Text Extraction and Recognition System for Myanmar Warning Signboard Images

Kyi Pyar Zaw
Faculty of Computer Science
University of Computer Studies, Mandalay
Mandalay, Myanmar
kyipyarzaw@ucsm.edu.mm

Nu War
Faculty of Computer Systems and Technologies
University of Computer Studies, Mandalay
Mandalay, Myanmar
nuwar@ucsm.edu.mm

***Abstract*— Text extraction and character recognition are the computer vision tasks which became important after smart phones with good camera. Character recognition from scene text images is still challenging area, because the camera captured text images have various background noise and the text also varies in shape, font, color. In this research, a camera captured based text extraction and character recognition system is developed for Myanmar warning text images. One major challenge of this research is that Myanmar OCR system has been greatly under-researched on the camera captured images. This research therefore considers these challenges of Myanmar OCR system and proposes a new algorithm for segmentation and recognition of Myanmar script. In the character segmentation, zone-based character segmentation is performed using position and size of connected component objects. Combination of features with chain code, pixel density, a new shape-based features of boundary-centroid distance and centroid-boundary distance, are explored and exploited in recognition process. This system uses K-Nearest Neighbors (KNN) classifiers to recognize the segmented characters. From the experiment, this system achieves satisfied results 93.9% segmentation accuracy and 92.77% classification accuracy.**

Keywords—Character segmentation, Character classification, Shape-based features, Warning text sign

I. Introduction

Character Recognition (OCR) is one of the most fascinating and challenging areas of pattern recognition with various practical applications. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. Many commercial systems for performing character recognition exist for a variety of applications, although the machines are still not able to complete with human reading capabilities. Character recognition is performed on three types of image such as document text image, scene text image and overlay text image. All scene text is written in different styles, different fonts, different sizes and different colors.

To understand the need for recognition, a system is proposed that uses the extracted and recognized text in different languages (such as English, Chinese, Japanese, Korean, Kannada and Arabic). Text widely exists in natural text, mostly in the form of informative signage like warning sign, road sign, shop sign and traffic sign. Given the rapid growth of camera-based applications readily available on mobile phones, understanding natural text is more important than ever, and application that are able to answer questions such as, "What does this sign say?" are becoming increasingly popular.

Potential applications of natural text recognition include improving navigation for visually impaired people, converting recognized text into other languages when translating it, improving image search, and helping cars and robots navigate automatically. However, unlike images, which usually have standard glyphs on a standard background, structured text, warning sign images often contain dramatic lighting changes and may contain unusual or highly stylized fonts. Variable number of characters, making it harder to take advantage of language restrictions or learn repeating appearance styles. In this system, the character recognition technique is evaluated using the characters extracted from warning text signboard images having Myanmar scripts and achieved encouraging results. The rest of this paper includes literature reviews presented in Section II, characteristics of Myanmar scripts are described in Section III; text line extraction process is presented in Section IV; proposed character segmentation processes are described in Section V, proposed feature extraction is presented in VI, classification method is described in Section VII, data collection is explained in Section VIII, Experimental results are shown in Section IX and finally, conclusion is placed in the Section X.

II. Literature Reviews

The literature numbers for the past OCR system of Myanmar scripts can be counted. In the Myanmar text area identification system, the method of labelling connected components and geometric properties (such as aspect ratio) used to detect and locate Myanmar text areas from video text. The authors also used Gaussian filters to remove noise from the video text [1]. A method that can determine the area of interest of the license plates of Myanmar vehicles obtained under different environmental conditions, such as different types of license plates, image shooting angles and actual environmental conditions was proposed [2]. This method uses horizontal and vertical expansion, tilt angle detection, and automatic bounding boxes to determine the license number based on the input image. The system is conducting experiments on 33 Myanmar car images, which include not only Myanmar license plates but also English alphabet license plates, and has achieved an average detection accuracy of 99%.

Recently, Text Image Recognition System [3] was developed for mobile environment using Myanmar Character Dataset. The system mainly used Connected Labelling Algorithm for character segmentation and Convolutional Neural Network was used to recognize the words in a warning text sign. The authors used 250 syllables warning text sign words to train this system and achieved 0.95 precision and 0.94 recall rate after 50 iterations. Myanmar text extraction system using a license plate is published as an experimental design. The author uses a border to separate the background from the foreground and emphasizes the foreground. The system achieves 90% extraction accuracy with 48 letters and 52 digits [4].

A new segmentation and recognition algorithms for Myanmar script inputted from offline printed images considers horizontal and vertical zones; it is applied to segment letters according to their roles such as primary or peripheral characters. In doing so, statistical and structural features of segmented characters are explored and exploited in recognition process. Hidden Markov model is used for recognition of primary characters while Kohonen self-organization map is used for peripheral characters. The recognized characters by each model are then combined, and finally are recognized by k-nearest neighbours' algorithm with the help of lexicon is composed of all common Myanmar characters. The authors investigated on 2380 characters and achieved 89.28% recognition accuracy [5]. A method based on efficient binarization and improvement techniques, followed by a suitable program for analysing the relevant components was proposed [6]. Image binarization successfully processes natural text images with shadows, uneven lighting, low contrast, and signal-related noise. The system was studied based on the ICDAR 2003 dataset.

. A text extraction and recognition system for warning text images was created by K. P. Zaw in 2018. Characters are segmented in that method only using the bounding box techniques. Therefore, all connected component characters in the Myanmar text script are used to train and recognize characters or words extracted from the bounding box using the features of the chain code and other structural features [7]. H.P.P. Win also proposed a method for the conversion of printed Myanmar using multi-class support vector machine in [8]. They used segmentation methods for related components, 25 partition functions, and 60 horizontal and vertical section methods. The document has a resolution of 98.89% for six Myanmar printed documents.

P. N. Sastry [9] used Zonal based features to recognize the Telugu handwritten characters. The character image is divided into predefined number of zones. By using these features, the system achieves 78% recognition accuracy. Features of Bengali characters have been carried out which is presented in a hierarchical structure. The first few layer deals with features that broadly classify the characters into small size groups. The lower-level features are more specific to each character within a group. While the higher-level features can be identified based on pixel density and arrangement, the lower level features have been identified using chain code technique [10]. Q. Zhang extracted the shape of a shell image using the CCD method [11]. They calculated the center point from the boundary of the shell. And before calculating the distance between the center point and the point boundary of each step, the point P at the shell boundary starting from 0 ° moves 5 ° to 360 ° every step (using the interval angle α = 5 °, only 360/5 = Step 72). Finally, 72 distance points are generated for a single shell image. The system selects 1340 samples for verification. The KNN classifier can achieve 67% shell recognition accuracy.

III. Characteristic of Myanmar Scripts

In the Myanmar text line, characters are written with three zones as in Indian script. The illustration of three zones and real text line (L) with red color bounding boxed characters is shown in Fig 1.

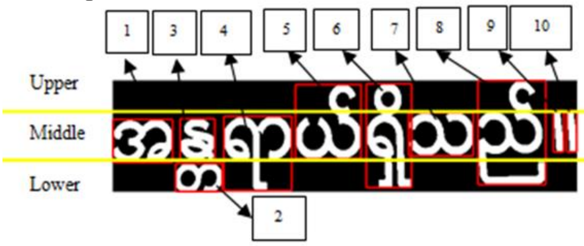
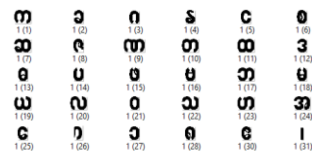
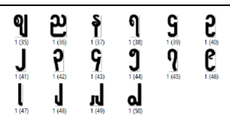
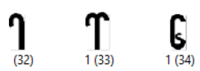
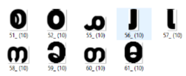
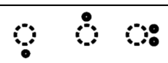


Fig 1. Myanmar text line

This system considers 72 single characters that are commonly found in the warning text lines. These characters can be clustered as small dot-shape symbols and six types of located zones as shown in TABLE I.

TABLE I. Description of single characters in Myanmar scripts

Group	Existed Zone in the text line	Character Description	No characters
1	Middle zone		30
2	Middle zone and lower zone		16
3	Upper zone and middle zone		3
4	Full zones		6
5	Lower zone		9
6	Upper zone		5
Very small character symbols			3

IV. Text line extraction

In this process, candidate texts are detected using canny edge detector. After obtaining the Canny edge-map, connected component labeling is performed and edge pixels are grouped into connected components (CCs) such as candidate text regions using morphological dilation. All regions that are smaller than 5 pixels are ignored in this step. Since images of text in warning text image contain very complex backgrounds, most of the candidate text regions (green text box) correspond to the non-text areas (red text box) as shown in Fig 2. Non-text regions are filtered using (1) and (2).

$$O(x,y)=\begin{cases}0, & (O_w>0.35*I_w)\ \textbf{or}\ (O_h>\ 0.5*I_h)\ \textbf{or}\ (0.2\le\ O_r\le 4.5)\\ 1, & Otherwise\end{cases} \quad (1)$$

$$O_r\ =\ \frac{O_h}{O_w} \quad (2)$$

where, O_w is width of object and O_h is height of object, O_r is aspect ratio of object, I_w is width of the whole image and I_h is height of the image

Fig.2 Non-text region identification (left) and Non-text filtering image (right)

And then, morphological filling and erosion is used for removing the remaining non-text candidate edges that cannot filter by heuristic rule. CC candidates are grouped into a text line by dilating the image I with horizontal line shape structuring element (B_H) that containing 15 pixels and vertical line shape structuring element (B_V) that containing 5 pixels using (3) and (4) respectively.

$$I_H\ =\ \mathrm{I}\oplus B_H \quad (3)$$

$$I_V\ =\ \mathrm{I}\oplus B_V \quad (4)$$

where, B_h = [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
B_v = [1 1 1 1 1]

False positive non-text lines are eliminated according to the width and size of Myanmar text line. The grouping text image is masked on the original color image for localizing and extracting the text lines. Localized color text lines (L) are cropped and binarized based on the gray thresholding. RGB image “I” is converted gray image firstly. The gray image is further converted into binary image by filling ‘1’ in the places where gray intensity values are greater than gray threshold value and filling ‘0’ in other places as follow:

Fig. 3 Grouping binary text line (left) and Localized text line (right)

$$\mathrm{L(x,y)}\ =\ \begin{cases}1 & if\ G(x,y)>threshold\\ 0 & Otherwise\end{cases} \quad (5)$$

$$\mathrm{L}=\begin{cases}1-L, & if\ \sum(L(x,y)==1)>\sum(L(x,y)==0)\\ L, & Otherwise\end{cases} \quad (6)$$

V. CHARACTER SEGMENTATION

In the Myanmar text, the glyphs may be touched or non-touched with each other. In this Myanmar character segmentation process, connected component labeling is used as a pre-segmentation process. Myanmar connected component characters contain one basic character and zero or more modifiers. In real world text image, basic character and its modifiers may be touched. Therefore, it needs to segment the touched characters into individual character that save the number of training characters in the classification process. The Myanmar character segmentation process is performed using the following steps:

(i) Zone Identification
(ii) Touching character identification and segmentation

A. *Zone Identification*

According to the Myanmar script nature, modifiers attach in left, right, top and bottom of the basic character. This identification process is performed by using the following conditions:

S1: $\frac{H_t}{6}\le y<\frac{H_t}{2}$ and $\frac{H_t}{4}<$ h $\le(\frac{H_t}{2}$+5) and d> $(\frac{H_t}{2}-5)$ and $\frac{H_t}{6}<$d-y< $(\frac{H_t}{2}-5)$

S2: $(\frac{H_t}{4}-5)\le y<\frac{2H_t}{3}$ and d-y< $\frac{H_t}{4}$

S3: $\frac{H_t}{3}\le y<\frac{2H_t}{3}$ and d-y < $\frac{H_t}{7}$ and r <12

S4: y $\le\frac{H_t}{4}$ and d> $(\frac{2H_t}{3})$-6

S5: y> $(\frac{2H_t}{3})$-10 and d-y$\le\frac{H_t}{7}$

S6: y$\le\frac{H_t}{6}$ and $(\frac{H_t}{5}-3)\le d<\frac{H_t}{2}$

S7: $\frac{H_t}{5}\le y<\frac{H_t}{2}$ and d-y $\le\frac{H_t}{6}$

where,

d = difference between text line height and character high
y = minimum y-coordinate of character
h = height of character
H_t = height of text line

Zone identification can be described as follow:

$$Zone\ =\begin{cases}MZ, & S1\ or\ S2\ or\ S3\\ UZ, & S4\\ LZ, & S5\\ MUZ, & S6\\ MLZ, & S7\\ FZ, & Otherwise\end{cases} \quad (7)$$

where,

MZ = Middle Zone Character, (eg. က, စာ ကာ …)

UZ = Upper Zone Character, (eg. ◌် ◌ဲ ◌ိ ◌ီ ◌ိ)

LZ = Lower Zone Character, (eg. ◌ု ◌ွ ◌ှ ◌ု …)

MUZ = Middle-Upper Zone Character, (eg. ဖ ◌ါ ပါ ကိ မာ် တာ် …)

MLZ = Middle-Lower Zone Character, (eg. ရ ည ကု ရာ စွာ မျာ …)

and

FZ = Full Zone Character, (eg. ကို နှ် ကြ ရှိ မျိ …)

B. Touching Character Identification and Segmentation

In Myanmar script, touching characters locate in middle zone, middle-upper zone, middle-lower zone and full zone and such touching characters are identified based on the following conditions that used the statistical and structural features:

C1: 0.4<= Rm <=0.8 and Hml=2 and Em=1
C2: 0.4<= Rm <=0.8 and Hml=1 and Sm_t = Sm_b =2
C3: 0.4<= Rm <=0.8 and and Hml=1 and Em=3
C4: Rm <0.4
C5: Sm_t = 1 and Sm_b =4
C6: Sm_t = Hm=1 and Smb=3
C7: Sm_t = Hml=1 and Sm_b =3 and Em=2
C8: Sm_t = Hml= Em=1 and Sm_b =2
C9: Sm_t = Hml=1 and Sm_b =Em=2
C10: Sm_t = Hm=1 and Smb=3 and Hml=0 and Em>=3
C11: Sm_t =1 and Sm_b =3 and Hm=0 and Em=4
C12: Sm_t =2 and Sm_b =3 and Hm=0 and Em>=3
C13: Sm_t = Sm_b =2 and Hm=0 and Em=4
C14: Sm_t =3 and Smb=1 and Em = 4
C15: Smu_m = 2 and Pmu13~=0
C16: Smu_m = 1 and (C1 or C2 or C3)
C17: Smu_m = 1 and (C4 or C5 or C6 or C7 or C8 or C9 or C10 or C11 or C12 or C13 or C14)
C18: Pmu13==0 and Rm>2 and Hmu=0
C19: Smu_l = 1 and Smu_r=2 and Emu<=2 and Emu_{tl}=0 and Emu_{bl}=0
C20: Sml_m = Hm=2 and Rm_r>1.5 and Rm_l>=0.8 and Eml=1
C21: Sml_m =2 and Rm_r>1.5 and Rm_l>=0.8
C22: Sml_m =2 and Rm_r>1.5 and Rm_l<0.8
C23: Sml_m =2 and 0.8< Rm_r<1.5 and Rm_l<0.8
C24: Sml_m =2 and 0.8< Rm_r<1.5 and Rm_l>=0.8
C25: Sml_m = 1 and (C1 or C2 or C3)
C26: Sml_m = 1 and (C4 or C5 or C6 or C7 or C8 or C9 or C10 or C11 or C12 or C13 or C14)
C27: Pml_{13} = 0 and Hm=1 and Eml = 2 and Sm_t=1 and Sm_b=3
C28: Pml_{16} = 0 and Hml = 0 and Rb<1.5
C29: Pml_{15} = 0 and Pml_{16} = 0 and Rb >1.5 and Sm_t=1 and Sm_b=2
C30: Pml_1 = 0 and Pml_{16} = 0 and Sm_t=1 and Sm_b=2 and Sb_t =2 and Sb_b = 1
C31: Rm<0.7 and Sm_t = 2 and Sm_b= 2 and Hml =0 and Eml>=4
C32: Sm_t=1 and Sm_b=2 and (Sb_b=2 or Hml=2) and Hm = 1
C33: Pml_{16} = 0 and Hm=1 and Eml = 2 and Sm_t=1 and Sm_b=3 and Sb_b=2
C34: Sb_b=2
C35: Sm_t=2 and Sm_b=1 and Sml_r=2 and Hm=0 and Hb=1
C36: Hml=0 and Sm_t<=2 and Sm_b=1 and and Sml_l<=2 and Sml_r=2 and Eml<=3
C37: Hm=Hml >=1 and Pml_{13} = 0 and Pml_{16} = 0 and Sml_l=2 and Sml_r=
C38: Hm=Hml=1 and Pml_{13}=Pml_{14}=0 and Sm_t=1 and Sm_b=2
C39: Hm=Hml =1 and Sm_t=1 and Sm_b=2 and Sml_l=2 and Sml_r=1
C40: Hml=0 and Pml_{13}=Pml_{14}=0 and Sm_t=1 and Sm_b=2 and Eml=2
C41: Sm_t=1 and Sm_b=2 and Sml_r=2
C42: Rm>2.5
C43: Sf_m=2 and Rm_l>1.5 and H_t=0 and Pf_{16}~=0
C44: Sf_m=2 and Rm_l>1.5 and H_t>=1 and Pf_{16}~=0
C45: Sf_m=2 and Rm_l>1.5 and Sm_{rt}=1 and Rm_r<0.5 and H_t=0 and Pf_{16}==0
C46: Sf_m=2 and Rm_l>1.5 and 0.5 <=Rm_r<=0.8 and H_t=0 and Pf_{16}==0
C47: Sf_m=1 and (C36 or C37 or C38)
C48: Sf_m=1 and C35
C49: Sf_m=1 and C32
C50: Sf_m=2 and C21
C51: Sf_m=2 and C22
C52: Sf_m=2 and C23
C53: Sf_m=2 and C24
C54: Sf_m=1 and Pf_1=Pf_2=Pf_{15}=Pf_{16}=0
C55: Pf_1=Pf_2=0 and H_t=0 and Sf_l=2

The details of attributes used in the above touching character identification conditions are described in TABLE II.

TABLE II. Detail description of Attributes of Touching Character Identification Conditions

Feature label	Description of features
Rm	$Rm = \frac{hm}{wm}$, Aspect ratio of the middle object (hm = Height of middle object, wm=Width of middle object)
Hm	Number of holes of the middle object
Hm_l	Number of holes at the left side of middle object
Sm_t	Number of segments *in the top part* of middle object.
Sm_b	Number of segments *in the bottom part* of middle object.
Em_{tl}	Number of end point in top-left part of middle object
Em_{tr}	Number of end point in top-right part of middle object
Em_{bl}	Number of end point in bottom-left part of middle object
Em_{br}	Number of end point in bottom-right part of middle object
Em	Number of end point of the whole middle object. $Em = Em_{tl} + Em_{tr} + Em_{bl} + Em_{br}$
Pmu_1, …, Pmu_{16}	Number of white pixels in each block of 16 equally separated blocks of middle-upper object
Hmu	Number of holes of in the middle-upper object
Smu_m	Number of segments in the middle zone in horizontally separated of middle-upper object
Emu_{tl}	Number of end point in top-left part of middle-upper object that 4 parts equally separated
Emu_{tr}	Number of end point in top-right part of middle-upper object that 4 parts equally separated
Emu_{bl}	Number of end point in bottom-left part of middle-upper object that 4 parts equally separated
Emu_{br}	Number of end point in bottom-right part of middle-upper object that 4 parts equally separated.
Emu	Number of end point of the middle-upper object $Emu = Emu_{tl} + Emu_{tr} + Emu_{bl} + Emu_{br}$

Rm_r	Aspect ratio of right part of middle zone object object
Eml	Number of end points of middle-lower object
Hml	Number of holes of middle-lower object
Rb	Aspect ratio of bottom zone object
Sml_l	Number of segments in the left zone of the middle-lower object
Sml_r	Number of segments in the right zone of the middle-lower object
$Pf_1, Pf_2, \ldots, Pf_{16}$	Number of white pixels in each zone of 16 equally separated zones of character
Sm_{rt}	Number of segments in the top part of right side of middle zone in Full character
Ht	Number of holes in top zone of the full object that horizontally three zones separated
Hb	Number of holes in top zone of the full object that horizontally three zones separated

Based on the above zone identification and touching character identification conditions, character segmentation rules are generated as in TABLE III.

TABLE III. Proposed Myanmar Touching Character Segmentation Rules

No	Condition-1	Condition-2	Statements
R1:	Middle	C1 to C8	Vertically project between two middle characters
R2:	Middle-Upper	C9	Horizontally project between middle character and middle-upper character.
R3:	Middle-Upper	Not C9 to 13	Horizontally project between middle character and upper character
R4:	Middle-Upper	C10, C11	Horizontally project as statement of R3 and vertically project as statement of R1.
R5:	Middle-Lower	C22, C23	Horizontally project between middle character and middle-lower character
R6:	Middle-Lower	C24 to C28	Vertically project between middle-lower character and middle character.
R7:	Middle-Lower	C29, C30	Horizontally project as statement of R5 and vertically projected as statement of R6
R8:	Middle-Lower	Not C14 to 35	Horizontally project between middle character and lower character.
R9:	Middle-Lower	C31, C32	Horizontally project as statement of R8 and vertically projected as statement of R1.
R10:	Middle-Lower	C33	Horizontally project between middle-lower character and lower character
R11:	Middle-lower	C34	Vertically project as statement of R6 and horizontally project as in R10.
R12:	Middle-lower	C35	Horizontally project as statement of R8 and vertically project between two lower characters.
R13:	Full	C36	Horizontally project between middle character and full character
R14:	Full	C37	Horizontally project as statement of R13, R3 and horizontally project between full character and upper character.
R15:	Full	C38, C39	Horizontally project as statement of R36 and vertically project as statement of R1.
R16:	Full	C40	Project as statement of R15 and projected between full character and lower character.
R17:	Full	C41	Horizontally project between middle-lower character and upper character
R18:	Full	C42	Horizontally project as statement of R17 and R10
R19:	Full	C43, C44	Horizontally project as statement of R17 and R5
R20:	Full	C45, C46	Horizontally project as statement of R17, R5 and vertically project as statement of R6.
R21:	Full	C50	Horizontally project as statement of R17 and R8 and vertically project as statement of R12
R22:	Full	C51	Horizontally project as statement of R13 and project between full character and lower character.

VI. FEATURE EXTRACTION

In the feature extraction system, the existing features chain code, pixel density hole or loop and new proposed centroid distance features are extracted. As the accuracy of classification / recognition depends on the features, this is one of the most important components of any identification system. The 193 feature vectors containing 128-zone-based chain code features, 16-zone-based pixel density features, 1-hole features, 24-centroid to boundary distance features and 24-boundary to centroid distance features are categorized and recognized by the individual segmented character.

A. *Chain Code Features Extraction*

Chain codes based from this scheme are known as Freeman chain codes. Chain code are extracted by identifying start-point and traversing from one pixel point to next pixel point in close-wise direction using Fig 4.

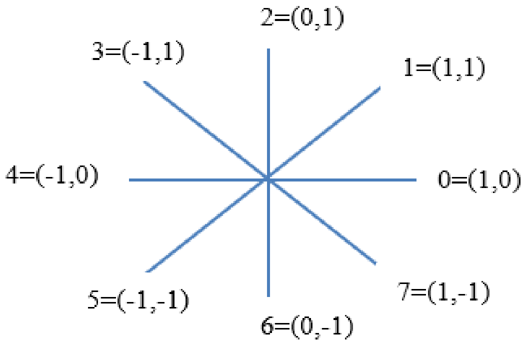


Fig 4. Eight-direction chain code

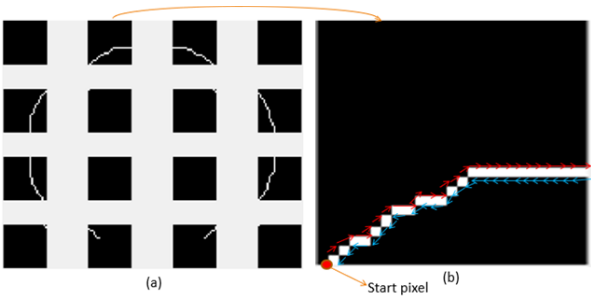


Fig 5. (a) Myanmar character sample with sixteen zones and (b) Traversing of chain code

Chain code extraction algorithm [12] used in this proposed system can be described as follow:

Begin

Input: Pre-processed (90×90 resizing and thinning) character image

Output: 128 normalized chain codes

1) Divide the input character into 16 blocks.

For each block

(i) Find the pixel in the top row of the object with the leftmost value; label this pixel as P0.

(ii) Traverse the current pixel's 3x3 neighborhood counterclockwise.

(iii) Stop when the new P_n border element is equal to the second P1 border element and the previous P_{n-1} border pixel is equal to the first P_0 border element.

2) Normalize chain codes by extracting the frequency of each direction and obtain 8 frequency features for 8 directions.

3) Concatenate the 8 frequency features from each block and obtain 128 frequency features called normalized chain codes

End

B. *Centroid Boundary Distance*

Centroid Boundary Distance (CBD) is one method of extracting form elements. The distance between the gravity center and the boundary point, which is useful for extracting targets beyond the boundary, can be identified by this method. The middle of C can be seen as the object's center of mass, and one of the points on the boundary is P. The distance from the gravity center is the distance between point P and center point C. This method is commonly used in close-shape object recognition such as leaf recognition.

In this article, this method is modified to recognize Myanmar characters that have open-shapes and close-shapes. By using the modified method, 24 centroid-to-boundary-distance features and 24 boundary-to-centroid-distance features are extracted from each Myanmar character.

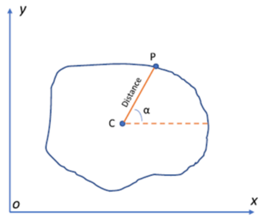


Fig 6. Principle of Centroid Boundary Distance

Proposed centroid distance feature extraction algorithm can be described as follow:

Input: Isolated Myanmar character

Output: 48 centroid features

Step (1): Preprocess the character by thinning and resizing as 90x90

Step (2): Locate 24 points $(r_1, c_1), (r_2,c_2), \ldots, (r_{24},c_{24})$ on the rectangle boundary by separating 15 pixels distance between two points, where topmost-leftmost point is defined as start point.

Step (3): Define the midpoint (R_c,C_c) of the character image using (8).

Step (4): Find 24 white boundary points $(r_1, c_1), (r_2,c_2), \ldots, (r_{24},c_{24})$ of character.

(i) By visiting from each located point of step (1) to the center point to get all outer boundary points as in Fig 7.

(ii) By visiting from the center point to each located point of step (1) to get all inner boundary points as in Fig 8.

Step (5): Find the distances between each new boundary points and the center point using equation (9) for the above two ways.

$$Midpoint(Rc, Cc) = \left(\frac{row}{2}, \frac{Column}{2}\right) \tag{8}$$

$$R(i) = \sqrt{(\hat{r}_i - Rc)^2 + (\hat{c}_i - Cc)^2} \tag{9}$$

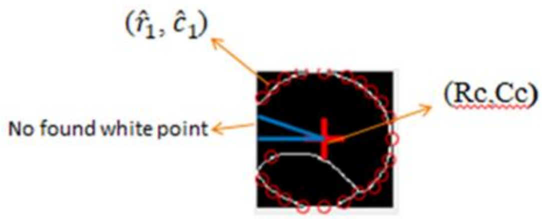


Fig 7. Illustration of Outer Boundary Point

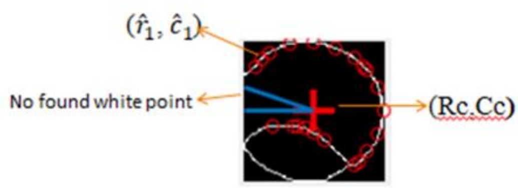
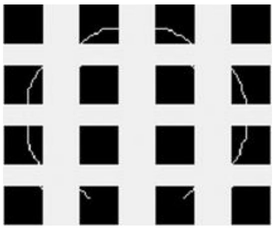


Fig 8. Illustration of Inner Boundary Point

C. Zone based Pixel Density Feature Extraction

In this feature extraction step, 16-pixel density features are extracted from 16 blocks of character image and feature illustration is shown in Fig 9.

$$np_i = \sum_{r=1}^{25} \sum_{c=1}^{25} x_{r,c} \quad (10)$$



0	24	23	0
24	1	2	23
25	0	0	25
2	8	8	1

Fig 9. Character image width 16 blocks (left) and pixel density features of 16 blocks (right)

VII. CLASSIFICATION

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour. In this character classification, KNN is used with k=1 and Euclidean distance.

$$Dis(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots (x_n - y_n)^2} \quad (11)$$

In this character classification of proposed system, 340x194 dimension matrix for training dataset is created to classify the segmented characters.

- 340 rows contain 68 characters with 5 samples
- The first 193 columns contain 193 features and the last column contains class labels.

VIII. DATASET COLLECTION

In this system, 260 common warning text signboard images are collected for text extraction and character recognition. The images are captured with mobile phone camera from various places such as offices, schools, pagodas, streets, universities and bus gates. To recognize each character resulted from the character segmentation process, the training character images are created by using Zawgyi One, Pyidaungsu, Winkalaw, 03Double and 64artHouse.

XI. EXPERIMENTAL RESULTS

In this paper, 260 Myanmar warning text images that contains 481 text lines and 7169 Myanmar isolated characters including dot shape characters are investigated using various feature sets on KNN classifier and SVM classifier. The text line extraction result, character segmentation accuracy and recognition accuracy are measured and performances of this system are described in TABLE IV, TABLE V, TABLE VI and TABLE VII. Where, the text line extraction result is evaluated by using equation (11), equation (12) and equation (13). The results of character segmentation, character classification and character recognition are evaluated using equation (14), equation (15) and equation (16) respectively.

$$P = \frac{TP}{TP+Fp} \quad (11)$$

$$R = \frac{TP}{TP+FN} \quad (12)$$

$$F = 2 \times \frac{P \times R}{P+R} \quad (13)$$

$$Segmentation\ Accuracy = \frac{CSC}{RCC} \times 100 \quad (14)$$

$$Classification\ Accuracy = \frac{CCC}{CSC} \times 100 \quad (15)$$

$$Recognition\ Accuracy = \frac{CCC}{RCC} \times 100 \quad (16)$$

where, TP is true positive, FP is false positive, FN is false negative, P is precision, R is recall, F is f-score measurement, CSC is correctly segmented count, RCC is real character count, CCC is correctly classified count.

The comparison of classification accuracy based on feature types is shown in Table VIII. This system was implemented with Matlab-2018 that installed on Window 10 of Core i-7 that have 4 GB of main memory and 1000 GB Hard Disk.

TABLE IV. Text Line Extraction Result

No. of images	Total text lines	TP	FP	FN	P	R	F-Measure
260	527	524	12	3	0.98	0.99	0.98

TABLE V. Character Segmentation Result

Evaluated on TP =524 text lines	Total	Small vowel	Zone character
Real character count (RCC)	7738	643	7095
Correctly segmented characters (CSC)	7266	639	6627
Segmentation Accuracy (%)	93.9	99.38	93.4

TABLE VI. Character Classification Result of Two Classifiers

Correctly segmented count (CSC=7266)	Correctly classified count (CCC)	Accuracy (%)	Time (s)
KNN	6741	92.77	12.13
SVM	6163	84.82	170.68

TABLE VII. Overall Character Recognition Result using KNN

Number of Images	Total Character	Correctly recognized character	Character recognition accuracy (%)
260	7738	6741	87.12

TABLE VIII. Comparison of Character Classification Result using Various Features -sets

Feature Method, Number of features	Number of correctly segmented zone characters (CS = 6627)		
	Correct classification	Accuracy (%)	Time (s)
Pixel Density, 16	5452	82.27	0.051
Boundary-Centroid, Centroid-Boundary distance, 48	5472	82.57	0.054
Chain Code, 128	5836	88.06	0.106
Chain Code and Pixel Density,144	5919	89.32	0.108
Combination of Boundary-Centroid distance, Centroid-Boundary distance, Chain Code, Pixel Density and Hole, 193 (Proposed)	6152	92.83	0.135

X. CONCLUSION

This paper is intended for developing Myanmar character recognition system for warning text images. After investigating on the 260 warning text images, the system achieves the acceptable results such as 93.9% segmentation accuracy, 92.77% classification accuracy by KNN classifier and 87. 12% recognition accuracy. From the result of TABLE VIII, it is observed that the result of combination of methods is better than the result of each method because proposed centroid distance feature extraction technique is effective for the character shapes that have loop and existing chain code features and pixel density features are useful for different opening-shapes. And also, the classification time of combination methods is not very different with the classification time of each method. In the future, the system is improved by extraction more robust features of Myanmar script. Multi-lingual character recognition system for warning text images will also be considered in the future.

REFERENCES

[1] T.Z.Tint and N. Aye, "Myanmar text areas identification from video scenes," presented at the *International Conference on Advances in Engineering and Technology, Singapore*, March-2014, pp. 389–394, doi: 10.15242/IIE.E0314187.

[2] O. Khin, M. Phothisonothai, and S. Choomchuay, "License plate detection of Myanmar vehicle images captured from the dissimilar environmental conditions," in *2017 International Conference on Advanced Computing and Applications (ACOMP)*, Ho Chi Minh City, Nov-2017, pp.127–132, doi: 10.1109/ACOMP.2017.31.

[3] S. Z. M. Maung and N. Aye, "Convolutional neural network based recognition of Myanmar text warning sign for mobile platform," *International Journal of Engineering and Research*, vol. 8, no. 01, pp.127–132, Jan. 2019, doi: 10.17577/IJERTV8IS010102.

[4] O. Khin, M. Phothisonothai, and S. Choomchuay, "Myanmar character extraction from vehicle images using aspect ratio and bounding box," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, Chiang Mai, Jan. 2018, pp. 1–4, doi: 10.1109/IWAIT.2018.8369714.

[5] C. S. Lwin and X. Wu, "Zone-Wise segmentation and lexicon-driven recognition for printed Myanmar characters," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 162–180, Dec. 2018, doi: 10.32628/CSEIT183844.

[6] B. Gatos, I. Pratikakis, and S. Perantonis, "Towards text recognition in natural scene images," *In Proceedings of International Conference on Automation and Technology*, 2005, pp. 354-359.

[7] K. P. Zaw and Z. M. Kyu, "Character Extraction and Recognition for Myanmar Script Signboard Images using Block based Pixel Count and Chain Codes", *17th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2018)*, June 6–8, 2018, Singapore.

[8] H.P.P.Win, P.T.T. Khine and K.N.N.Tun, "Bilingual OCR System for Myanmar and English Scripts with Simultaneous Recognition", *International Journal of Scientific & Engineering Research*, Volume 2, Issue 10, October-2011, ISSN 2229-5518.

[9] P. N. Sastry, T. R. V. Lakshmi, N. V. K. Rao, T. V. Rajinikanth, and A. Wahab, "Telugu handwritten character recognition using zoning features," in *2014 International Conference on IT Convergence and Security (ICITCS)*, Beijing, China, Oct. 2014, pp. 1–4, doi: 10.1109/ICITCS.2014.7021817.

[10] A. Sikdar, S. Banerjee, P. Roy, S. Mukherjee, and M. Das, "Bengali printed character recognition using a feature-based chain code method," *Adv. Image Video Process.*, vol. 2, no. 3, pp. 1–9, Jun. 2014, doi: 10.14738/aivp.23.189.

[11] Q. Zhang, J. Zhou, J. He, X. Cun, S. Zeng, and B. Zhang, "A shell dataset, for shell features extraction and recognition," *Science Data*, vol. 6, no. 1, pp. 1-9, Dec. 2019, doi: 10.1038/s41597-019-0230-3.

[12] K. P. Zaw and N. War, "Y-Position based Myanmar Touching Character Segmentation and Sub-components-based Character Classification," in *2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA)*, Honolulu, HI, USA, 2019, pp. 76–83, doi: 10.1109/SERA.2019.8886810.