International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:11, 2017

Myanmar Character Recognition Using Eight Direction Chain Code Frequency Features

Kyi Pyar Zaw, Zin Mar Kyu

***Abstract*—**Character recognition is the process of converting a text image file into editable and searchable text file. Feature Extraction is the heart of any character recognition system. The character recognition rate may be low or high depending on the extracted features. In the proposed paper, 25 features for one character are used in character recognition. Basically, there are three steps of character recognition such as character segmentation, feature extraction and classification. In segmentation step, horizontal cropping method is used for line segmentation and vertical cropping method is used for character segmentation. In the Feature extraction step, features are extracted in two ways. The first way is that the 8 features are extracted from the entire input character using eight direction chain code frequency extraction. The second way is that the input character is divided into 16 blocks. For each block, although 8 feature values are obtained through eight-direction chain code frequency extraction method, we define the sum of these 8 feature values as a feature for one block. Therefore, 16 features are extracted from that 16 blocks in the second way. We use the number of holes feature to cluster the similar characters. We can recognize the almost Myanmar common characters with various font sizes by using these features. All these 25 features are used in both training part and testing part. In the classification step, the characters are classified by matching the all features of input character with already trained features of characters.

***Keywords*—**Chain code frequency, character recognition, feature extraction, features matching, segmentation.

I. Introduction

RECENTLY, Myanmar character recognition is an active research area for various applications. The authors used various methods for Myanmar character recognition. Some researchers recognized Myanmar handwritten characters based on structural features and statistical features. Myanmar character recognition were developed for the applications of car plate number recognition, date and amount number recognition on the bank check, printed character recognition on standard application form, bilingual Myanmar-English character recognition form document image, handwritten Myanmar consonant character recognition. Nature of Myanmar characters is circular shape and similar. Therefore, it is difficult to properly recognize the Myanmar characters. To solve this problem, various features of Myanmar characters are extracted by the researchers.

OCR System has been developed maturely for different languages such as English, Japanese or Chinese etc. but Myanmar OCR is still infancy because there are still few researchers available, the difficulty of the language itself and also because some researchers attempt to solve basic problem Myanmar OCR only.

In Myanmar character, there are 33 consonants, 12 vowels, four medials, 10 digits and 12 independent vowels. And then, Myanmar characters can be divided into two types such as basic characters (က (ka), ခ (kha), ဂ (ga), …, အ (a)) and extended characters (ဿ (thawayhtoe), ဉ (yachar), ဳ (lonegyitin), ံ (sankhat), ျ (yapin), ြ (yayit), ွ (waswe), ှ (hahtoe), ၎ (hnachaungpin), ၍ (tachaungpin)). In each basic character (consonant), there can have zero or more extended characters to create a compound word. The extended characters may be at left or right or top or bottom of the basic character. Segmentation between basic character and extended characters is the difficult problem in the printed and typed face character recognition system. Therefore, the segmentation of touching compound characters in Myanmar script (e.g.: ကာ (kar), ကု (ku), ကြ (kya), ကျ (kya), ကိ (ki), ကွ (kwa), ကံ (kathat), etc) is still challenged. Almost previous paper related Myanmar character recognition system recognized only consonants, digits and some papers recognized on non-touching handwritten characters based on statistical features and structural features. It is not found the chain code frequency features extraction in Myanmar character recognition systems. Therefore, we use this feature extraction method for Myanmar character recognition. Figs. 1-5 represent all characters in Myanmar script.

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		

Fig. 1 Myanmar Basic Consonants

အ	အာ	အိ	အီ	အု	အူ
အေ	အဲ	အော	အော်	အံ	အား

Fig. 2 Myanmar Basic Vowels

Kyi Pyar Zaw is with the University of Computer Studies, Mandalay (UCSM), Myanmar (e-mail: kyipyarzaw08@gmail.com).

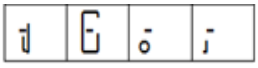


Fig. 3 Medials

Fig. 4 Independent Vowels

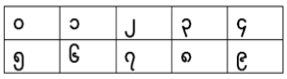


Fig. 5 Myanmar Digits

II. Related Works

A very few work of character recognition has been reported on Myanmar script. Most of the work is found on Machine Printed document Images and handwritten document images of Myanmar script. In 2005, Swe and Tin proposed a Myanmar printed character recognition and translation system using a hoped field neural network. They displayed experimental result using a standard application form. They achieved 97.56% detection rate [1]. Also in 2005, Mon and Sein developed Myanmar Handwriting Text Recognition System based on Hidden Markov Model. This paper also transforms the recognized handwritten text into printed characters [2]. In 2006, Phyu and et al. proposed online handwritten Myanmar compound words recognition system based on Myanmar Intelligent Character Recognition (MICR). They achieved 95.45% and 93.81% recognition rate for typeface and handwritten compound words respectively [3].

In 2008, Aye and et al. proposed a Myanmar Voice Mixer (MVM) System through Myanmar Intelligent Character Recognition (MICR). They achieved 94.57% MVM recognition rate for compound words [4]. In 2009, Theingi and et al. designed a Recognizer for both Online and Offline Handwritten Myanmar Pali Characters using MICR. They achieved 98.45% classification rate for Myanmar Pali Characters [5]. In 2010, Thein and Yee contributed an effective Myanmar Handwritten Characters Recognition System using MICR and back propagation neural network. This system only takes 3 seconds average processing time for 1000 word samples and 93% recognition rate for 1000 samples of noise free image [6]. In also 2010, they developed an effective recognition and detection erratum approach for Myanmar Handwritten Compound Words. In this paper, 90% recognition rate and 88.60% erratum detection rate for over 50 handwritten compound words. Myanmar Intelligent Character Recognition (MICR) used the statistical and structural features of Myanmar characters. In this MICR system, voting system is mainly used according to the nature of MICR [7].

In 2011, Win proposed a Bilingual OCR System for both Myanmar and English script using multiclass- Support Vector Machine (SVM). They used connected component segmentation method, 25 features of zoning, 60 features of horizontal and vertical profiles methods. This paper achieved 98.89% segmentation rate for six Myanmar printed documents [8]. In 2013, Htike and Thein proposed Myanmar Handwritten Character Recognition System based on Competitive neural trees (Cnet). In this paper, region based methods are used for feature extraction and 18 features are extracted for entire character image. They achieved 97% recognition accuracy rate for their test data 330 [9]. In 2014, Tint and Aye proposed Myanmar Text Area Detection and Localization from Video Scenes using connected component labeling approach and geometric properties such as aspect ratio. In this paper, Gaussian filter is used for eliminating noise from video scenes [10]. Almost Myanmar character recognition systems used small font size (10, 12, 14) characters in document images. In this paper, we use large font size (32, 40, 48, 56 and 64) since we intend to recognize the text on the displayed board in later.

III. Proposed Model

This paper is based on the Zawgyi-One font characters. In the proposed system, 550 characters are collected using snipping tool since there is no standard character dataset in Myanmar script. These 550 characters include 57 isolated characters and 493 popular compound words. Preprocessing and feature extraction steps are used in both training and testing parts before the character classification. In the training part, only the features of font size 40 characters [12] are stored by clustering the characters into five classes based on the number of holes. In the testing part, the various font size (32, 40, 48, 56, 64) characters are tested. In this paper, not only isolated character images and compound word images but also text-lines images can be recognized.

In this character recognition system, the following three main steps are performed.

A. Segmentation

For an OCR system for text-lines images, segmentation phase is an important phase and accuracy of any OCR heavily depends upon segmentation phase. In this paper, horizontal cropping method is used for line segmentation in two or more text-lines images and vertical cropping method is used for character segmentation in one text-line images.

The horizontal cropping (line segmentation) steps are as follow:

1. Count the black pixel in each row of the image [11].
2. Find the rows containing no white pixel [11].
3. Crop each text-line.
4. Input the cropping text-line to the character segmentation step.

The vertical cropping (character segmentation) steps are as follow:

1. Count the black pixel in each column of the image [11].
2. Find the columns containing no white pixel [11].
3. Crop each character
4. Input the cropping character to feature extraction step.

B. Feature Extraction

The feature extraction is described about the characteristics of an image. It is one of the most important components for

any recognition system, since the classification/recognition accuracy is depending on the features. In this character recognition system, 8-direction chain code frequency features on the whole character and sixteen blocks based TCCF features are extracted from each character. The illustration of 8-direction chain code extraction is shown in Fig. 6 and described as an algorithm. Furthermore, one of the structural features such as the hole number of characters is also extracted for clustering the characters in training part. Therefore, 25 features are extracted in both training part and testing part. According to the hole number feature, we can cluster the 550 characters into five groups such as zero-hole group, one-hole group, two-holes group, three-holes group and four-holes group in the training part. By clustering the characters, we can reduce the matching time in the testing part. In this system, average character recognition time is 0.03 sec. After clustering the characters, zero-hole group contains 120 characters, one-hole group contains 193 characters, two-holes group contains 159 characters, three-holes group contains 65 characters and four-holes group contains 13 characters. Illustrations of three sample characters in each group are shown in Figs. 8 (a)-(e).

(i) Eight Direction Chain Code Algorithm

Begin
Input: Image Boundary
Output: Chain Code
Step1. Define sp of Image Boundary.
Step2. Let sp_x=0, sp_y=0.
Step3. cp_x = sp_x and cp_y= sp_y.
Step4. If np_x==1 and np_y==0, cc = 0.
Elseif np_x==1 and np_y==1, cc=1.
Elseif np_x==0 and np_y==1, cc=2.
Elseif np_x==-1 and np_y==1, cc=3.
Elseif np_x==-1 and np_y==0, cc=4.
Elseif np_x==-1 and np_y==-1, cc=5.
Elseif np_x==0 and np_y==-1, cc=6.
Elseif np_x==1 and np_y==-1, cc=7.
End if
Step5. Repeat step 3 and step 4 until ep of Image Boundary.
End

where, sp = start point, cp= current point, np_x= x coordinate of next point, np_y= y coordinate of next point, cc= chain code, ep= end point.

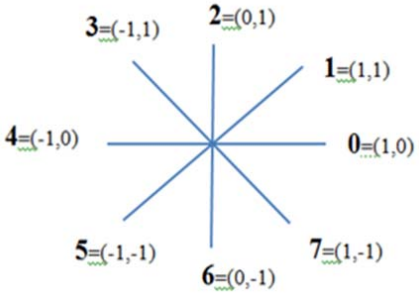


Fig. 6 Eight-direction chain code

(ii) Proposed Feature Extraction Algorithm

Begin,

Input: Pre-processed (100x100) size normalized character image as shown in Fig. 7 (a)

Output: 25 features.

1. Find the number of hole for the input character.
2. Find the chain code frequency of eight directions on the whole character by (1) and outcomes eight features.
3. Divide the input character into 16 non-overlapping blocks as Fig. 7 (b).

a. Find the chain code frequency of eight directions on each block as in step 2.
b. Sum the frequencies of eight directions by (2) and these total frequency value is assumed as one feature for one block. Therefore, 16 features are obtained for 16 blocks.

4. Finally, 25 features are extracted from step 1, step 2, step 3.

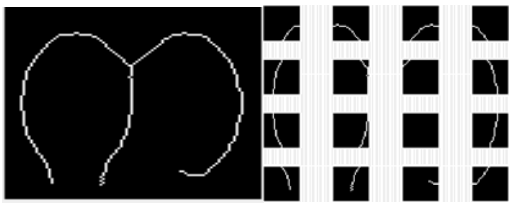
End

Note that, C_i = Count of i direction and F_i = Frequency of i direction on the whole character; where, i = 0, 2, 3, …., 7.

$$F_i = C_i \quad (1)$$

Note that, B_n = Total frequency of eight directions on Block n of character, where n= 1, 2, 3, …, 16.

$$B_n = \sum_{i=0}^{7} F_i \quad (2)$$



(a) (b)

Fig. 7 (a) Preprocessing character and (b) 16-blocks character

Fig. 8 (a) Three sample characters in zero-hole group

Fig. 8 (b) Three sample characters in one-hole group

Fig. 8 (c) Three sample characters in two-holes group

Fig. 8 (d) Three sample characters in three-holes group

Fig. 8 (e) Three sample characters in four-holes group

C. Features Matching

Features matching or generally image matching is a part of computer vision application such as object recognition and is the task of establishing correspondences between two images. A common approach to features matching consists of detecting a set of interest points each associated with image descriptors from image data. Once the features and their descriptors have been extracted from two or more images, the next step is to establish some preliminary feature matches between these images. [13].

D. Classification

In this paper, to classify a character, features matching method is used. Characters are recognized by matching the features of input character with the features of all characters in corresponding groups that are already clustered according to the number of holes. The distance values between input character and the characters in the clustered group are calculated using Euclidean distance (3). The minimum distance value is selected and returns the position of that value.

$$dist\big((x,y),(a,b)\big) = \sqrt{(x-a)^2+(y-b)^2} \qquad (3)$$

Finally, the Myanmar Zawgyi-One code of that position is generated as output and displayed in the editor (notepad or word file) as editable characters. The purpose of this paper is to successfully recognize Myanmar characters from the text document image and store them with ease. Examples of isolated-character images, compound word images and text line images are shown in Figs. 9-11.

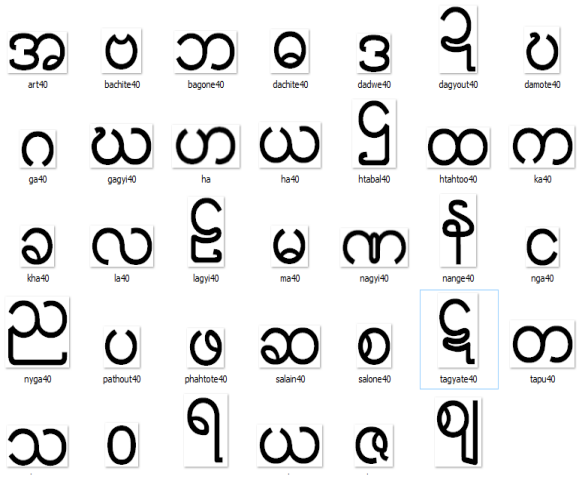


Fig. 9 Examples of isolated-character images

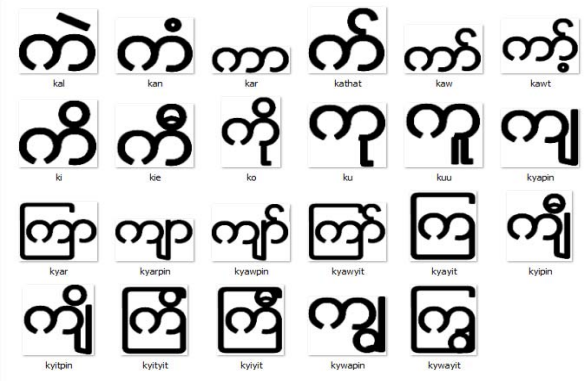


Fig. 10 Examples of compound word images related with 'ka'

Fig. 11 Examples of text line images with various font size

IV. Experimental Results

In this experiment, 550 Myanmar common characters with Zawgyi-One font size 40 are used for training features. Isolated-character images, compound word images that do not need to segment and text-lines images that need to segment are tested using various font size. The classification results are shown in Tables I-III. According to the experimental results, we can see that the larger font size is the higher recognition rate and the smaller font size is the lower the recognition rate compared with font size 40.

TABLE I
Classification Accuracy of Isolated-Character Images

Font Size	Number of tested characters	Number of correct characters	Rate of recognition Accuracy
32	57	43	75.44%
40	57	57	100%
48	57	54	94.74%
56	57	53	92.98%
64	57	56	98.25%
Average recognition rate			92.28%

TABLE II
Classification Accuracy of Compound Word Images

Font Size	Number of tested characters	Number of correct characters	Rate of recognition Accuracy
32	493	381	77.28%
40	493	493	100%
48	493	425	86.21%
56	493	428	86.82%
64	493	426	86.41%
Average recognition rate			87.34%

TABLE III
CLASSIFICATION ACCURACY OF TEXT LINES IMAGES

Font Size	Number of text line image	Number of included characters	Number of correct characters	Rate of recognition Accuracy
32	11	206	175	84.95%
40	11	207	187	90.34%
48	11	207	179	86.47%
56	17	204	173	84.80%
64	17	206	175	84.95%
Average recognition rate				86.31%

V. CONCLUSION

This paper has presented Myanmar character recognition using eight direction chain code frequency features based on the whole character and 16-blocks character. Features matching method is used for classification and recognition. After classifying with the use of proposed features, 92.28% on isolated character images, 87.34% on the compound word images and 86.31% on text line images are achieved. Therefore, the proposed feature extraction method achieves the acceptable classification accuracy on both Myanmar isolated characters and compound words from the text images. In further extension, we will try to recognize the text from natural scene images such as signboard text by training the various font styles and font size. In this paper, we have used the large font size of 32, 40, 48, 56 and 64 since we are trying to recognize the text displayed on notice boards and in advertising billboards.

REFERENCES

[1] T. Swe and P. Tin, 2006, "Recognition and Translation of Myanmar Printed Text based on Hopefield Neural Network" IEEE.
[2] S. S. Mon and M. M. Sein, 2006, "Recognition of Myanmar Handwriting Text Based on Hidden Markov Model".
[3] E. E. Phyu, Z. C. Aye, E. P. Khaing and Y. Thein, 2007, "Recognition of Myanmar Handwritten Compound Words based on MICR".
[4] Z. C. Aye, E. E. Phyu, Y. Thein and M. M. Sein,2008, "Myanmar Intelligent Character Recognition (MICR) and Myanmar Voice Mixer (MVM) System".
[5] E. Theingi, E. K. Khine, T. W. K. kyaw, Y. Thein, 2009, Enhance the Handwritten Myanmar Characters Recognition System for Pali based on MICR".
[6] Y. Thein and S. S. S. Yee, 2010, "High Accuracy Myanmar Handwritten Character Recognition using Hybrid approach through MICR and Neural Network", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November, ISSN(online):1694-0814.
[7] Y. Thein and S.S.S Yee, 2010, "Online Myanmar Handwritten Compound Words Recognition and Erratum Detection with MICR".
[8] H. P. P. Win, P. T. T. Khine and K. N. N. Tun, 2011 "Bilingual OCR System for Myanmar and English Scripts with Simutaneous Recognition", International Journal of Scientific & Engineering Research Volume 2, Issue 10, October, ISSN 2229-5518.
[9] T. Htike and Y. Thein, 2013, "Handwritten Character Recognition Using Competitive Neural Trees", IACSIT International Journal of Engineering and Technology, Vol. 5, No. 3, June 2013.
[10] Thuzar Tint, and Nyein Aye, 2014, "Myanmar Text Area Identification from Video Scenes", International Conference on Advanced in Engineering and Technology (ICAET2014), March, Singapore.
[11] Emmanuel, Rosemol, and Jilu George. "Automatic detection and recognition of Malayalam text from natural scene images." *IOSR Journal of VLSI and Signal Processing* 3.2 (2013): 55-61.
[12] Sok, Pongsametrey, and Nguonly Taing. "Support Vector Machine (SVM) Based Classifier For Khmer Printed Character-set Recognition." *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014.
[13] Hassaballah, M., Aly Amin Abdelmgeid, and Hammam A. Alshazly. "Image Features Detection, Description and Matching." Image Feature Detectors and Descriptors. Springer International Publishing, 2016. 11-45.