count, vertical black stroke count, width and height ratio, number of loops, histogram values, end points etc. When the optimum output is made, the voting system is used. MICR has been successfully developed for Off-line Myanmar characters recognition applications such as car license plates recognition system, reading characters in vouchers, digit recognizers and road-sign recognition system. This paper is the extension of MICR to recognize on-line handwritten Myanmar compound words. In this paper, Tablet device is used to capture handwritten words for on-line acquisition. As soon as these input are recognized, the editable text (Unicode/ ASCII) is produced by using MICR system.

1. INTRODUCTION

Many researchers become more and more interested in character recognition system all over the world. As a result of intensive research and development efforts, systems are available for English language (Bozinovic and Srihari, 1989), Chinese language (Liu Huang and Suen, 1999) and Japanese language (Okamoto and Yamamoto, 1999) (Fernando, Kodikara and Hewavitharana, 2003). There are many papers on handwritten system proposed various features such as chain code sequence, chain code histogram, height and width ratio, number of transition from black and white pixel (Bounnady, Kruatrachue and Wangsiripitak, 2005; Veltman and Prasad, 1994; Tay et al., 2001).

Myanmar language is not as easy as English but not so difficult as Chinese and Japanese languages. Myanmar script has been developed from the Mon script and adapted from southern Indian Pali script. In Myanmar, there are many languages such as Myanmar, Karen, Rakhine, Shan, etc. Myanmar is spoken by (32) million as a first language. Myanmar numbers and characters are of round shape and very similar to each other. That is the reason why it is difficult to recognize with some recognition methods. Although there are many Myanmar research to recognize Myanmar characters, they are not completely finished. Most of the systems are developed through the use of OCR (Optical Character Recognition) like Hopfield, Backpropagation neural network etc. Nevertheless, these systems occur many errors and misrecognition. Some characters which have similar patterns can be misrecognized when using OCR. MICR (Myanmar Intelligent Character Recognition) system can easily overcome these problems. MICR is a kind of ICR (Intelligent Character Recognition) system. Both on-line and off-line image acquisition can be used in MICR. However, it can work well for noise free images, isolated and not broken characters.

In this paper, we propose on-line handwritten Myanmar compound words recognition system based on MICR (Myanmar Intelligent Character Recognition) system. The input image is acquired from Tablet device and that image is recognized via MICR engine and then the editable text (Unicode/ ASCII) will be produced word by word. This paper focuses only on isolated characters.

This paper is organized as follows. In Section 2, we present background knowledge of Myanmar language and Myanmar Unicode system. In Section 3, we describe overview of the system. Myanmar Intelligent Character Recognition (MICR) system is discussed in Section 4. Algorithm for proposed system can be seen in Section 5. Finally, Experimental Results and conclusions are provided in Section 6.

2. BACKGROUND KNOWLEDGE OF MYANMAR LANGUAGE

Myanmar Language (formerly known as Burmese) is a member of Sino-Tibetan language (Krahnstover and Paulhamus, 1999; Swe and Tin, 2005; Htut, 2003). Myanmar alphabet has (33) consonants, (12) basic vowels and (4) medials or semi-vowels and (10) digits. They can be seen in Table (1).

က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ယ	ရ	လ	ဝ	သ
	ဟ	ဠ	အ	

(a) Basic Myanmar Consonants

အ	ဋ	ဌ	ဍ	ဎ
ဏ	တ	ထ	ဒ	န

(b) Myanmar Letter

၁	၂	၃	၄
---	---	---	---

(c) Simi-vowels or Medials

	၁	၂	၃	၄
အ	၁	၂	၃	၄
	၅	၆	၇	၈

(d) Vowels

၀	၁	၂	၃	၄
၅	၆	၇	၈	၉

(e) Digits

Table 1 Myanmar Character Patterns

Some basic consonants may stand as one Myanmar word and combining consonant with one or more extended character become new words. These consonants only exist at the middle position of the words. And there are certain consonants that cannot be combined with extended characters because they do not exist. Figure (1) shows an example of Handwritten Myanmar Compound Word.

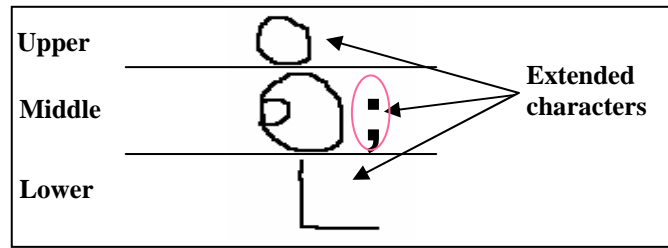


Figure 1 Example of Handwritten Myanmar Compound Word

Myanmar Unicode

The Unicode standard is the Universal Character encoding scheme for written characters and text (Seethalakshmi et al., 2005). The aim of the standard is to provide a universal way of encoding characters of any languages, regardless of the computer system or platform. The Unicode (U+1000 ~ U+109F) is the internationally standardized character set encoding for Myanmar script. The Unicode characters are comprised of hexa decimal values in nature. For example, the Unicode for the character '၀' is 'U+1000', and so on. Some Myanmar Characters Unicode are under research.

3. SYSTEM OVERVIEW

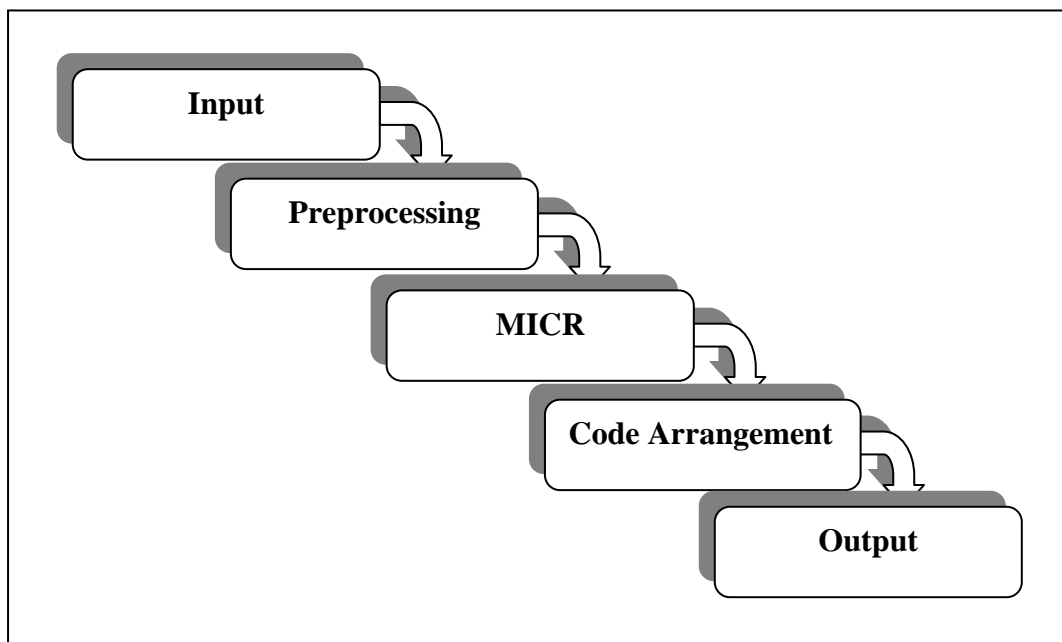


Figure 2 Block diagram of the system

There are five main steps in this system. They are input, preprocessing, MICR, Code Arrangement and Output. In the input stage, on-line images are acquired from Tablet device. In the preprocessing stage, gray scale converting, noise filtering, binarization, extraction and normalization are needed. The preprocessed characters are recognized via MICR engine. In Code Arrangement stage, the recognized characters are arranged according to Unicode/ ASCII Code Order. In Output stage, the editable Myanmar compound words are produced word by word.

4. MYANMAR INTELLIGENT CHARACTER RECOGNITION (MICR)

There are two main methods in character recognition: Intelligent Character Recognition (ICR) and Optical Character Recognition (OCR). OCR is the process which reads text from printed documents and converts them to a machine readable form. ICR is pattern based character recognition and is also known as Hand-Print Recognition. MICR (Myanmar Intelligent Character Recognition) system is one kind of ICR (Intelligent Character Recognition). It is an interested algorithm to recognize Myanmar characters that has been developed recently in Myanmar. MICR is easy to implement and it is simply developed without using any pattern matching techniques. MICR is based on the statistical and semantic methods and voting system. Statistical and semantic information includes horizontal black stroke count, vertical black stroke count, width and height ratio, number of loops, histogram values, end points, etc.

A statistical approach to image character recognition would suggest that one look for a typical spatial distribution of the pixel values that characterize each character. In general, one is searching for the statistical characteristics of various characters. These characteristics could be very simple, like the ratio of black pixels to white pixels, or more complex (News & Events White Papers, 2002). Statistical information (eg. Black Stroke Count) for some Myanmar characters are shown in Figure (3).

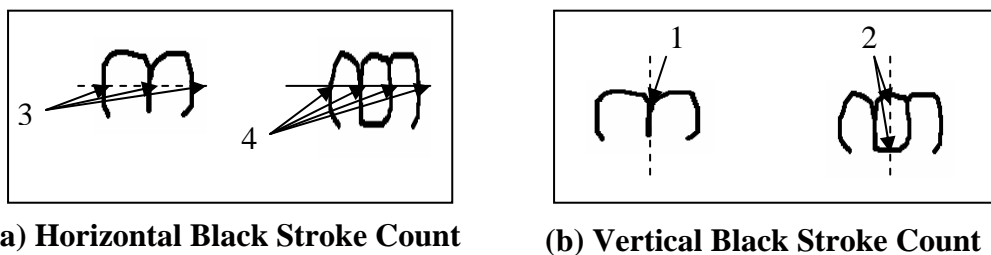


Figure 3 Black Stroke Count

The essential point of the semantic approach to character recognition: first recognize the way in which the contours of the characters are reflected in the pixels that represent them and then try to find typical characteristics or relationships for each character (News & Events White Papers, 2002). The following figure (4) shows semantic information (eg. number of loops and open).

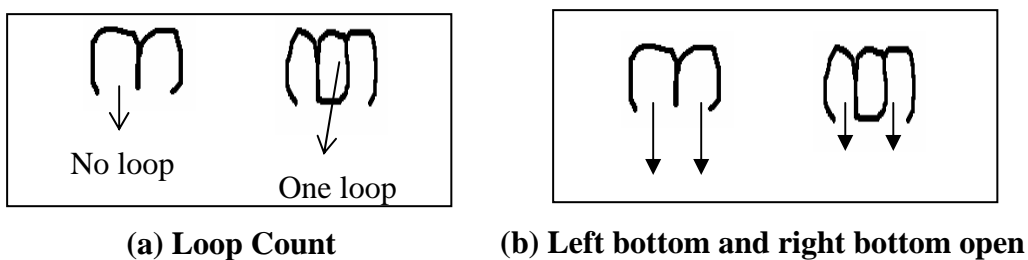


Figure 4 Number of Loops and Open

This statistical and semantic information can be used as various glossary parameters that can be controlled via MICR engine. Data from sub-algorithms are stored in B array (for basic characters) and E array (for extended character). The incoming data are compared with predefined database. If the incoming data matches with predefined database, the voting system produces the specified code number for each character. This code numbers are stored in a code buffer. Otherwise, reject message is produced. MICR engine is run for all characters in an input image. These code values come from MICR are changed into their relative Unicode or ASCII

code and then they are arranged in Code Arrangement stage. In Output stage, the editable Myanmar Compound Words are produced.

5. ALGORITHM FOR PROPOSED SYSTEM

Step 1. Read an image from Tablet device.

Step 2. Preprocessing (image).

Step 3. For all no. of characters :

(a) Find statistical and semantic information.

Type(); /* Assign type (Basic or Extended character)*/

W:H(); /* Width & Height ratio*/

HC(); /* Horizontal Black Stroke*/

VC(); /* Vertical Black Stroke*/

.....

(b) Compare the input array with arrays in predefined database.

(c) The voting system produces code value of recognized characters according to its policies.

Step 4. (a) Code changing ();

(b) Arrangement ();

Step 5. Output the editable texts.

6. EXPERIMENTAL RESULTS AND CONCLUSION

Myanmar Characters	Typeface	Handwritten
Digits	98.75%	97.28%
Consonants	97.86%	95.34%
Extended Characters	96.24%	94.66%
Compound words	95.45%	93.81%

Table 2 Experimental Result of Myanmar characters

This system focuses only on isolated characters. Overlapped characters are intended for future work. If the input image is noise free and not broken characters, the accuracy rate using ICR is higher than OCR. OCR can recognize only typeface characters. ICR can overcome OCR's problems and vice versa. That is the reason why combination of ICR and OCR can achieve the best recognition and accuracy rate. We tend to develop a system combining ICR and OCR in near future.

REFERENCES

Bounnady, K., Kruatrachue, B. and Wangsiripitak, S., 2005. On-line Lao Handwritten Recognition with Proportional Invariant Feature. Proceedings of the World Academy of Science, Engineering and Technology, Volume 5, ISSN 1307-6884.

Bozinovic, R. M. and Srihari, S. N., 1989. Off-line cursive script word recognition. In IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, pp. 68-83.

Fernando, H. C., Kodikara, N. D. and Hewavitharana, S., 2003. A Database for Handwriting Recognition Research in Sinhala Language. Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003).

Htut, Z., 2003. Features of Myanmar Language Document Styles. Executive Committee Member, MCSA, Myanmar Computer Federation (MCF).

Krahnstover, N. and Paulhamus, B., 1999. Development of a Korean OCR System Term Project in CSE 581 – Patterns Recognition. Computer Vision Laboratory, Department of Computer Science and Engineering, Pennsylvania State University, 220 Pond Lab, University Park, PA 168029.

Liu, K., Huang, Y. S. and Suen, C. Y., 1999. Identification of fork points on the skeletons of handwritten Chinese characters. In IEEE Trans. Pattern Anal. Mach. Intell., vol.21, pp. 1095-1100.

News & Events White Papers, 4.10.2002, http://www.tis.co.il/html/news_white_pcr.shtm

Okamoto, M. and Yamamoto, K., 1999. On-line handwriting character recognition using directional change features that consider imaginary strokes. Pattern Recognition, 32, pp. 1115-1128.

Seethalakshmi, R., Sreeranjani, T. R., Balachandar, T., Singh, A., Singh, M., Ratan, R. and Kumar, S., 2005. Optical Character Recognition for printed Tamil text using Unicode. Journal of Zhejiang University Science, ISSN 1009-3095, 6A(11): 1297-1305.

Swe, T. and Tin, P., 2005. Recognition and Translation of the Myanmar Printed Text Based on Hopfield Neural Network. In Proc. of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT 2005), pp. 99-104, Yangon, Myanmar.

Tay, Y. H., Lallican, P. M., Khalid, M., Viard-Gaudin, C. and S. Kneer, 2001. An offline cursive handwritten word recognition system. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, TENCON 2001, Vol. 2, pp. 519-524.

Veltman, S. R. and Prasad, R., 1994. Hidden Markov models applied to on-line isolated character recognition. In IEEE trans. Image Processing, Vol. 3, no. 3, pp. 314-318.