Mixed-Myanmar and English Character Recognition with Formatting

Dr. Yadana Thein, Cherry Maung

University of Computer Studies Yangon (U.C.S.Y)
Yangon, Myanmar

University of Computer Studies Yangon (U.C.S.Y)
Yangon, Myanmar

Abstract

This paper proposed Myanmar and English typeface Character Recognition with their related format. The system converts Portable Document Format (.pdf) to Machine Editable Word Document (.doc). It includes two parts; recognition and formatting. The recognition of Myanmar and English character can be done by MICR (Myanmar Intelligent Character Recognition) which is one kind of ICR. Statistical and semantic information is used in MICR. Final decision of is made by voting system. MICR has become successful in character recognition area recent years. MICR can produce character recognition with high accuracy rate and faster speed. Table classification is used for the recognition of table format. Hough Transformation is used to detect lines in table recognition. This system can perform not only paragraph format but also text format. Paragraph format includes alignment (left, right and center). Text format includes font color, font size and bold, etc. The system use image processing and Matlab programming.

Keywords: *Character Recognition, MICR, Table Forma, Hough Transformation, Text Format, Paragraph Format.*

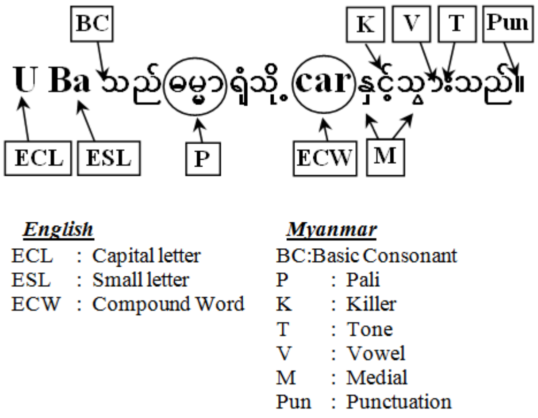
1. Introduction

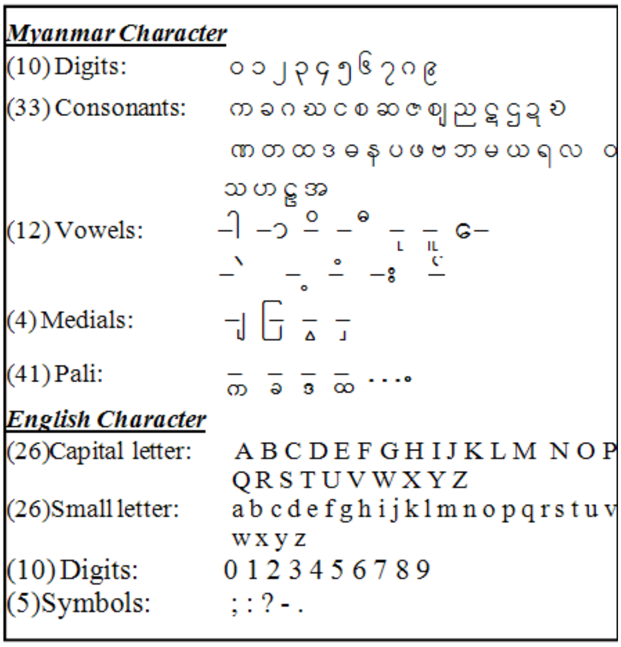
The character recognition has been one of the most interesting and important fields in research world because it is a kind of communication medium between the human and computer machines. Several different methods such as artificial neural networks, multiple classifier combination, support vector machine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

There are more than 30,000 languages all over the world. Among them English is an international language. Even some of the Myanmar words are adopted from English language. Therefore, it is necessary to recognize both English and Myanmar characters. Myanmar language includes Kachin, 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. Most of the Myanmar characters are round in shape. Myanmar characters are more complex than English alphabets and less complex than Chinese character. Software developers considered Myanmar script as a complex script. Myanmar and English character combination with formatting which is not widely popular in Myanmar computer environment will be presented. This system can produce excellent recognition and formatting rate with faster speed.

2. Myanmar and English Language Characteristic

Myanmar language includes (10) digits, (33) basic characters, (12) vowels, (4) medial and other extended characters. English language consists of (26) English Capital letter and (26) small letter.





Myanmar Character	
(10) Digits:	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉
(33) Consonants:	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
(12) Vowels:	ါ ာ ိ ီ ု ူ ေ ဲ ့ ံ း ်
(4) Medials:	ျ ြ ွ ှ
(41) Pali:	္တ ္ခ ္ဒ ္ထ . . . .
English Character	
(26)Capital letter:	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
(26)Small letter:	a b c d e f g h i j k l m n o p q r s t u v w x y z
(10) Digits:	0 1 2 3 4 5 6 7 8 9
(5)Symbols:	; : ? - .

Fig. 1 Characteristics of English and Myanmar languages

3. Motivation and Previous Work of MICR

In Myanmar character research area, only a few works for Myanmar character recognition have been studied. It is still in research because the existing works are not complete enough. The recognition of mixed Myanmar and English characters with their related format is rarely found in research environment. The recognition of characters can be done by using MICR. MICR has successfully developed in the following applications.

- Car license plate reader
- Myanmar digits recognizer
- Recognition of speed limited road signs
- Recognition of account papers and vouchers
- On-line Handwritten Myanmar combined words recognition system
- Voice production of handwritten Myanmar combined words, etc.

4. System Design

The input of this system is Portable Document Format which is converted to Joint Photographic Expert Group (.jpg). Image acquisition can be done by on-line or off-line technique. After image acquisition, the classification of table from the background image is carried out. If the interesting region is inside the table, table format recognition and cell extraction is performed. Table format recognition consists of the recognition of table properties such as table border color, the number of rows and columns, width and height of each cell. If the interesting region or passage is not inside the table, extract the passage outside the table. Moreover, the extraction of paragraph, row and character from both inside and outside of the table takes place. Furthermore, document format recognition is performed. The extracted characters are recognized by using MICR. MICR uses statistical and semantic information to make final decision. It produces related code in the code buffer. The UNICODE or ASCII codes are arranged in the code arrangement stage and related code of each character is assigned to the Word Document. The document format; table format, paragraph format and text format that have been recognized is applied to the Editable Word Document. The output of this system is in the form of Editable Word Document.

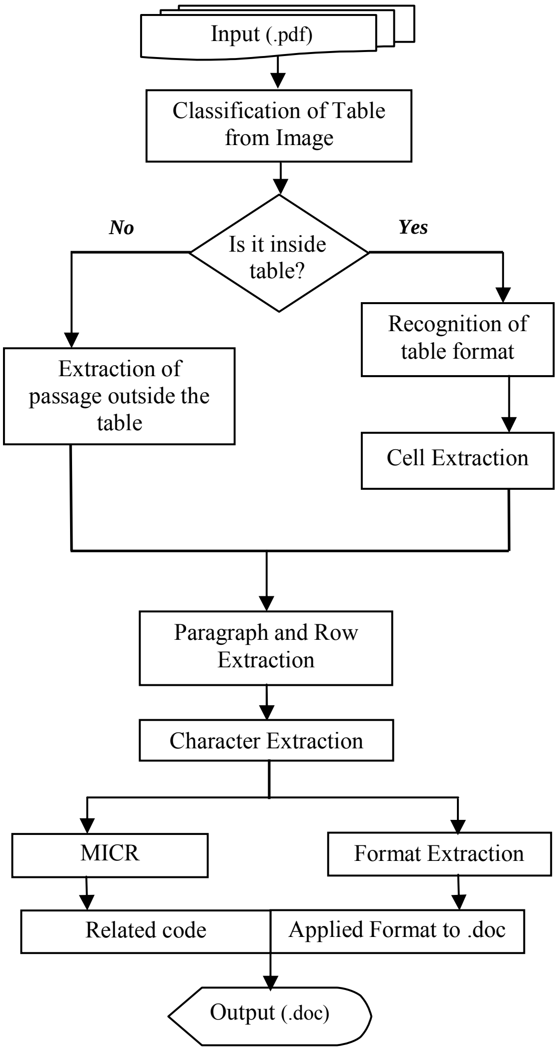


Fig. 2 Proposed system design

5. Portable Document Format (.pdf)

Image acquisition is completed either by on-line or off-line technique. In on-line image acquisition 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 highly compressed and reliable reproduction of published material. In this system, (.pdf) document of A_4 standard paper size with document format is converted to Joint Photographic Expert Group (.jpg) with PDFCreator 0.9.9.

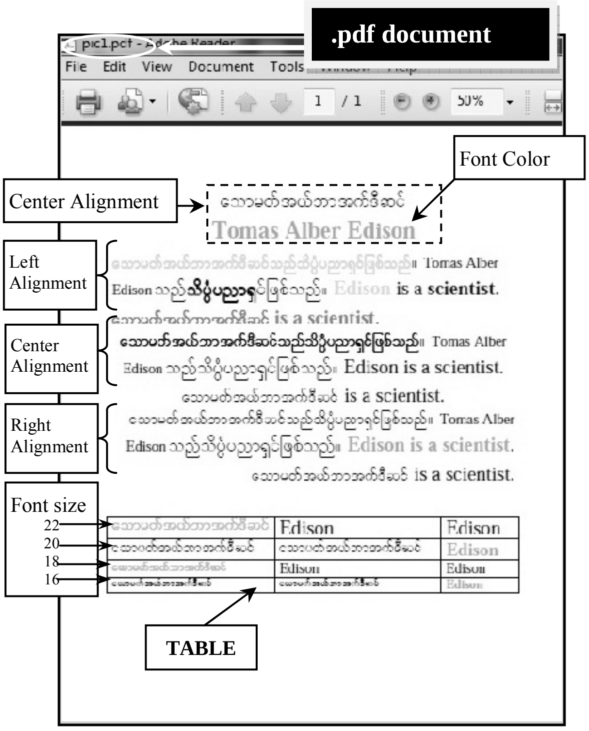


Fig. 3 Input document (.pdf) with formatting

6. Classification of Table from Image

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.

6.1 Hough Transformation

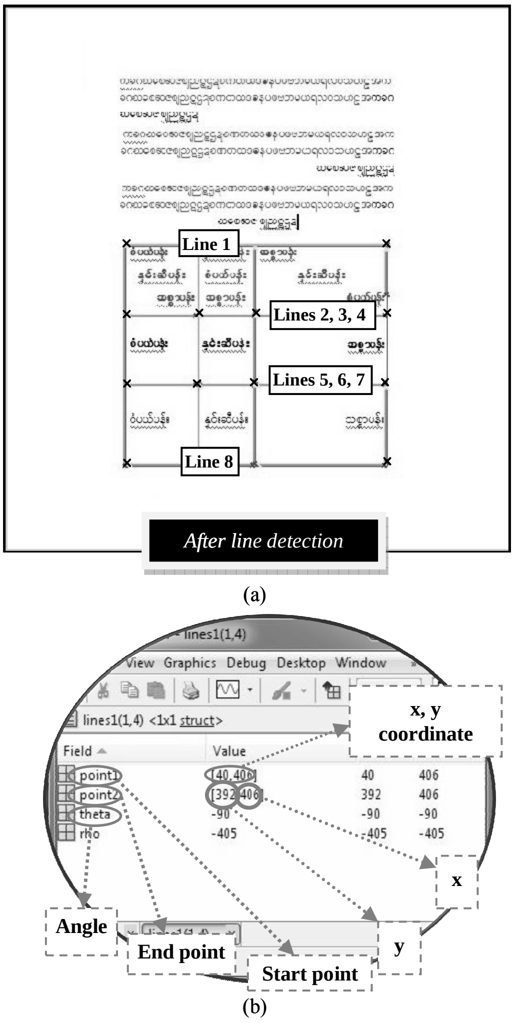
Hough Transformation is essential to detect line segments from table. It uses evidence gathering approach. All the collinear points in line detection are stored in an accumulator array. The major advantage is that it can deliver the same result as template matching, but faster. Two methods of line detection are:

- ❖ Cartesian parameterization

$$y=mx + c; \qquad (1)$$

- ❖ Polar parameterization

$$p=x\cos(ø) + y\sin(ø); \qquad (2)$$



Field	Value		
point1	[40,406]	40	406
point2	[392,406]	392	406
theta	-90	-90	-90
rho	-405	-405	-405

(a)

(b)

Fig. 4 Information of each line from the table

After line detection, all the lines information are stored in an array structure. The information of each line consists of x,y coordinates of starting and ending points, angle and rho. The starting points and ending points are represented by cross illustrated in Fig.4 (a).

6.2 Recognition of table format

The total number of rows can be acquired by increasing row count when the y coordinates of the first line is equal to the y coordinates of other line.

The total number of columns can be achieved by adding one to column count each time when the x coordinates of the first line is equal to other lines in the first row.

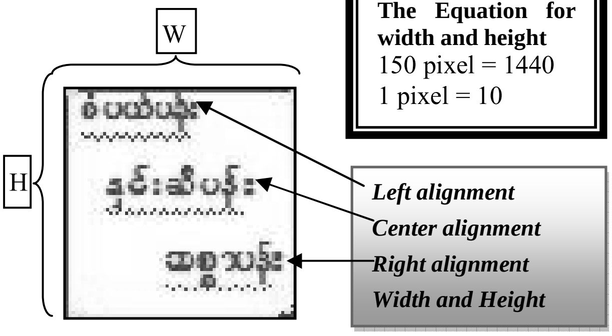
The width and height of each rectangle is measured by inches. The image size (1240x1754) is converted with the standard A4 paper size (8.27"x11.69"), then the equation becomes (150px=1inch).

This system can recognize other table format such as table border color, paragraph alignment and text format for each character within each cell. The border color can be recognized by indexing rgb color of detected line. For paragraph alignment, if the paragraph is near to the left side of the cell, it is left alignment. If the paragraph is near to the right side of the cell, it is right alignment. Otherwise, the paragraph of within the cell is center alignment. Text format inside the table is the same as that of outside the table and it will be fully explained in other section.

For instance,
W = 324 pixel
In Word Document
W = 324 pixel * 10 = 3240
H = 70 pixel * 10 = 700

Equations Derivation

From pixel to inch
150 pixel = 1 inch
In Word Document
1inch = 1440
The Equation for width and height
150 pixel = 1440
1 pixel = 10



6.3 Cell Extraction

The first line contains only one starting point for all three columns. Therefore, add width of the lower first line of first row to the y coordinate of the upper first line of first row. Perform the same procedure for the third column.

A= starting point of upper first line of first row;

$$A1=A+W; \quad (3)$$

$$A2=A1+W1; \quad (4)$$

The last line also contains one starting points for three column. The upper first line of the last row can be achieved by subtracting the number of column from total number of lines. Next, add the width of upper first line last row to the y coordinate of the last line. The same procedure is carried out to find the ending point of third column of the last line.

B= starting point of last line;

$$B1=B+W2; \quad (5)$$

$$B2=B1+W3; \quad (6)$$

Then, each cell is extracted by cropping at diagonal points shown in Fig. 5; the starting point of first line and ending point of the lower first line of first row. The same procedure is performed for the extraction of other cells. It is vital to extract each cell without border line.

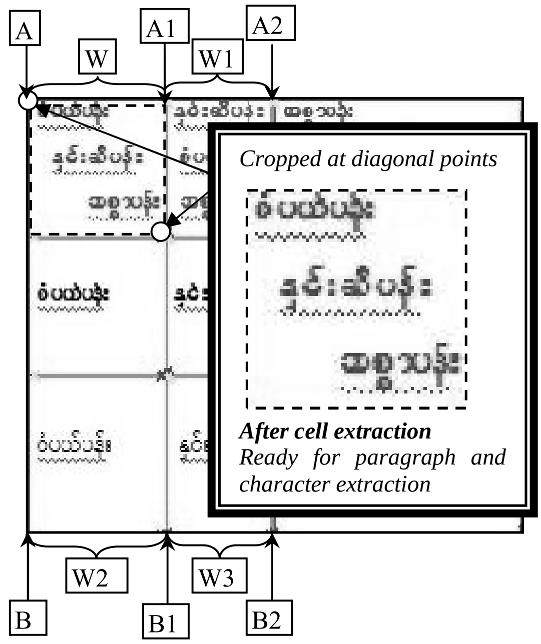

A
A1
A2
W
W1
Cropped at diagonal points
After cell extraction
Ready for paragraph and character extraction
W2
W3
B
B1
B2

Fig. 5 Cell extraction from table

7. Extraction of passage outside the table

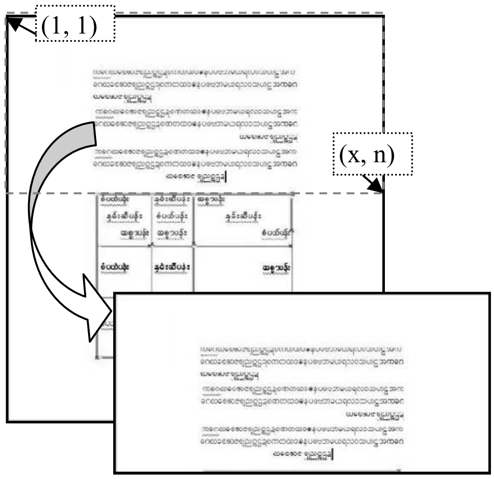


Fig. 6 Extraction of passage outside the table

Let (m, n) be the size of the image. The passage outside the table is extracted by cropping at the point of (1, 1) coordinate and x coordinate of the first line of the table and n.

8. Paragraph Extraction

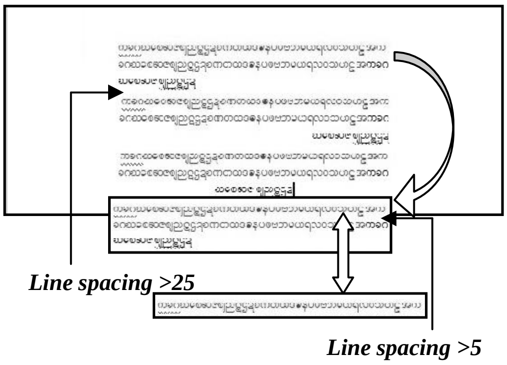


Fig. 7 Paragraph and row extraction

There are two kinds of paragraph extraction. They are extraction of paragraph outside the table and extraction of paragraph inside each cell. If the line spacing is greater than 25, paragraph extraction takes place. If the line spacing is greater than 5, row extraction is carried out.

9. Character Extraction

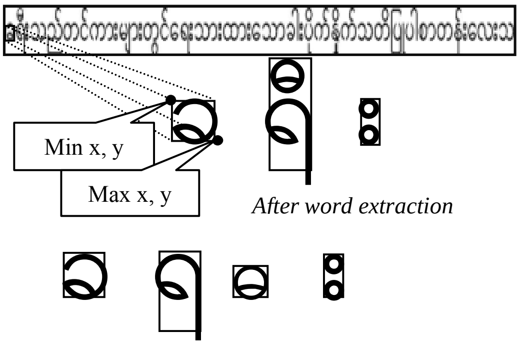


Fig. 8 Character extraction

Character Extraction is done by using labeling method.

10. MICR (Myanmar Intelligent Character Recognition)

Myanmar intelligent character recognition (MICR) is one kind of ICR. The input of MICR includes 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 resulting statistical or semantic 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 relevant code. This code is put into the code buffer. MICR has potential of improving efficiency in the recognition of characters.

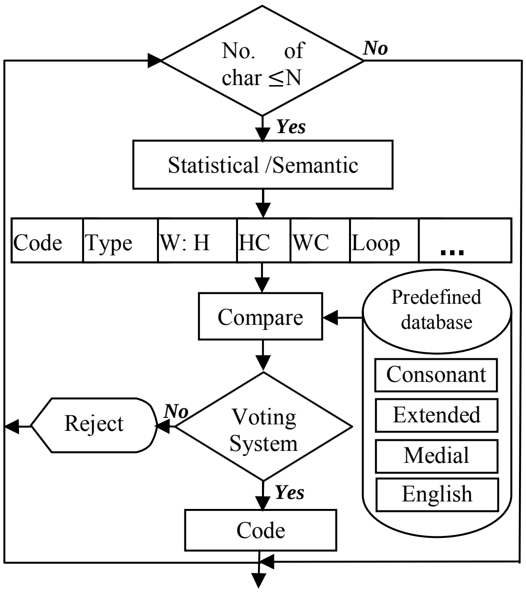


Fig. 9 MICR system flow chart

Table 1: Recognition of similar patterns

Myanmar characters	*English Capital characters*
က	**M**
င	**C**
ဝ	**O**
ပ	**U**

10.1 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 of each character.

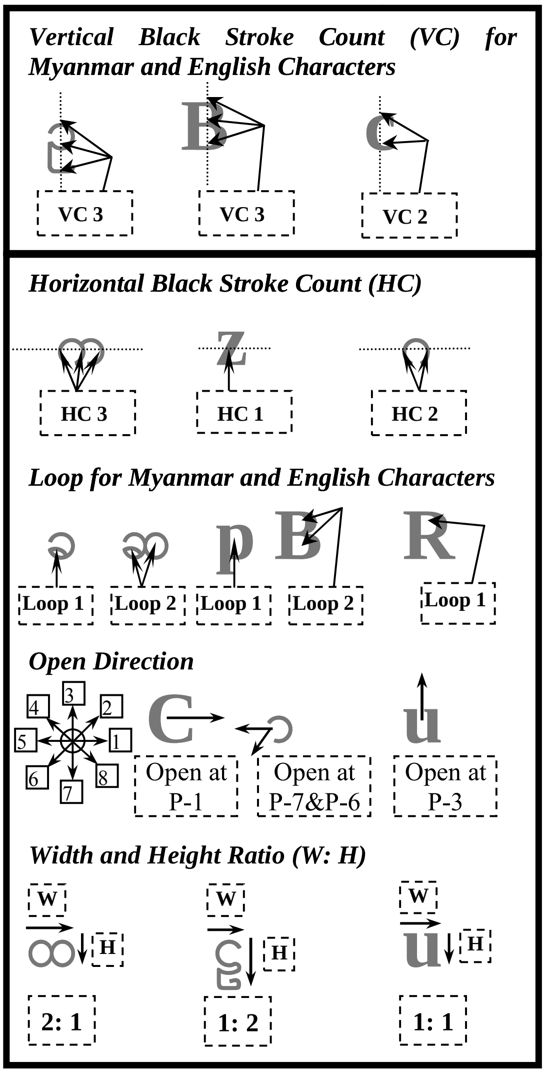

Vertical Black Stroke Count (VC) for Myanmar and English Characters
B
C
VC 3
VC 3
VC 2
Horizontal Black Stroke Count (HC)
Z
HC 3
HC 1
HC 2
Loop for Myanmar and English Characters
p
B
R
Loop 1
Loop 2
Loop 1
Loop 2
Loop 1
Open Direction
1
2
3
4
5
6
7
8
C
u
Open at P-1
Open at P-7&P-6
Open at P-3
Width and Height Ratio (W: H)
W
H
u
2: 1
1: 2
1: 1

Fig. 10 Statistical and semantic information

11. Format Extraction

11.1 Alignment

The alignment of the passage depends on the margin. The boundary of the passage is assumed as the margin. Minimum x is the left margin. Maximum x is the right margin. The center point of these points is (maximum x-minimum x)/2+minimum x. If the incoming paragraph is near to left margin, it is left alignment or near to right margin right alignment, otherwise center alignment.

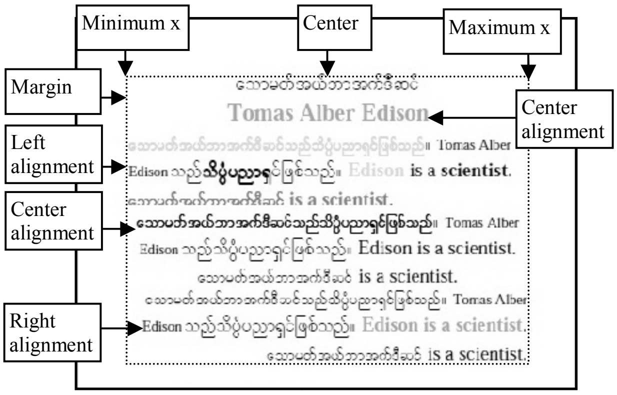


Fig. 11 Alignment

11.2 Font size

The recognition of font size depends on the height of its basic consonant character. Basic characters can be divided into two groups for font size. They are one row characters (တ, a, c, e, m, n, o, r, s, ပ, သ, က, င, လ, etc.) and two rows characters (န,ရ,ဍ ,ဋ , ဠ, ဌ, ဉ, ည, k, M, l, S, etc). For one row characters, their font size is equal to their height. For two rows character, the font size is nearly equal to half of its height.

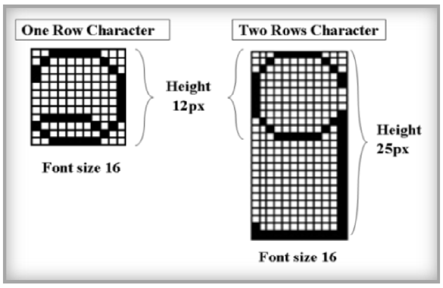


Fig. 12 Font size of each character

Table 2: Relation between font size and height

Font size	*Myanmar Characters*	*Myanmar and English Characters*	*English Characters*
	Height of the first group	*Height (second group or Capital letter)/2*	*Height of small letter*
16	11/12	12	13
18	13	13	14
20	14/15	14	16
22	16	15	17
24	17/18	17	19
26	19	18	20
28	20/21	20	22
36	27	25	28
48	36	34/35	37
72	54	52/53	55

11.3 Font Color

The recognition of font color can be done by indexing RGB image. After indexing, the image will contain two rgb colors; one for background color and one for foreground color. It has two options that are 'dither' and 'no-dither'. Dither option can cause noise and distortion of an image. Therefore, 'no-dither' option is needed to be set.

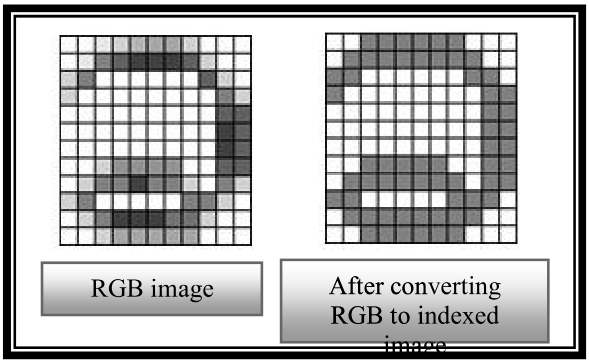


Fig. 13 Font color of each character

11.4 Bold

Whether the character is bold depends on the font size and the pixel count at the black stroke. The pixel count of bold character is greater than normal character.

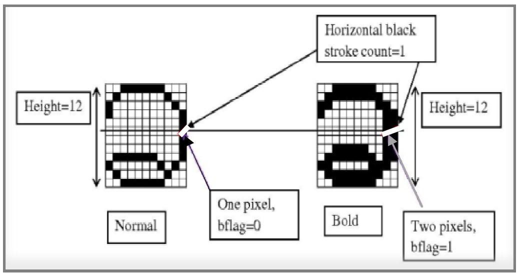


Fig. 14 Bold for each character

Table 3: Pixel information at black stroke for bold and font size

Font size	*Normal*	*Bold*
16	2px	4px
18	3px	5px
20	3px	5px
22	3px	5px
24	3px	5px
26	3px	5px
28	4px	6px
36	4px	6px
48	5px	7px
72	8px	9px

12. Applied Format

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, table format is applied to the array. Moreover, paragraph alignment is assigned to the string array. Furthermore, font size and bold are added to the array.

13. Related Code

In this system, the voting system produced the code number for each recognized character. These code numbers are changed into their relative Unicode or ASCII code. For example, 'u' is the code number for 'က' in the Unicode system. Then, these codes are arranged according to Unicode. English characters have code number. The following table shows the Unicode Sequence or ASCII code sequence for some Myanmar and English words.

In Word Document, various types of font face are stored in a table. The related code is appended to the string array by indexing the related font face. The whole string array is transferred to the word document. The output of the system is Editable Word Document.

13.1 UNICODE

In most of the world's writing systems, Unicode is allowed as a computing industry standard to represent and manipulate text. The Unicode Standard consists of a repertoire of more than 100,000 characters, a set of code charts for visual reference, an encoding methodology and set of standard character encodings, an enumeration of character properties such as upper and lower case, a set of reference data computer files, and a number of related items, such as character properties, rules for normalization, decomposition, collation, rendering and bidirectional display order.

For example, "ကိုး", 'က' ' ိ' ' ု' 'း' ⇨'U+1000' 'U+102D' 'U+102F' 'U+1038'

Table 4: Unicode Sequence of Myanmar character

Myanmar Characters	*Unicode Sequence*
က…ည… အ	U+1000…U+100A…U+1021
ျ ြ ွ ှ	U+103B U+103C U+103D U+103E
ါ ာ ိ	U+102B U+102C U+102D
ီ ု ူ	U+102E U+102F U+1030
ေ ဲ း	U+1031 U+1032 U+1038

13.2 ASCII Code

Coding of characters have been standardize to exchange recorded data between computers efficiently. The most popular standard is ASCII (American Standard for Information Interchange). It consists of 7bits for each character.

For example, ‘c’ ‘a’ ‘t’ ⇨‘99’ ‘97’ ‘116’

Table 5: ASCII codes of English alphabets

English Characters	*ASCII Sequence*
A….Z	65….90
a….z	97….122

14. Limitation

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.

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

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.

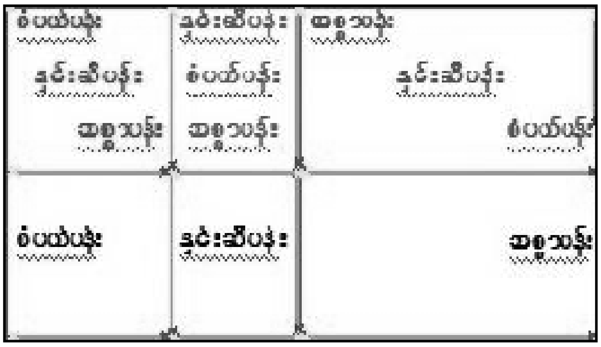
Font size : In Myanmar character, the smallest font size is 16. If the font size is smaller than 16, there will be noise in it.

Similar pattern : MICR misrecognize for similar pattern.

Table 6: Misrecognition of some similar patterns

Myanmar characters	*English characters*
တ	m
င	c
ဝ	o
ပ	u
ဂ	n
ဓ	Q
၂	J

15. Experimental Result



စပယ်ပန်း နှင်းဆီပန်း [illegible]	နှင်းဆီပန်း စပယ်ပန်း [illegible]	[illegible] နှင်းဆီပန်း စပယ်ပန်း
စပယ်ပန်း	နှင်းဆီပန်း	[illegible]

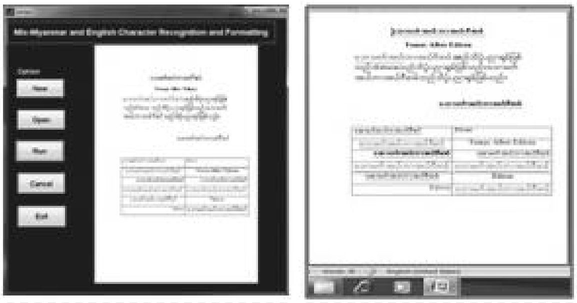
Fig. 15 2rows, 3columns table with format

Table 7: Experimental Result of 2rows, 3columns Table

Cell	*Original Document*		*JPG image in pixel*		*Output Editable Word Document*	
	Width	*Height*	*Width*	*Height*	*Width*	*Height*
1	2241	1880	230.9	203	2309	2030
2	1519	1880	159.1	203	1591	2030
3	4700	1880	482	203	4820	2030
4	2241	1109	230.9	198	2309	1980
5	1519	1109	159.1	198	1591	1980
6	4700	1109	482	198	4820	1980

Table 8: For Character Recognition

Samples	*Recognition accuracy*		
	Myanmar character	*English character*	*Mix-Myanmar and English*
10 words	98.62%	98.68%	97.67%
50 words	95.23%	95.87%	94.98%
Over 50	92.16%	92.42%	90.34%


Mix-Myanmar and English Character Recognition and Formatting
Option
New
Open
Run
Cancel
Exit

INPUT, Before Recognition

OUTPUT, Editable Word

Fig. 16 Final result of the proposed system

16. Conclusion

In conclusion, the main contribution of this system is the recognition of table format with Myanmar character in Portable Document Format (.pdf) document. There is no research work with table recognition in Myanmar. This system produces high accuracy rate for table format such as width and height, boundary color and etc. It can also extract other format; paragraph format and text format. The recognition of characters is based on MICR. MICR can cause misrecognization for similar character pattern; ‘c’ and ‘ၔ’, ‘m’ and ‘ဏ’. Experimental result of the system mainly depends on MICR and font size. Accuracy rate of bold depend on PDFCreator. Although there is minor error, the system can produce the nearest value of the input image.

Acknowledgments

I am very grateful to my supervisor, my family and my friends for their immerse love and encouragement of the research. I want to thank all my senior friends who have done some of the previous applications of MICR. I would like to thank the reviewers and editors of IJCSI. Finally, I appreciate all the readers who spend their precious time to read this paper.

References

[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 Technomathematics 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

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

[10]www.En.Wikipedia.org/wiki/Unicode

Dr. Yadana Thein I achieved M.Sc (Master Computer Science) in 1996 and PhD (I.T) in 2007. I am now associated professor of U.C.S.Y (University of Computer Studies, Yangon). I have written about 25 papers altogether. About 10 of them are local papers and 15 are foreign papers. My first paper is “Recognition of Myanmar Handwritten Digits and Characters “for ICCA conference in 2007. My current research interest is MICR (Myanmar Intelligent Character Recognition) field.

Cherry Maung I am a Master Thesis Student. I got B.C.Tech in 2008 and B.C.Tech (Hons.) in 2009. I got two distinctions (English and Image Subjects) in Master course work exam.