Myanmar Portable Document Format Recognition with Formatting

Cherry Maung, Dr. Yadana Thein
University of Computer Studies Yangon, Myanmar
c.maung2@gmail.com, *yadanaucsy@gmail.com*

Abstract

This paper contributes Myanmar Printed Character Recognition with format. This system consist recognition and formatting. It recognizes for Myanmar Portable Document Format (.pdf) such as font size, font style, alignment and table. It converts the existing document to Machine Editable Word Document (.doc). It contains paragraph and table classification. In table classification, table recognition and formatting can also be performed. The extraction of text format, paragraph format and table format can be done efficiently. The system is based on MICR (Myanmar Intelligent Character Recognition) which is a kind of ICR (Intelligent Character Recognition). MICR uses statistical and semantic information which includes width and height ratio, black stroke counts, number of loops, open directions and histogram value, etc. The final decision is made by the voting system. The system use image processing and Matlab programming.

1. Introduction

The character recognition has been one of the most interesting and important fields in research world because it is a kind of communication medium between the human and computer machines. Several different methods such as artificial neural networks, multiple classifier combination, support vector machine and statistical methods have been used to recognize characters.

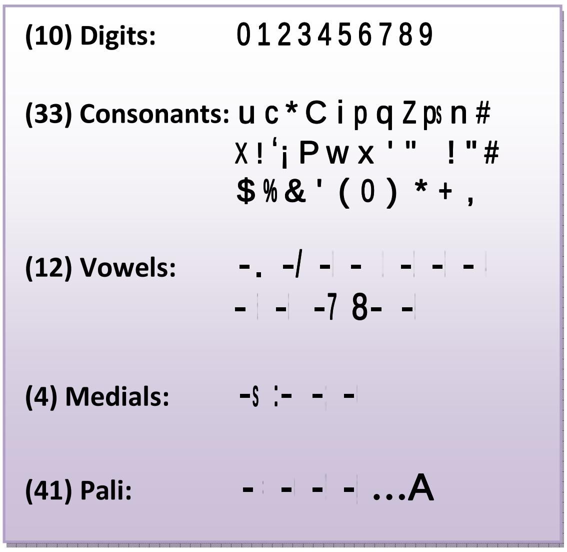
Two main methods of Character Recognition are OCR (Optical Character Recognition) and ICR (Intelligent Character Recognition). OCR is a process of converting images of characters to ASCII data or machine readable characters. The disadvantage of OCR is that its misrecognition of similar pattern. ICR, the pattern based method, has the ability to turn images of handwritten or printed characters into ASCII data. ICR technique is very convenient in character recognition.

There are many languages in Myanmar such as Kachin, Kaya, Kayin, Chin, Mon, Myanmar, Rakhine and Shan, etc. Among them Myanmar language is the most commonly used. According to international language family tree, Myanmar language is a member of Sino-Tibetan language family.

Most of the Myanmar characters are round in shape. Myanmar characters are more complex than English alphabets and less complex than Chinese characters. However, software developers considered Myanmar script as a complex script. In this paper, Myanmar character recognition with related formatting will be presented. Moreover, Myanmar character formatting which is not widely popular in Myanmar computer environment will be processed with excellent recognition and formatting rate in this application.

2. Myanmar Language Characteristics

Myanmar language consists of (10) digits, (33) basic characters, (12) vowels, (4) medial, (41) pali characters and other extended characters.



(10) Digits:	0 1 2 3 4 5 6 7 8 9
(33) Consonants:	u c * C i p q Z ps n # X ! ' ¡ P w x ' " ! " # $ % & ' (0) * + ,
(12) Vowels:	-. -/ - - - - - - - -7 8- -
(4) Medials:	-s :- - -
(41) Pali:	- - - - ...A

Figure 1. Myanmar language characteristics

3. Portable Document Format

Image acquisition is completed either by on-line or off-line technique. On-line image acquisition becomes a popular area. The data input is carried out by means of Tablet or PDA (personal data assistant). Off-line image acquisition is done by scanner.

The input of this system is Portable Document Format (.pdf) which is self-contained cross document format. It allows the reliable reproduction of published material on many different platforms. In this system, pdf format of A4 standard paper size is converted to Joint Photographic Expert Group (.jpg) with PDFCreator 0.9.9.

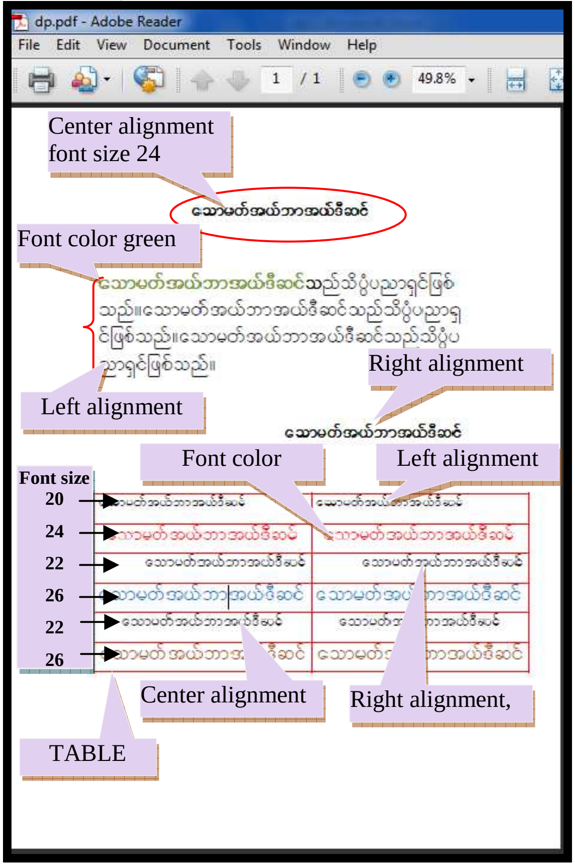


Figure 2. Portable document format with formatting

4. Preprocessing

Preprocessing stage includes grayscale converting, binarization and normalization and feature extraction. Gray-scale converting is converting RGB image to gray-scale. Gray-scale eliminated hue and saturation. Binarization is binarizing the image to 0 or 1. If a pixel has data, it will be represented with 0, otherwise 1.

5. Myanmar Intelligent Character Recognition (MICR) Architecture

There are two kinds of character recognition OCR and ICR. OCR can recognize continuous or disjoint characters. It is rule-based system. Thus, misrecognition of character pattern can be occurred. ICR is the pattern-based recognition system and so can recognize isolated characters.

Myanmar intelligent character recognition (MICR) is one kind of ICR. The input of this system is isolated characters. It is vital to perform preprocessing stage for MICR. Statistic and Semantic information is acquired after preprocessing stage. MICR used statistic and semantic information. The information of each character is compared with the information in the predefined database. Final decision is made by voting system. The outcome of MICR is related code which is stored in the code buffer.

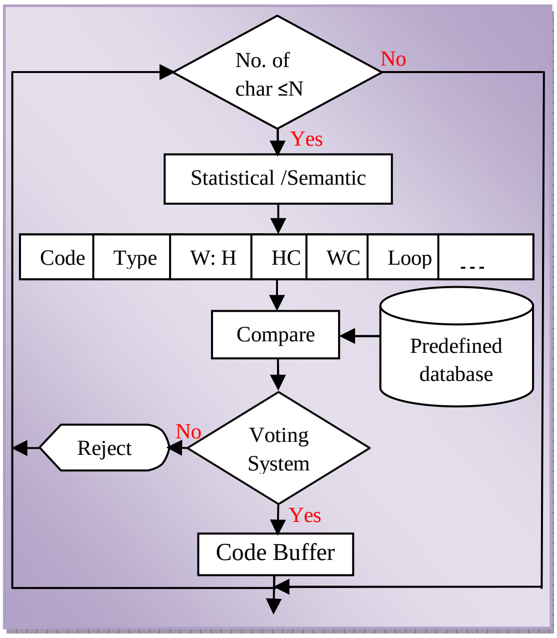


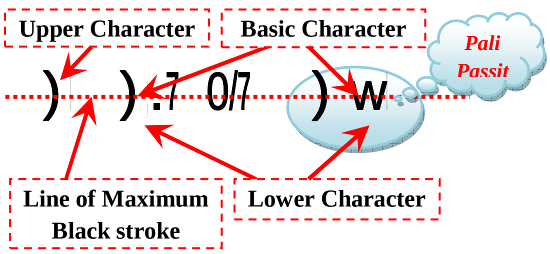
Figure 3. MICR system architecture

5.1. Statistical and Semantic Information

The statistical information of typical spatial distribution of the pixel values in image can be recognized. In semantic information, some of the pixels in the image may be formed in lines or contour. Statistical and semantic information includes the ratio of black pixel to white pixel, histogram value and pixel density, black stroke count, loop count and open position.

The position of each character is also necessary both for Myanmar and Pali characters. It is determined by the line of maximum black stroke.

Position of Character (P)



6. System Design

The input of this system is Portable Document Format which is converted to Joint

Photographic Expert Group (.jpg). After input stage, preprocessing stage takes place. After preprocessing stage, classification of table from the background image is carried out. If it is table, sorting lines, defining rows and columns and extracting cells take place. If it is not table, each row (sentence) from the paragraph is extracted. Then, character extraction of each cell and each row (sentence) is performed. The format of each character is extracted and the information of each character is assigned to MICR. The related code produced from MICR is converted to ASCII code or Unicode. The features of text format and table format are applied in format stage. The output of the system is Editable Word Document.

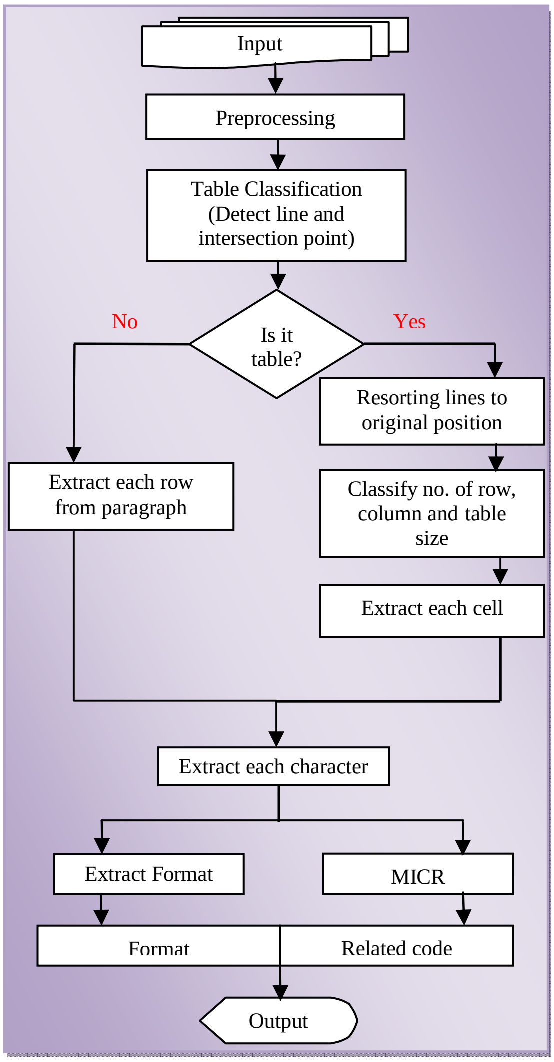

Input
Preprocessing
Table Classification (Detect line and intersection point)
No
Is it table?
Yes
Resorting lines to original position
Extract each row from paragraph
Classify no. of row, column and table size
Extract each cell
Extract each character
Extract Format
MICR
Format
Related code
Output

Figure 4. Architecture of the proposed system

7. Table Classification

It is necessary to detect lines and intersection points in order to classify table from image. Threshold is also important for table classification because the image may include other lines. Therefore, minimum length and pixel gap are assigned as threshold value.

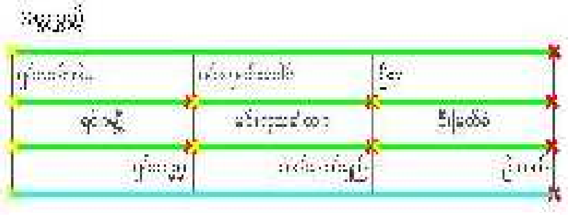
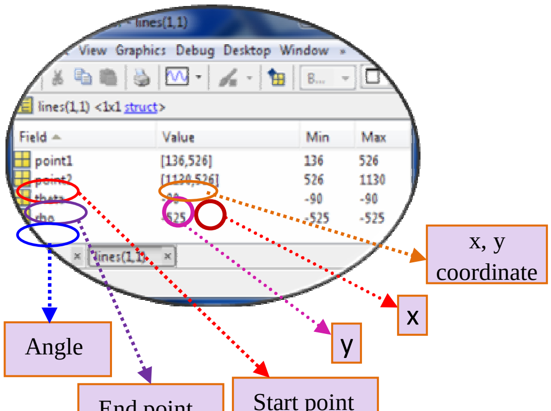


Figure 5. After line detection



lines(1,1) <1x1 struct>

Field	Value	Min	Max
point1	[136,526]	136	526
point2	[1130,526]	526	1130
theta	-90	-90	-90
rho	-525	-525	-525

Figure 6. Line information

7.1. Resorting lines to original position

If the image has line segments and intersection points, it can be assumed as table; otherwise, it is a background paragraph. After line detection, all the lines information is stored in a lines array. Line information includes start point, end point, rho and angle. Start point and end point are presented with their respective coordinates (x, y) value. All these lines are sorted according to their row (x) position. If a line with greater (x) position, that line will be further from the top of the image.

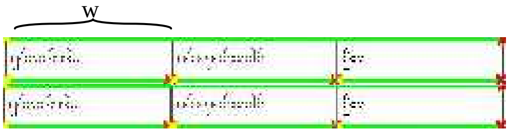
7.2. Classify row, column and table size

If the (y) coordinate of second line is equal to the y coordinate of first line increase row count. It is necessary to perform the same procedure for other lines. The total number of row is row count minus one. In the first row, if the x coordinate of other lines are equal to the x coordinate of the second line, increase the column count.

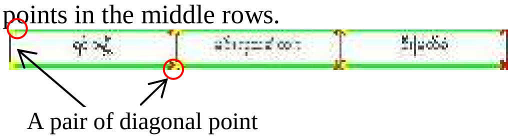
The width and height of each rectangle is measured by inches. Therefore, the unit of the image is required to convert. The image size (1240x1754) is converted with the standard A4 paper size (8.27"x11.69") then the equation becomes (150px=1inch).

7.3. Extract each cell

Cell extraction consists of three phases. The first phase include the first row is cropped with the (x, y) position of start point of the first line and (x, y) position of end point of second line. Then, add width to the start point of the first line. The other rectangles are cropped with this procedure.

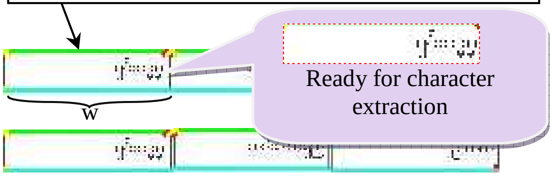


In the second phase, the cells are cropped at diagonal points in the middle rows.



In the final phase, the final row will be cropped at the (x, y) position of start point of L column and y position of end point of L, (x) position of start point of the last line.

Upper line of final row first column, L = Total no. of line, (Tl) – no. of column



Then, add one to L in order to process the next rectangle. The remaining rectangles are cropped with the same procedure as the previous one.

It is vital to extract all rectangles without boundaries so that they will be ready for character extraction.

8. Extract each row from paragraph

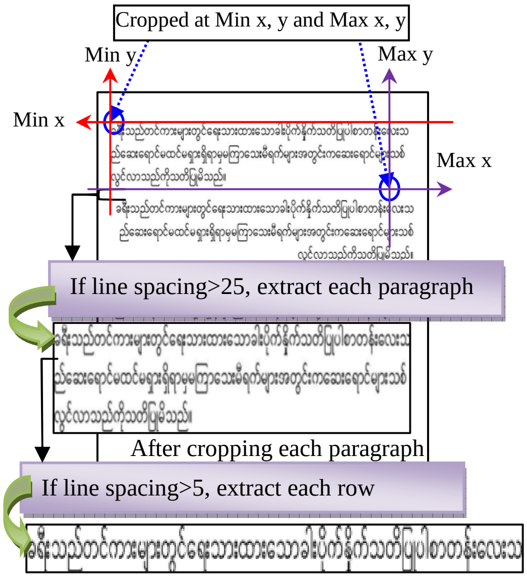

Cropped at Min x, y and Max x, y
Min y
Max y
Min x
Max x
If line spacing>25, extract each paragraph
After cropping each paragraph
If line spacing>5, extract each row

Figure 7. Row exctraction

9. Extract each character

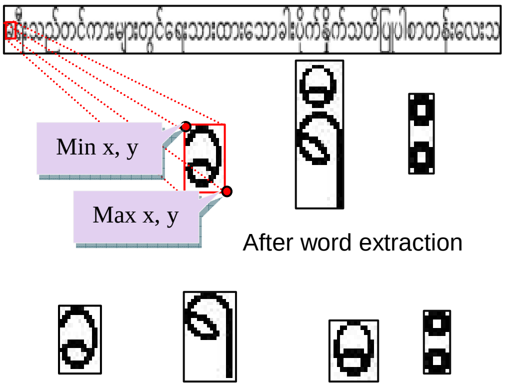


Figure 8. Character extraction

10. Myanmar Intelligent Character Recognition (MICR)

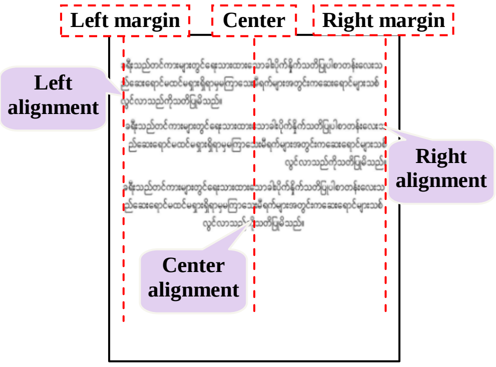
The input of MICR includes isolated characters. The information for MICR can be achieved by statistical approach or semantic approach. The data of statistical or semantic information include width and height ratio, black stroke count and loop and their position, etc. The resulting information is compared with the data in the predefined database. There are three types of predefined database; (i) Basic database, (ii) Vowels database, (iii) Medial database. The final decision is made by voting system. The output of the voting system includes related code. This code is put into the code buffer.

11. Extract Format

Format extraction includes alignment, bold, etc.

11.1. Alignment

Minimum x is the left margin. Maximum x is the right margin. The center point, c= (max x-min x)/2+min x. If the paragraph is near to left margin, it is left alignment or near to right margin right alignment, otherwise center alignment.



each row

Figure 9. Alignment

11.2. Font size

The recognition of font size depends on the height of its basic consonant characters. Basic characters can be divided into two groups for font size. The first group contains one row characters (w, !,), u, i, & ,etc.) and the second group includes two rows characters(,',! ,# , +, X, C, n).

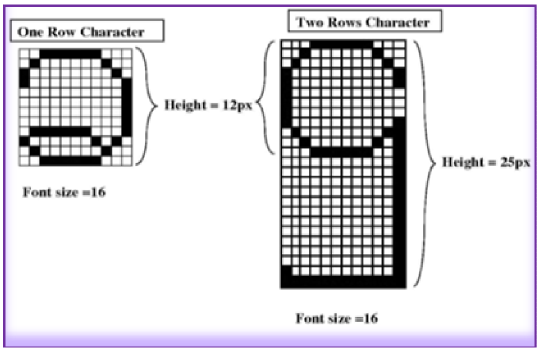


Figure 10. Font size

Table 1. Font size of the character and their related height

Font size	Height of first group	Height of second group
16	12	11/12
18	13	13
20	14	14/15
22	15	16
24	17	17/18
26	18	19
28	20	20/21
36	25	27
48	34/35	36
72	52/53	54

11.3.Font Colour

The recognition of font color can be done by converting RGB to indexed image. Dither option can cause noise and distortion of an image. Therefore, ‘no-dither’ option is needed to be set. After converting RGB to indexed image, a single RGB (red, green and blue) colour code can be easily acquired.

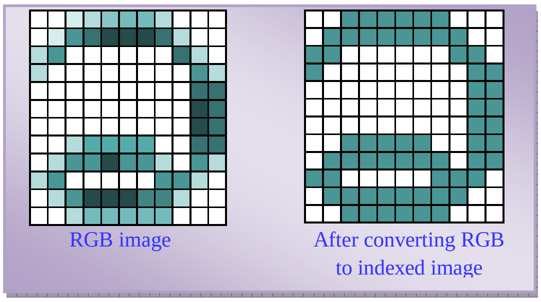


Figure 11. Font colour

11.4.Bold

The recognition of bold depends on the font size and the pixel count in each black stroke of a character. The pixel count in each black stroke of bold character is greater than normal character. If the character is bold, the bold flag will be one. Otherwise, the bold flag will be zero.

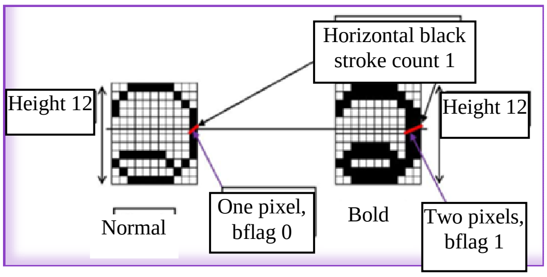


Figure 12. Bold

Table 2. Pixel information of normal and bold character

Font size	Pixels of Normal	Pixels of Bold
16	1px	2px
18	2px	3px
20	2px	3px
22	2px	3px
24	2px	3px
26	2px	3px
28	3px	4px
36	3px	4px
48	4px	5px
72	7px	8px

12. Format and Related Code

The format that has been extracted is exported to Editable Word Document with string array. The font color is stored in the string array first. Then, paragraph alignment is assigned to the string array. Then, font size and bold are added to the array.

MICR produces related code in the code buffer. These codes are changed to UNICODE or ASCII code and then append to the string array.

The whole string is transferred to the word document. The output of the system is Editable Word Document.

13. Rejection Criteria

In Table Recognition : If the line weight of table boundary is less than 1pt, Hough Transformation cannot recognize all line segments. If the line weight is greater than 3pt, Hough Transformation recognizes extra line segments. Double border line cannot be recognized.

Bold : If the font size is 24, its height is 18. When it is bold its height become 19. If the font size is 26, its height is 19. Therefore a character of font size 24 with bold, it become font size 26 with bold.

Color : If the color is very soft, the character will be misrecognized.

Font size : In Myanmar character, the smallest font size is 16. If the font size is smaller than 16, there will be noise in it.

14. Experimental Result

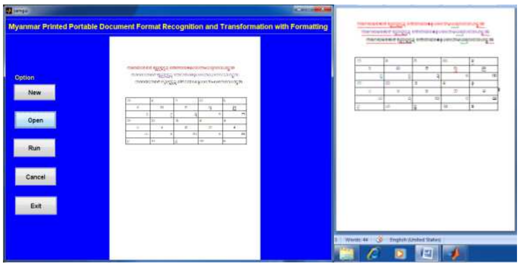
r= row; c= column; W =width; H= height; T =time; B=bold; FS = font size; AR= accuracy rate; px=pixels;

Table 3. Experimental Result of table format

TABLE					
(r, c)	W (px)	AR	H (px)	AR	T
3x3	324	98.56%	70	99.94%	11s
4x2	465.5 506.3	98.42%	50.6	99.95%	12s
7x5	184.8 195.6	98.36%	44.6 46.1	99.1%	13s
21x5	187.1 197.9	98.27%	46.1 47.6	99 %	15s

Table 4. Experimental Result of character recognition with font size and bold

FS	Character Samples	AR (without B)	AR (with B)
20	100	97.21%	97.13%
	200	97.16%	96.89%
24	100	97.46%	97.37%
	200	97.32%	97.23%
28	100	97.73%	97.64%
	200	97.53%	97.49%
48	100	98.45%	98.31%
	200	97.87%	97.96%


Myanmar Printed Portable Document Format Recognition and Transformation with Formatting
Option
New
Open
Run
Cancel
Exit

Input, Before Recognition **Output, Editable Word Document**

Figure 13. Final result of the proposed system

15. Conclusion

In conclusion, the main contribution of this system is the recognition of table format with Myanmar character in (.pdf) document. There is no research work with table format in Myanmar. This system produces high accuracy rate for table format such as width and height, boundary color and etc. This system cannot recognize table border with double line. It can also extract other format; paragraph format and text format. The accuracy rate of character recognition depends on MICR. MICR cannot recognize broken or continuous character. It recognizes isolated character with high accuracy rate. MICR depends on font size. The bigger font size can produce the higher accuracy rate for each character. The accuracy rate for bold depends on PDFCreator. Although there is minor error, this system can produce the nearest value of the input image.

16. Reference

[1]Dipti Deodhare, NNR Ranga Suri, R.Amit, "Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System", International Journal of Computer Science & Applications Vol. 2, No. 2, pp. 131-144, © 2005 Techno mathematics Research Foundation.

[2]Tay Zar Ko Ko and Dr.Yadana Thein, "Converting Myanmar Portable Document Format (.pdf) to Machine Editable Text with format",

[3] Ei Ei Phyu, Zar Chi Aye, Ei Phyu Khaing, Yadana 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

[4] 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.

[5] 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.

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

[7]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.*

[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