Converting Myanmar Portable Document Format (pdf) to Machine Editable Text with format

Tay Zar Ko Ko and Dr.Yadana Thein
yadanaucsy@gmail.com,tayzarkoko1@gmail.com,

Abstract

This paper proposed a system that can convert Myanmar Portable Document format to machine editable word document with format. It uses Myanmar Intelligent Character Recognition (MICR) to recognize character. MICR is one kind of ICR (Intelligent Character Recognition) system. It is based on statistical and semantic information of the characters. The required statistical and semantic information can be obtained by measuring width and height ratio, black stroke counts, number of loops, open directions, histogram values, etc. The final decision is made by the voting system. MICR has been successfully developed in the following applications such as car license plate recognition system, speed limited road sign recognition system, recognition of vouchers, digit recognizer and online handwritten Myanmar combined words recognition system, etc. The main idea of this paper is to format the page like the original (pdf) document including alignment (left, right, center), Bold, Italic, and underlined color and picture, etc.

1. Introduction

Every language in the world has different and distinct character recognition methods. This class of methods includes statistical methods, artificial neural networks, support vector machines, multiple classifier combination, etc. Character recognition is the most important research in the world of computer science because it acts as a communication medium between the human and computer machines. 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. Intelligent Character Recognition (ICR) is pattern based character recognition and is also known as Hand-Print Recognition. ICR is applied in the acquiring of handwriting images from documents like forms and applications.

In Myanmar, Myanmar characters are more complex than English characters and less complex than Chinese characters. Myanmar script has been developed from the Mon script and adapted from southern Indian Pali script. Although there are many languages in Myanmar, such as Myanmar, Karen, Rakhine, Shan, Mon, Chin, etc. Myanmar is spoken by 32 million as a first language. So, this paper proposed Myanmar characters to recognize and format. Moreover, character formatting is not widely popular in Myanmar Computer Environment. The recognition rates and formatting rates are very excellent in this application.

2. Nature of Myanmar language

Myanmar character is descended from the Mon script and Southern Indian Pali script. In Myanmar, Myanmar language is mostly used. In Myanmar language, Myanmar character has (33) basic consonants, (12) basic vowels, (4) medials, compound words and other extended characters. By combining basic character with medial is become meaningful word. Typically of Myanmar character are round shape and nearly to each other. Patterns of Myanmar characters are shown in the following figures. Myanmar characters are more complicated than English characters. But it is less than Chinese and Japanese characters. Some information is given below.

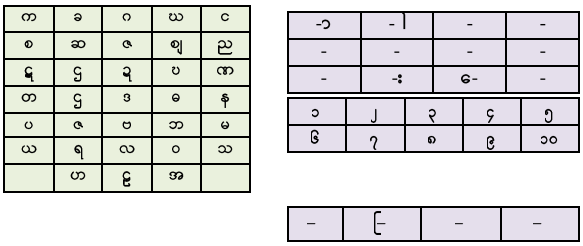


Figure1. Myanmar Character (Consonant, Medial, Vowels and Digits Tables)

3. Previous works of MICR

Myanmar Intelligence Character Recognition (MICR) is successfully developed in the following applications such as

- Car license plate reader
- Myanmar digits recognizer
- Recognition of speed limited road signs
- Recognition of account papers and vouchers

- Handwritten Myanmar combined words recognition system
- Voice production of handwritten Myanmar combined words, etc

4. Motivation for this system

In Myanmar, character recognition methods are popular among the researchers nowadays. Characters are now recognized in many techniques like Optical Character Recognition (OCR), Intelligent Character Recognition (ICR) etc. Successful character recognition applications are described above [3] and there is no application for formatting the page like document. Characters formatting are not seen very much in Myanmar as well as in the world. So, this paper has been proposed. The formatting rate of this application is very high in real.

5. MICR System Architecture

Five main steps are verified in this system: ***input, preprocessing, MICR, Code arrangement, Output.*** First the image acquisition step takes on-line and off-line image from scanner. Then, the acquire images are passed through a preprocessing step including Gray scale converting, Noise filtering, Binarization, Extraction, and Normalization. The third step, MICR engine starts to recognize the input characters of each word. And then the recognize words are pass into code arrangement in which each word is change into Unicode. The changed characters are arranged in format. Finally, a word processing document with formatting is produced.

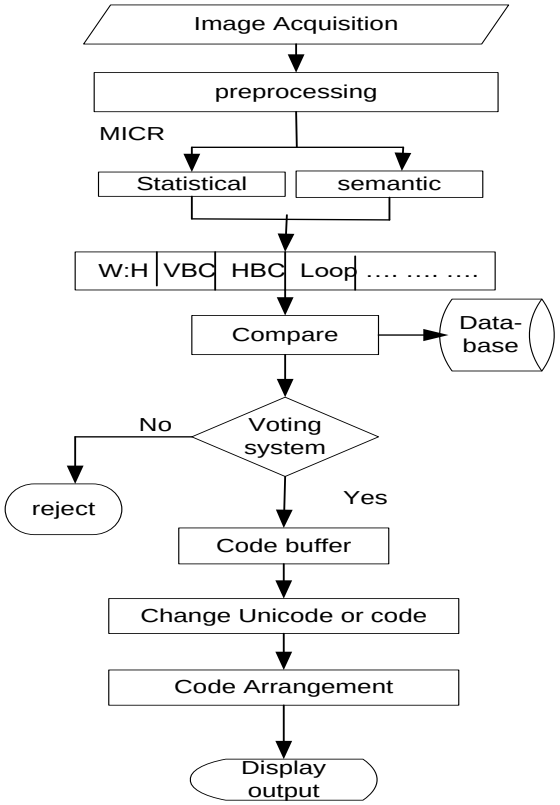


Figure5. Design of the proposed system

6. Myanmar Intelligent Character Recognition (MICR)

Myanmar Intelligent Character Recognition (MICR) is one kind of ICR methods and it can recognize both online and offline characters but it is more suitable for noise free image, not broken and isolated characters. It is an interested algorithm to recognize Myanmar characters that has been developed recently in Myanmar. This system is based on statistical and semantic information from each character and final decision is made by the voting system. MICR can successful overcome problem such as misrecognition Myanmar character 'ခ' with 'ဓ' compared to back propagation neural network.

6.1. Statistical Approach

Since every electronic image of a digit consists of pixel values that are represented by a spatial configuration of "ခ" and "ဓ", a statistical approach to image character recognition would suggest that one look for a typical spatial distribution of the pixel values that characterize each digit. In general, one is searching for the statistical characteristics of various digits. These characteristics could be very simple, like the ratio of black pixels to white pixels, or more

complex, like higher order statistical parameters such as the third moments of the image.

6.2. Semantic Approach

Digitized images of handwritten characters indeed consist of pixels. However, a fact that most statistical methods ignore is that the pixels also form lines and contours. This is the essential point of the semantic approaches to character recognition: first recognize the way in which the contours of the digits are reflected in the pixels that represent them and then try to find typical characteristics or relationships for each digit.

7. Character Recognition and Formatting

Formatting the page is very important in character formatting and need to classify types of fonts, size of fonts, line spacing of the paragraph. Also paragraph formatting includes left alignment, right alignment, and center alignment and justify of the paragraph. This information can be obtained by applying the MICR algorithm, statistical and semantic information. But certain information is difficult to obtain like font size of the character because there are many font sizes that can be used. Also font styles of character are not similar in Myanmar language. So, this information is unable to find easily.

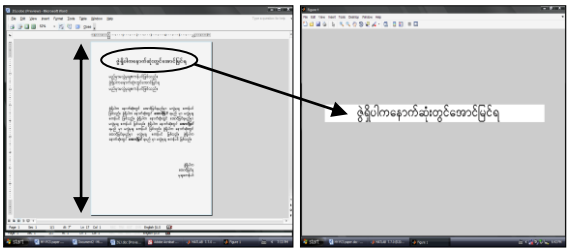
But applying the MICR algorithm, nearly equal values of such information are obtained. However, some limitations are still remained because of the complexity of the character and font sizes. In Myanmar language, some basic character and extended character are connected in some types of fonts and Unicode such as (‘ကာ’,’ကြ’). But some are not connected in types of fonts such as (–Win—Researcher).

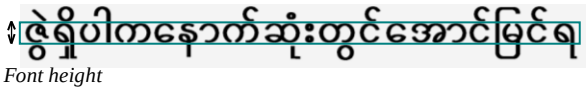
This paper proposed a system that can convert Myanmar Portable Document Format to machine editable word document including various sizes of the character, kinds of fonts, line spacing between two lines and so on. It also calculates information for bold character, left alignment, right alignment, center alignment, left and right justify of the page. In analyzing the height of the character, various sizes of the character can be calculated and can be converted to Rich Text Format. In calculating the line spacing, it is necessary to calculate the vertical space count between the two rows of the sentence. Recognizing and analyzing the bold character is simple by using statistical and semantic information. This information can be obtained by calculating pixel counts in one horizontal or vertical black stroke.

7.1 Some Information for Formatting

7.1.1 Font Size

The required information for font size of a character can be obtained by calculating the number of pixel counts from the height of a basic character. First, extracting a sentence from the page and then extract a character form the sentence. The font size of a character can be obtained by the following way.





Font height

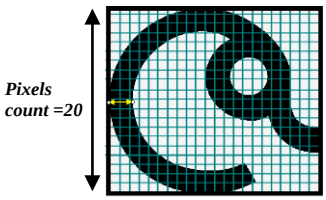


Figure6. Acquiring the height of a character

7.1.2 Bold

By calculating the pixel count in a black stroke of a character, the information for bold or normal character can be classified.

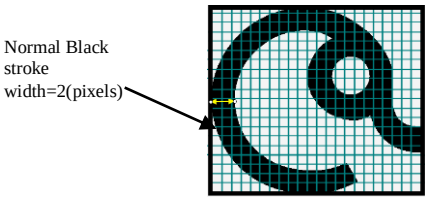


Figure7. Pixel information for bold or normal

7.1.3 Alignment

The alignment style of the paragraph can be calculated by space count between sentence and margin and center point. By comparing left edge point, right edge point and midpoint of a sentence, the required information for paragraph alignment can be concluded.

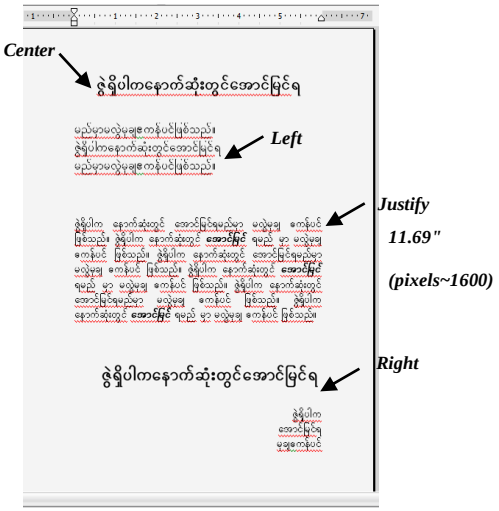


Figure8. Example of Alignment form of a page

8. Experimental Results

The accuracy rates and processing time, that have been discovered in applying proposed system is shown in the following table.

Number of characters	Processing time	Accuracy Rates
1000	10s	95%
2000	20s	95%
5000	50s	94%
10000	1.3min	90%

Table1. Practical Experimental results

9. Conclusion

Converting Portable Document Format to Machine Editable Text is a new contribution for Myanmar Computer Environment. Some limitations are still included in this system such as limitation from PDF, limitation from characters, etc. But it can recognize Myanmar Typefaces in 100% and can format left alignment, right alignment, center alignment and justify of the paragraph. It also format bold and italic characters and can produce various font sizes for character. Some extensions are still remaining to find out and to classify.

There are some difficulties in typeface character recognition because some basic characters are connected with extended character or medials characters. Because of this, there are some difficulties in recognizing all kinds of font styles. Like say in above, some character such as basic character 'က' is connected with 'ာ' in 'ကာ' and 'က' is connected with medial ' ' in 'ကျ' in (-Win---Innwa) and also in the Unicode fonts.

10. References

[1] Ei Ei Phyu, Zar Chi Aye, Ei Phyu Khaing, Yadand Thein and Myint Myint Sein, "Recognition of Myanmar Handwritten Compound Words based on MICR", the 29th Asian Conference on Remote Sensing (ACRS), Colombo, Sri Lanka, 2008

[2] Zar Chi Aye, Ei Ei Phyu, Yadana Thein and Myint Myint Sein, "INTELLIGENT CHARACTER RECOGNITION (MICR) AND MYANMAR VOICE MIXER (MVM) SYSTEM", the 29th Asian Conference on Remote Sensing (ACRS), Colombo, Sri Lanka, 2008.

[3] 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.

[4] chavdhuri, B. B., Pal, U. And Mitra, M., "Automatic Recognition of Printed Oriya Script", Sadhana, 2002, Vol. 27, Part I

[5] R. K, Rajapakse, A. R. Weerasinghe and E. K.Seneviratne, "A Neural Network Based Character Recognition System for Sinhala Script," South East Asian Regonial Computer Confederation, Conference and Cyberexhibition *(SEARCC'96), Bangkok, Thailand, July 4-7th 1996.*

[6] LI Guo-hong, SHI Peng-fei.2003. An approach to offline handwritten Chinese character recognition based on segment evaluation of adaptive duration, ISSN 1009-3095

[7] Nafiz, A., Fatos, T.Y., 2002. Optical character recognition for cursive handwriting. IEEE Trans. On Pattern Recognition and Machine Intelligence, 24(6):801-813

[8] LI Guo-hong, SHI Peng-fei.2003. An approach to offline handwritten Chinese character recognition based on segment evaluation of adaptive duration, ISSN 1009-3095