Character Segmentation and Recognition for Myanmar Warning Signboard Images

Kyi Pyar Zaw[1,*], Zin Mar Kyu[2]

[1]*Faculty of Computer Science Department, University of Computer Studies, Mandaly, Patheingyi, Mandalay 1001, Myanmar*
[2]*Faculty of Computer Science Department, University of Computer Studies, Mandaly, Patheingyi, Mandalay 1001, Myanmar*

ARTICLE INFO

ABSTRACT

This paper publicizes the character segmentation and recognition of the Myanmar warning text signboard images taken by a mobile phone camera in natural scene. In this system, two templates are created. The first template that contains both connected pixel words and characters are used for character segmentation and the second template that contains only the connected pixel characters are used for character classification. Color enhancement process is first performed to extract the text regions. By preprocessing the color enhancement, the system can overcome the some illumination conditions. To remove the background noises on the binary images, color threshold based filtering, aspect ratio based filtering, boundary based filtering and region area based filtering techniques are used. As a next step, line segmentation and character segmentation are done. Line segmentation is performed using horizontal projection profile and character segmentation is done using vertical projection profile and bounding box methods. In the character segmentation process, template matching method is used by training connected pixel words. These connected component characters are recognized using 4×4 blocks based pixel density and total chain codes, four rows-based pixel density, four columns-based pixel density and count of eight directions chain code on the whole character image and on each block of character image. This system is investigated by feature-based approach of template matching on 160 camera-captured Myanmar warning signboards.

1. INTRODUCTION

Optical character recognition system for Orientals languages have been studied extensively in the literature. There are a number of reports available for the recognition of scripts of oriental languages as compared with other non-Latin scripts. Signboard text analysis and understanding in these scripts faces specific challenges because of:

 (i) The existence of large number of characters in the scripts,

 (ii) The degree of complexity of character formation, and

 (iii) The existence of various visually similar characters in the scripts

Thus, character recognition becomes a touch task due to the large number of classes and high degree of inter-class similarity.

Warnings about the display of signs are important in several areas of application: machine language translation, text-to-speech, help for visually impaired and foreigners. There are some methods and are described in the literature. Some researchers recognize Myanmar handwritten characters based on structural and statistical characteristics. Myanmar's character recognition system was developed to apply license plate recognition, date identification and number of numbers on bank checks, typed character recognition in the standard application form, bilingual design of Burmese symbols, with English symbols, Burmese handwritten character recognition. They are primarily based on pattern matching and normalized correlation with large databases of storage patterns.

Although data sets in many different languages have been developed, the Myanmar text extraction and recognition system does not have a standard data set. So we created 160 Signage Image Datasets to capture signage images from offices, hospitals, universities, pagodas and streets. In this system, the foreground text region is extracted by filtering the background noise based on the color threshold, the aspect ratio threshold and the boundary tangent object, and the character region and the text grouping region. Since Myanmar text does not have an embedded Optical Character Recognition (OCR) reader, we generate important functions of the Burmese symbol. Due to the nature of Burmese text, such as italics, circles, and consisting of one or more extended characters or modifiers or vowels with basic characters forming words, it is difficult to find the basic features of Burmese characters. Burmese text is more complex than English and has 62 characters (26 uppercase letters, 26 lowercase letters and 10 numbers). In the Myanmar scenario, there are not only 75 basic characters, but also a total of more than 1881 glyphs [8]. This number is many and there are many similarities in this language. The longest component used to form the glyph is 8. Due to the contour of the horizontal projection, the vertical projection and the outline of the bounding box are used for the segmentation of the symbols, so it is sufficient for the system to train only 64 symbols. Most experiments in the Myanmar Character Recognition System are based on their

*Corresponding author. Email: kyipyarzaw08@gmail.com

own file (handwritten or printed) datasets, using structural features, semantics and statistical functions. Myanmar's character recognition system has been developed for license plate recognition applications where only 34 consonants and 10 digits are required [9] to identify date numbers and numbers in bank checks [10]. The character of Burmese characters is round, left and right, and so on. However, this script lacks the concept of uppercase/lowercase characters. In the warning SMS image of Myanmar, you can get different numbers of characters and different numbers of text lines, because they are written according to the emotions of those who want to prohibit doing something. Therefore, it is difficult to correctly extract and recognize the text on Myanmar warning signboard image in natural scene. Most authors divide text recognition in symbolic images into three phases. These steps include foreground text extraction, text segmentation (segment and symbol segmentation), and character recognition. Intensive pre-processing is required to complete the task at each stage. In this system, color enhancement processing is first performed on the original image of the logo. The foreground text is extracted by filtering noise based on color thresholds, aspect ratios of characters, boundary tangent objects, and character regions and text grouping fields. In this paper, we estimate the recognition rate of related component characters in the warning text tag image captured by the mobile device. The rest of this paper includes related previous research works presented in Section 2, Myanmar script nature are described in Section 3, dataset is explained in Section 4, proposed research including text extraction, line and character segmentation, feature extraction technique and classification are described in Section 5. Experimental results and comparison is shown in Section 6. Findings and discussion is described in Section 7. Finally, conclusion is placed in Section 8.

2. PREVIOUS RESEARCH WORKS

Most of the algorithms for detecting text in the literature can be divided into methods based on space-based and connectivity (CC). The regional approach uses a sliding window scheme, which is basically a powerful method. The second way is to localize individual symbols using local parameters of the image (intensity, area, color, gradient, etc.). The selection of a function also plays a decisive role in the image positioning process. The main goal of object extraction is to maximize recognition speed. Many researchers have made research related to this but no technique is almost perfect and they found need to improve the work in more areas at different instants and techniques.

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. The literature numbers for the past OCR system of Myanmar scripts can be counted. Thaung [1] was for the recognition of handwritten data on bank cheques using directional zoning method. The goal of this research was to develop the recognition of handwritten dates on Myanmar bank cheques using zoning method. The system aimed for 0–9 digits. Each sub character was divided into 16 zones, so 4 × 4 features were obtained. And then 8 directions for each zone, so that the final feature vector had (16 × 8) and got 128 components. In 2004, Thein developed Myanmar Cars License Plate Recognition System. In this system, the authors used Discrete Cosine Transform to locate the license plate and identified the symbols of license plate

images using back-propagation neural network [2]. This system correctly identified 91.5% on the 200 plate images. As the system was for car license plate, it considered the characters for and digits "0–9". The sub image was normalized into 32 × 32 image and got 128 dimension "က to အ" ns for 32 × 4 sides and 16 dimension for 4 × 4 sides. Mar presented the identification of the handwritten Myanmar characters on exhibit and specimen documents based on the shortest distance method. Characters were extracted by fast Fourier transform method and Euclidean distance approach was used for character identification [3]. The system tested five different Myanmar characters; "က", "မ", "အ", "န", "လ" and resulted 90.8% for classification. In 2004, Nge et al. [4] proposed a new histogram labeling method for recognizing the handwritten Myanmar digits. This system recognized the 2100 digit symbols using two methods (Method I and II) based on number of features. Method I used 377 features and Method II used 544 features. The Method I achieved 75.52% recognition rate and Method II achieved 98.19% recognition rate. In 2004, Sandar [5] proposed a system to recognize offline Myanmar handwriting using discrete Hidden Markov Models. In this system, Myanmar characters are auto-segmented using recursive algorithm and normalized contour chain codes are extracted by counting direction categories along the chain code.

In 2008, E.E. 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 [6]. In 2010, Y. Thein and S.S.S. 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 [7]. In 2011, H.P.P. 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 6 Myanmar printed documents [8]. In 2014, Angadi and Kodabagi presented a robust segmentation method for line, word and character extraction from Kannada Text in low resolution display board images. They used projection profile features and pixel distribution statistics for segmentation of text lines. They also used *k*-means clustering to group inter-character gaps into character and word cluster spaces. This method achieved text line segmentation accuracy of 97.17%, word segmentation accuracy of 97.54% and character extraction accuracy of 99.09% [9]. In 2008, N.A.A. Htwe from Mandalay Technological University suggested the system about automatic Extraction of Payee's Name and Legal Amount from Myanmar Cheque by using Hidden Markov Model. This system proposed a model to register the payee's name and legal amount on Myanmar bank cheque by using Hidden Markov Model [10].

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. They used Gaussian filter for eliminating noise from video scenes [11]. In 2017, Zaw proposed a method of segmentation for Myanmar character recognition using blocked-based pixel count and aspect ratio. The segmented characters are classified using feature matching method. In this system, only printed character from the text images can be segmented and recognized by training 98 Myanmar

typed-face characters. In this system, 94.77% segmentation rate and 80.37% classification rate on 20 text line images of font size 32, 40, 48, 56 and 60 are achieved [12]. In 2018, Khn publicized Myanmar character extraction system using the license plate number. This system used aspect ratio and bounding box to settle the problem of identification of license plate. This system also used edge boxes in order to separate the background from the foreground and accentuate in the foreground. The experimental results showed that the accuracy rate of 90% [13].

The previous papers in related character extraction, segmentation, and recognition of other languages are observed. In 2016, S.A.B. Haji [14] introduced a new approach to segmenting Malayalam handwritten documents into components, words and symbols related to the problem. The system uses water flow techniques to extract text lines and proposes an algorithm for processing tangent and overlapping lines. Words from text strings are detected using the Spiral Run Length Smearing Algorithm (SRLSA). In addition, skew correction is performed on the extracted words, and corrected words are created for the division of the characters. A skew correction is included to facilitate the recognition phase in the handwritten Malayalam OCR. In 2013, Choudhary et al. [15] proposed a new handwriting recognition segmentation algorithm. In this method, the segmentation points are located after the sparse image words to obtain the width of a single pixel stroke. The author points out that the challenge lies in the autonomous segmentation of handwritten characters. The author achieved an accuracy of 83.5% in 200 words. In 2012, Roy et al. [16] suggested using a dynamic programming-based approach for multi-directional tangent segmentation of text characters. The algorithm first divides the touch characters into primitives, and then uses these primitive fragments to find the best character sequence based on the dynamic programming method. The author also specializes in the segmentation of touch characters based on a preliminary understanding of the number of characters in the touch line. The author achieved the best results in this case, but this is not possible in the real world because it is difficult to get a preliminary knowledge of the number of characters in the image.

3. CHARACTERISTIC OF MYANMAR SCRIPT

Myanmar has 33 consonants, 12 vowels, four drugs, 10 figures, three asatas, one kinzi and 12 independent vowels. You can then divide the characters into two types, such as basic characters and extended characters. Creating a compound word in each basic character (consonant) may be zero or more extended characters. The extended characters can be on the left or right side of the main character or above or below. Segmentation between primary and extended characters is a complex problem in printing and typing character recognition systems. Therefore, the segmentation of complex and complex characters in Burmese scripts is still challenging. Since a character is formed by adding different characters and labels, most letters have some common characteristics.

Myanmar characters have a number of notable characteristics. This is mainly attributed to character formation in the script. Each word can be formed by combining consonants vowels and various signs. It has own specified composition rules for combining vowels, consonants and modifiers. The longest component to form a glyph is 8. There are total of above 1881 glyphs, a large number, and has

many similarity scripts in this language. In Myanmar script, there is no distinction between upper case and lower case characters. Myanmar writing direction, just like Latin-based scripts, goes horizontally from left to right, then top-to-bottom. When writing text, space is used after each phrase instead of each word or syllable. The punctuation marks are used to indicate the end of a phrase with a single line called "pote htee" and for sentence, use two vertical lines called "pote ma". The shapes of Myanmar scripts are circular, consist of straight lines horizontally or vertically or slantways, and dots. The style of writing is done in upper, middle, and lower zones.

4. DATASET

In the proposed system, two training datasets are created. The first training dataset includes 177 characters (118 connected pixel words and 59 connected pixel characters). This dataset is used for individual character segmentation by matching training word and the input word. The second training dataset for character classification includes 64 characters. To save the time of segmentation process, the 177 training characters are grouped into 24 groups according to the 24 styles of segmentation process. Such groups are shown in Table 1. And also 64 training characters for character recognition are clustered as three groups such as 15 below characters, six top characters and 53 middle characters to distinguish the similarity of basic (middle) characters and modifiers (top or below) characters as shown in Table 2.

In Myanmar regular characters, the above and below modifiers of basic characters are smaller than basic (middle) characters. And also the basic characters may have left and right modifiers. The characters resulted from the bounding box process are segmented based on above 24 groups of training dataset. The segmentation styles of each group are shown in Table 3.

5. PROPOSED SYSTEM

Many early works [1,3,10] have shown that the combination of different methods can improve the performance of text extraction and recognition systems. This prompted us to incorporate threshold-related areas in our research, suggesting features and effective classifiers. In the proposed system, the following four basic steps are performed.

5.1. Text Extraction

In the text extraction phase, color reproduction is improved to overcome various lighting conditions. The color-enhanced image displayed in the first stage is converted to binary, as shown in the second stage of Figure 1. In this case, since the text color on the white mark is white, we fill in the threshold with "1" if the color value is >215 and <255 or fill "0" in other color value to get the binary image. Related components in a binary image may contain components that are too large and/or too small. These components are obviously not text. These single-ended related components must be detected and removed without deleting any characters. Here, a heuristic filter is used to filter out very large non-text connected component based on height, weight and

Table 1 | Twenty four groups of training dataset for character segmentation

Group	Symbols
1	ch-1 … ch-25
2	ch-26 … ch-53
3	ch-54 … ch-64
4	ch-65 … ch-71
5	ch-72, ch-73
6	(ch symbol)
7	ch-74, ch-75, ch-76, ch-77
8	(ch symbol)
9	ch-79, ch-80, ch-81, ch-82, ch-83, ch-84, ch-175, ch-176, ch-177
10	ch-86, ch-87
11	ch-88, ch-89
12	ch-90, ch-91
13	ch-92, ch-93
14	ch-94, ch-95, ch-96, ch-97, ch-98, ch-99, ch-100, ch-101, ch-102, ch-103
15	ch-104, ch-105
16	(ch symbol)
17	ch-107, ch-108, ch-109
18	ch-110, ch-111, ch-112
19	ch-113, ch-114
20	(ch symbol)
21	(ch symbol)
22	ch-117
23	ch-118
24	ch-119 … ch-174

Table 2 | Three groups of training dataset for character classification

Top characters		Middle characters						Below characters		
tr-1	tr-2	tr-1	tr-3	tr-3	tr-4	tr-5	tr-6	tr-1	tr-2	tr-3
tr-2	tr-4	tr-7	tr-8	tr-9	tr-10	tr-11	tr-12	tr-4	tr-5	tr-6
		tr-13	tr-14	tr-15	tr-16	tr-17	tr-18			
tr-5	tr-6	tr-19	tr-20	tr-21	tr-22	tr-23	tr-24	tr-7	tr-8	tr-9
		tr-25	tr-26	tr-27	tr-28	tr-29	tr-30			
		tr-31	tr-32	tr-33	tr-34	tr-35	tr-36	tr-10	tr-11	tr-12
		tr-37	tr-38	tr-39	tr-40	tr-41	tr-42	tr-13	tr-14	tr-15
		tr-43	tr-44	tr-45	tr-46	tr-47	tr-48			
		tr-49	tr-50	tr-51	tr-52	tr-53				

aspect ratio. The remaining small non-text connected component of the output image from this stage is further filtered based on the area connected component threshold (manually defined as <30) and the result is shown in step 4. The connected component value exceeding the regional area threshold is filled with white pixels, and such connected component is expanded into group text. After grouping the text regions, the remaining non-text regions are deleted based on the region thresholds of the text region grouping (manually specified <1000). The sample input signboard image and the illustration steps of text extraction process are shown in Figures 1 and 2.

5.2. Line Segmentation and Character Segmentation

In this character recognition system for warning signs, the segmentation phase is an important phase, and the accuracy of any OCR depends to a large extent on the segmentation phase. The image of the data center warning sign can contain 1–5 lines of text. Therefore, it is necessary to divide the line of text into several lines of text. Below is the horizontal projection outline of the line segmentation and vertical projection outlines of the vertically connected symbols. The contour of the split of the horizontal projection is as follows:

1. Calculates the black pixels in each line of the image.

2. Finds the line containing the white pixels.

3. Segments each line of text.

4. Enter an image of the trimmed text line in the outline vertical position. The implementation of line segmentation is shown in Figure 3.

The steps for dividing the vertical projection outline symbol and the bounding box method are as follows:

1. Calculate the black pixels in each column of the image.

2. Find the column that contains the white pixels.

3. Crop each vertical connected character (characters may not be horizontally connected).

**Table 3** | Twenty four segmentation styles for 24 groups

If the character resulted from bounding box process is match with the character of the group	Segment the character sample that must be (row > 30 and col > 30) as
1	(Row/2)-5
2	(Row/2)+5
3	2*col/3
4	Col/2
5	
6	
7	
8	Row/3
9	
10	
11	
12	
13	
14	Row/3 2*Row/3
15	
16	
17	Row/3
18	
19	
20	
21	
22	
23	
24	(no need to segment)

Figure 1 | Various implementation steps of text extraction.

Figure 2 | Input warning signboard image.

4. Use the bounding box method to remove the symbols of the single connection from the symbols using the labels of the related components. Here, we define a condition based on the ratio of height to format for the tonal symbol (:) to overcome the missing segmentation of the associated component tag with the bounding box. The symbols given after step 4 are shown in Figure 4a and 4b.

5. Enter the final extracted character image resulting from the boundary separation operation.

In Figure 4a, the second character is match with the character of group-22, the fourth character is match with the character of group-23 and the fifth character is match with the character
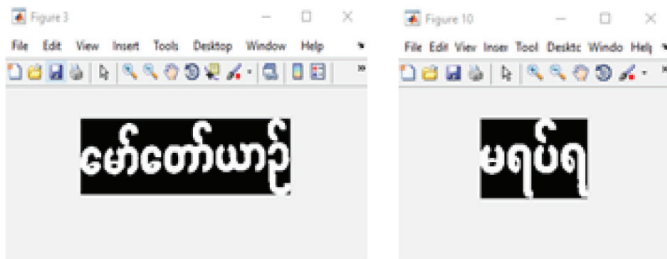

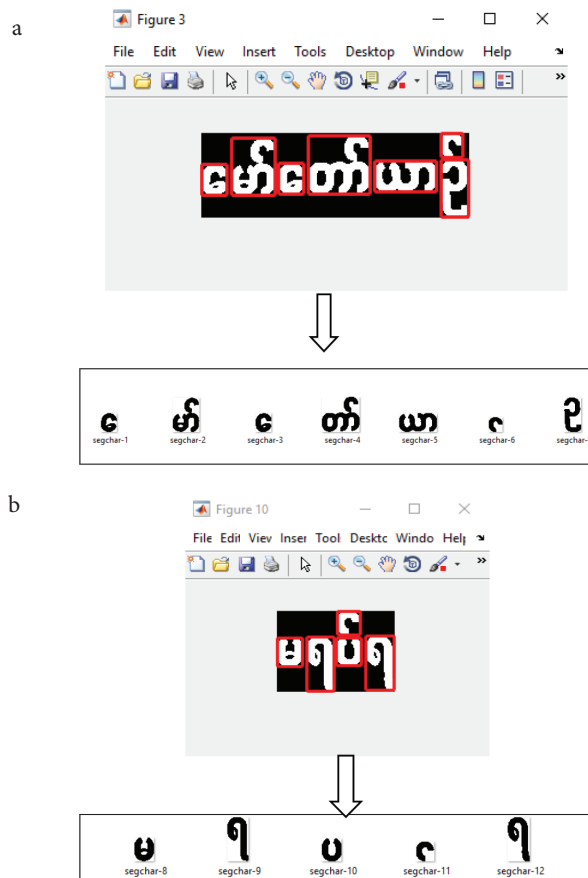**Figure 3** | Implementation of horizontal projection (Line segmentation).

Figure 4 | (a) Bounding box implementation of second line character segmentation. (b) Bounding box implementation of first line character segmentation.

of group-3 of training dataset created for segmentation process. Therefore, such characters are need to segment and the remaining characters are no need to segment. The segmentation of second, fourth, and fifth characters of Figure 4a is shown in Figure 5a–5c.

5.3. Feature Extraction

The characteristics of the image features are described. This is one of the most important components of any identification system because the accuracy of classification/recognition depends on the characteristics. In this case, the search function associated with the search character will reach the standard size of 100×100 before the search function. In this system, the total number of block symbols (16) and 16 integers blocks the image (each block has 25×25 binary pixels), 16 white pixel densities, four rows of pixel densities, and the pixel density of four columns of images. In this way, 160 functions are extracted for the connected characters of the training and related characters extracted from the image with the verified logo. Since the characters are evenly changed to 100×100, the extracted attribute values, some of the basic characters and modifiers shown in Figure 6, are very similar. Therefore, their original width, height and coordinates (x, y) are further used. Use the corresponding component label with a bounding box to adjust some missing characters based on the number of connected components horizontally or vertically in each of the two regions, as shown in Figure 7.

5.3.1. Eight directions chain code

Eight direction chain codes is one of the shape representations which are used to represent a boundary by a connected sequence of straight line segments of specified length and direction.
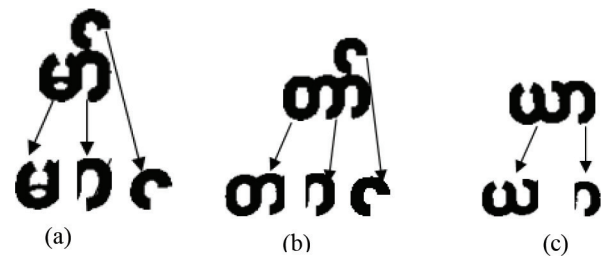

Figure 5 | (a) 22nd Group-based segmentation style of second character, (b) 23rd group-based segmentation style of fourth character and (c) 3rd group-based segmentation style of fifth character.
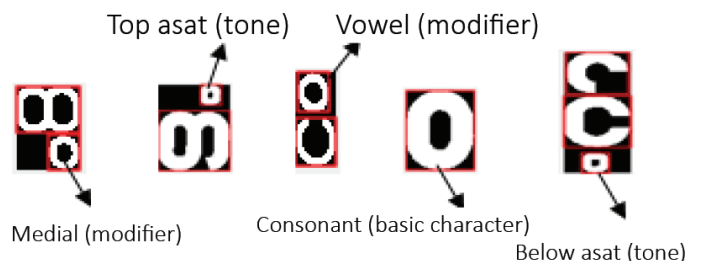

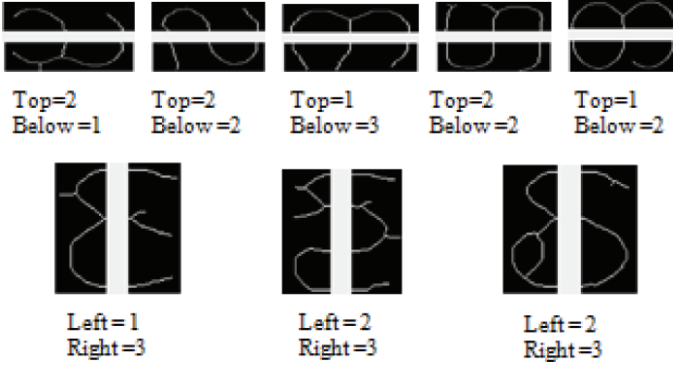**Figure 6** | Five similar characters with different positions.

Figure 7 | Sub-component features in two horizontal zones (first row) and sub-component features in two vertical zones (second zone).
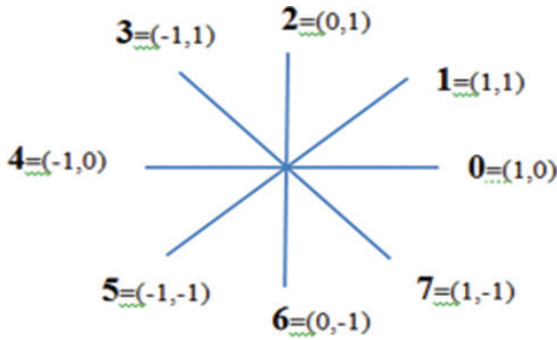

Figure 8 | Eight chain code directions.

The direction of each segment is coded by using a numbering scheme as shown in Figure 8. Chain codes based from this scheme are known as Freeman chain codes.

The circuit code can be generated clockwise after the object boundary and assigned to the line segment connecting each pair of pixels. First, we select the starting position of the pixel from the leftmost point of the object's border. Our goal is to find the next pixel on the border. One of the eight positions around the current boundary pixel should have adjacent boundary pixels. Taking into account each of the eight adjacent pixels, we find at least one and boundary pixels. Based on this, we assign a numeric code from 0 to 7. For example, if the found pixel is directly in the upper right corner, the code is "0". If the found pixel is directly in the upper right corner, the code "1" is assigned. Repeat the search process for the next border pixel and assign the code until we return to the original border pixel. The result is a list of code chains showing the direction from each pixel of the border to the next pixel.

5.3.2. Proposed feature extraction algorithm

Begin

Input: Pre-processed (Resizing and Thinning) normalized character image.

Output: 160 features.

1) Divide the input character into 4×4 blocks as shown in Figure 9.

   a. Find the total number of eight directions in each block to extract 16 features by Equation (1).

   b. Perform step 1 for each blocks and extracts 8×16 features (8 features of 16 blocks).

   c. Find the number of white pixels for 16 blocks to extract 16 features by Equations (2) and (3).

   d. Find the number of white pixel by merging four blocks in each row for four rows to extract four features by Equations (4)–(7).

   e. Find the number of white pixel by merging four blocks in each column for four columns to extract four features by Equations (8)–(11).

2) Finally, 160 features are extracted from step 1(a)–(e).

End

$$B_n = \sum_{i=0}^{7} F(i) \tag{1}$$

Note that, B_n = total frequency of eight directions on block n of character, where $n = 1, 2, 3, \ldots, 16$.

$$P_{row,col}(x) = \begin{cases} 1, & \text{if } x == 1 \\ 0, & \text{else} \end{cases} \tag{2}$$

$$W_n = \sum_{row=1}^{25} \sum_{col=1}^{25} P_{row,col}(x) \tag{3}$$

Let W_n be one block among 16 blocks of character, row = 1, 2, 3, …, 25, col = 1, 2, 3, 4, …, 25, where, $n = 1, 2, 3, \ldots, 16$.

$$R_1 = W_1 + W_2 + W_3 + W_4 \tag{4}$$

$$R_2 = W_5 + W_6 + W_7 + W_8 \tag{5}$$

$$R_2 = W_9 + W_{10} + W_{11} + W_{12} \tag{6}$$

$$R_4 = W_{13} + W_{14} + W_{15} + W_{16} \tag{7}$$

$$C_1 = W_1 + W_5 + W_9 + W_{13} \tag{8}$$

$$C_2 = W_2 + W_6 + W_{10} + W_{14} \tag{9}$$

$$C_3 = W_3 + W_7 + W_{11} + W_{15} \tag{10}$$

$$C_4 = W_4 + W_9 + W_{12} + W_{16} \tag{11}$$
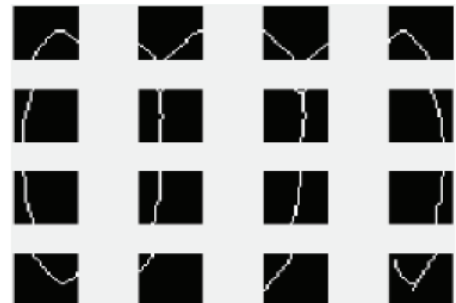

Figure 9 | Sixteen blocks characters.

5.4. Classification

The character recognition system is tested using a pattern based pattern matching method. Characters are identified by matching the functionality of the input characters to the functionality of all 62 characters already stored in the database. The Euclidean distance [Equation (12)] is used to calculate the distance value between the input symbol and the symbol in the cluster group. Select the minimum distance value and return the location of the value. The purpose of this article is to successfully identify Burmese symbols from text warning images and save them easily.

$$\text{dist}\big((x, y), (a, b)\big) = \sqrt{(x - a)^2 + (y - b)^2} \tag{12}$$

6. EXPERIMENTAL RESULTS

In this experiment, 64 characters were trained. 160 warning text signs, of which 4324 characters were corrected. Here, the classification accuracy is not very good, because the system only prepares one sample for each character in order to classify different font styles and handwritten characters from the warning text tag image. Symbol classification/extraction indicators and symbol classification characteristics are shown in Tables 4 and 5.

7. FINDINGS AND DISCUSSION

There are a few limitations in this system such as:

1. Since this system classifies the text and non-text regions based on region area ranges, some very small size characters and very large size characters are detected as non-texts. The very small noise in the text region may be mis-detected as characters shown in Figure 10.

2. Since Myanmar characters on the natural warning signboard images are various styles and various sizes in the same

Table 4 | Character segmentation performance of the system

Number of tested images	Total extracted characters	Truly segmented characters	Segmentation accuracy
160	4580	4324	94.4%

Table 5 | Character classification of the system

Number of images	Truly segmented characters	Truly classified characters	Classification accuracy
160	4324	2810	65%

Figure 10 | Miss-classified characters as non-text.

Table 6 | Miss classified character samples with various styles

Different styles of same characters	Correct classified character	Mis-classified character
(thawayhtoe)	G	C (nga) or ⊖ (dawachite)
(athat)	C	C (nga) or Ω (gange)
(yaychar)	Ɔ	Ɔ (one) or Ω (gange)

characters, the system may be mis-classified with similar characters as shown in Table 6. This problem can be solved by training more than five samples for each character.

3. If characters or connected words in the warning text signboards are overlapping or very touching this system may decrease the segmentation accuracy.

8. CONCLUSION

The main objective of this paper is to present Myanmar character segmentation and character recognition from warning text signboard using count of chain codes and number of pixel on 16-blocks of character and eight direction chain code features on the whole character and on each block of character image. The proposed system has used template matching method to classify 4324 characters extracted from 160 warning signboard text images and feature-based approach of template matching method. In this system, although segmentation accuracy is very good, the classification accuracy is not so good because the system trains only the 64 characters (one sample for each character). As further extension, the classification accuracy can be improved by training more training samples for every character for classification process.

CONFLICTS OF INTEREST

There is no conflict of interest.

REFERENCES

[1] H.H. Thaung, Recognition of handwritten data on bank cheques using directional zoning method, International Conference on Computer Application, 2004.

[2] N.L. Thein, Myanmar cars license plate recognition system, International Conference on Computer Application, 2004.

[3] S.H. Mar, Identification of the handwritten Myanmar characters on exhibit and specimen, International Conference on Computer Application, 2004.

[4] M.M. Nge, New histogram labeling method for recognizing the handwritten Myanmar digits, International Conference on Computer Application, 2004.

[5] K. Sandar, Off-line Myanmar handwriting using discrete hidden markov models, International Conference on Computer Application, 2004.

[6] E.E. Phyu, Z.C. Aye, E.P. Khaing, Y. Thein, M.M. Sein, Recognition of Myanmar handwritten compound words based on MICR, the 29th Asian Conference on Remote Sensing (ACRS), Colombo, Sri Lanka, 2008.

[7] Y. Thein, S.S.S. Yee, High accuracy Myanmar handwritten character recognition using hybrid approach through MICR and neural network, Int. J. Comput. Sci. 7 (2010), 22–27.

[8] H.P.P. Win, P.T.T. Khine, K.N.N. Tun, Bilingual OCR system for Myanmar and English scripts with simultaneous recognition, Int. J. Sci. Eng. Res. 2 (2011), ISSN 2229–5518.

[9] S.A. Angadi, M.M. Kodabagi, A robust segmentation technique for line, word and character extraction from Kannada text in low resolution display board images, 2014 Fifth International Conference on Signal and Image Processing, IEEE, Bangalore, India, South Korea, 2014, pp. 42–49.

[10] N.A.A. Htwe, S.S. Mon, M.M. Sein, Recognition on user-entered data from Myanmar bank cheque, International Conference on Computer Application, Yangon, 2008, pp. 114–118.

[11] T. Tint, N. Aye, Myanmar text area identification from video scenes, International Conference on Advanced in Engineering and Technology (ICAET2014), Singapore, 2014, pp. 389–394.

[12] K.P. Zaw, Z.M. Kyu, Segmentation method for Myanmar character recognition using blocked based pixel count and aspect ratio, 27th International Conference on Computer Theory and Application, 2017.

[13] O. Khin, M. Phothisonothai, S. Choomchuay, Myanmar character extraction from vehicle images using aspect ratio and bounding box, 2018 International Workshop on Advanced Image Technology (IWAIT), IEEE, Chiang Mai, Thailand, 2018.

[14] S.A.B. Haji, A. James, S. Chandran, A novel segmentation and skew correction approach for handwritten Malayalam documents, Procedia Technology 24 (2016), 1341–1348.

[15] A. Choudhary, R. Rishi, S. Ahlawat, A new character segmentation approach for off-line cursive handwritten words. Proc. Comput. Sci. 17 (2013), 88–95.

[16] S.B. Ahmed, S. Naz, M.I. Razzak, R. Yousaf, Deep learning based isolated Arabic scene character recognition, 2017 First International Workshop on Arabic Script Analysis and Recognition (ASAR), IEEE, Nancy, France, 2017.