IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 6 Issue 4, April 2019
ISSN (Online) 2348 – 7968
www.ijiset.com

Feature Extraction in OCR for Myanmar Old Printed Documents

Zu Zu Aung[1], Cho Me Me Maung[2] and Yadana Htun[3]

[1] Information and Communication Technology Department, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Mandalay, Myanmar

[2] Computer Engineering Department, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Mandalay, Myanmar

[3] Computer Engineering Department, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Mandalay, Myanmar

Abstract

Nowadays, Myanmar optical character recognition is an open area in research field. A great work has been done for Myanmar handwritten character recognition. But in case of Myanmar old printed characters recognition is limited. In character recognition, feature extraction is very important task for high recognition accuracy. This paper describes a relevant feature extraction method for Myanmar old printed characters recognition. Myanmar old printed character recognition performance is compared with feature extraction method and without feature extraction method.

Keywords: *Myanmar Character Recognition, Preprocessing, Feature Extraction, Classification.*

1. Introduction

Optical Character Recognition (OCR) system converts large amount of documents, either handwritten or printed text into machine encoded text so that it can be easily accessed and preserved [1]. OCR system can apply documents like passport application forms, examination question papers, language translation books, etc. and improve the speed of input operation, decrease some possible human errors, enable compact storage, perform other file manipulations, fast retrieval and automatic number plate recognition [2]. During the last decades a lot of research has been done in the field of Myanmar handwritten optical character recognition (OCR). Myanmar historical documents recognition is still an open problem for the research community. In OCR process, the text image is acquired and various preprocessing steps are operated. In these preprocessing steps, feature extraction is one of the important steps of OCR systems. Selection of feature extraction is directly affected accuracy of recognitions. In this paper, six structural features are extracted for Myanmar old printed characters. It is a challenge for recognition Myanmar old printed document in case of bad quality, absence of standard alphabets, presence of unknown fonts, ink through page, uneven background, broken characters, overlapped scripts and mixed scripts.

2. Related Works

Several approaches have been proposed to enhance OCR accuracy and text detection. The authors of [3], discuss clustering of similar words by using k-Mean clustering algorithm. Then, they used Support Vector (SVM) algorithm for classification problem. In Myanmar, one of our earlier works in [4] developed Handwritten Character Recognition Using Competitive Neural Trees. The performance of trained CNet depends on the global search method in order to improve its recognition accuracy. Although in this research only 33 consonants of Myanmar handwritten characters are recognized and Myanmar compound handwritten characters still remain as problem of recognition. In [5], they implemented High Accuracy Myanmar Handwritten Character Recognition using Hybrid approach through MICR and Neural Network. They stated that statistical and semantic information of MICR used to extract as the features and back-propagation neural network for recognition. The authors of [6] discuss the feature extraction methods based on zoning method and developed rule-based recognition system for Myanmar alphabet. But they can only show the 98.8% accuracy for the first five Myanmar alphabets and not yet done for all Myanmar alphabets and compound words. Converting

Myanmar printed document image into Machine understandable text format by using SVM recognizer is proposed in [7], and they used X_Y cut method on the use of histogram or a projection profile technique for segmentation. Then feature extraction method using zone and projection profile. Their goal is to satisfy the need for digitized information processing for local scripts. Although the researchers mentioned above have their advantages in Myanmar OCR, it still remains an inviting and challenging field of research.

3. Myanmar Character Sets

The Myanmar (formerly known as Burmese) script developed from the Mon script and more than (32) million people use Myanmar script. So Myanmar language is a major language. Myanmar language includes (10) digits, (33) basic characters, (12) vowels, (4) medial and (2) punctuation characters as shown in table 1. Myanmar script is 'round script', is written from left to right [2].

Table 1: Myanmar Character Sets

33 consonants	□ □ □ □ □ □ □ □ □□ □
12 vowels	- □ - □ - □ - □ - □ - □ □- □ - □ - □ -□ □
4 medials	□ □ □ □
10 Myanmar digits	□ □ □ □ □ □ □ □ □ □
Punctuation	□ □

4. Proposed System

We have collected Myanmar old printed documents such as different books, newspapers and magazines around 1938. Firstly, document images might be acquired by scanning or capturing. Preprocessing step involves grayscale conversion and noises are removed by applying median filter. In binarization process, gray scale images are converted into black and white image by using local thresholding method. Distortion alignment of input text, scanned or photo that needs to align it by performing skew correction method. And then lines and words are segmented. Characters are segmented by using block definition method. Six structural features such as aspect ratio, termination points, bifurcation points, horizontal strokes, vertical strokes, weight direction features are extracted. These six extracted features for each characters are trained with convolutional neural network. Proposed system design as shown in figure 1.

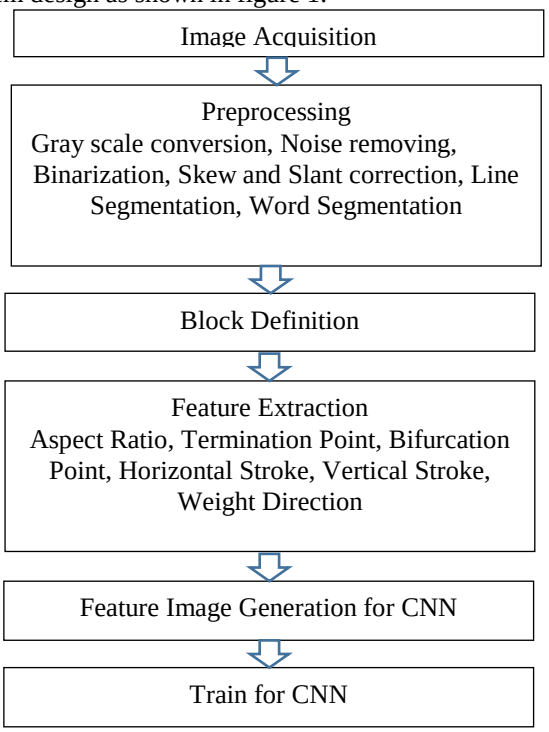


Fig. 1 Proposed system design for Myanmar old printed characters recognition system

4.1 Preprocessing

Preprocessing produces data that are easy for the OCR systems to recognize accurately. Firstly, the input image converts into grayscale and then noise is removed by using median filter. Next, binarization method is applied to convert grayscale into two tones. For binarization, proposed system computes threshold value using local thresholding methods. This method generally performs better for low quality images. The threshold is set at the midrange value, which is the mean of minimum $Seg_{(i,j)}$ and maximum $Seg_{(i,j)}$ of gray values. The binarized image is equally segmented into 6 columns and 20 rows. So, the window size has been set to 6x20. According to the resolution of image the window size can be changed.

$$TS_{(i,j)} = \frac{\max(Seg_{(i,j)}) + \min(Seg_{(i,j)})}{2} \quad (1)$$

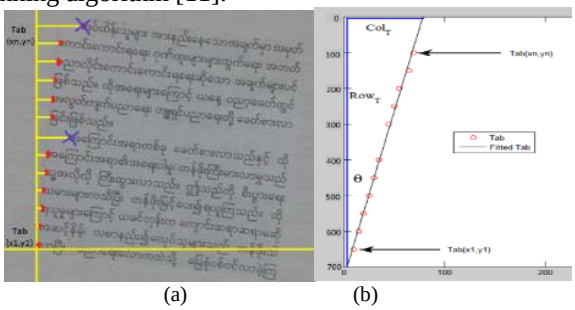
Where, = Threshold value of gray scale image at coordinate (i,j)

max $Seg_{(i,j)}$ = Maximum pixel value of segment (i,j)

min $Seg_{(i,j)}$ = Minimum pixel value of segment (i,j)

4.2 Skew and Slant Correction of Document Image

Distortion alignment of input text, scanned or photo that needs to align it by performing skew and slant angle correction method. Skew correction can be done by (i) estimating the skew angle, and (ii) rotating the image by the skew angle in the opposite direction. In Figure 2, the red point is calculated by using horizontal projection profile method. A series of horizontal projection profiles are obtained at a number of angles close to the expected orientation. This angle is skew angle. The skew angle is calculated by (2). After auto-rotation, removed some unnecessary pixels of binary image using morphological thinning algorithm [11].



(a) (b)

Fig. 2(a) Red points define the left boundary (b) Find the slope with horizontal projection profile

$$\theta = \tan^{-1}(\frac{Y}{X}) \quad (2)$$

4.3 Line segmentation

Lines are separated by using horizontal projection profile method. To detect the lines, assume that the value of the element in the r[th] row and the c[th] column of the character matrix is given by a function: $f(r, c) = BW$. Where, BW takes binary values (i.e., 0 for background black pixel and 1 for white pixel). Horizontal projection HP of the character matrix is calculated by the sum of white pixels in each row.

$$HP(r) = \sum_{c} BW(r,c) \quad (3)$$

And separate the lines on the HP values as shown in figure 3.

4.4 Word segmentation

Words segmentation from separated lines is divided by using vertical projection profile method. Similarly, the vertical projection VP of the character matrix is calculated by the sum of white pixels in each column of the line segment.

$$VP(c) = \sum_{r} BW(r,c) \quad (4)$$

And isolate words on the VP values as shown in figure 4.

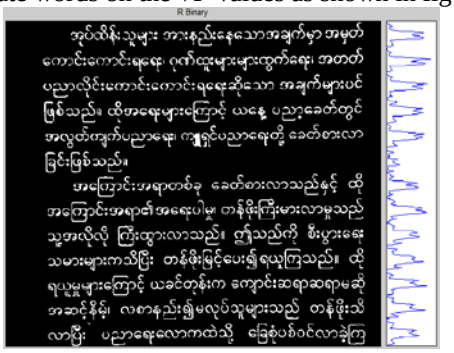


Fig. 3 Horizontal projection for line segmentation

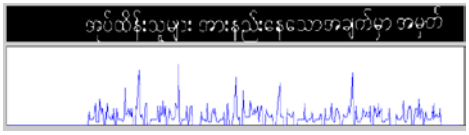


Fig. 4 Vertical projection for word segmentation

5. Feature Extraction

In this proposed system, six structural features will be extracted such as aspect ratio, how many termination points, bifurcation points, horizontal strokes, vertical strokes and which the weight direction of character turn to.

5.1 Aspect Ratio

Aspect ratio of character is proportional relationship between its width and its height. Aspect ratio is important feature of character recognition.

Fig. 5 Samples of aspect ratio for Myanmar character 'KA' and 'SA'

5.2 Termination Point

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 6 Issue 4, April 2019
ISSN (Online) 2348 – 7968
www.ijiset.com

Termination feature is one of most important feature of character recognition process. Most of Myanmar characters included termination points. Binary morphology mask of 3x3 structuring element is used for termination point detection. Mathematical model of termination point detection is described in equation 5.

$$SN = \sum_{i=1}^{8} N_i$$

$$TP_{(i,j)} = \begin{cases} 1, \text{if } SN_{(i,j)} = 1, \\ 0, \quad \text{otherwise} \end{cases} \tag{5}$$

Where, SN is the sum of neighborhood pixels and TP is termination point.

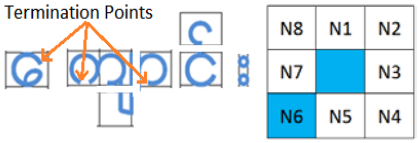


Fig. 6 Termination points detection

5.3 Bifurcation Point

Bifurcation feature is also one of most important feature of character recognition process. Some of Myanmar characters included bifurcation features. Binary morphology mask of 3x3 structuring element is used for bifurcation point detection. Mathematical model of bifurcation point detection is described in equation 6.

$$BP_{(i,j)} = \begin{cases} 1, \text{if } SN_{(i,j)} = 3, \\ 0, \quad \text{otherwise} \end{cases} \tag{6}$$

Where, BP is the bifurcation point of binary image.

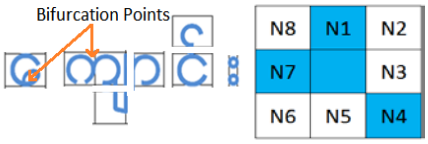


Fig. 7 Bifurcation points detection

5.4 Horizontal Stroke Point

Horizontal stroke is counted the number of binary points in mid row of isolated character. Horizontal stroke is also important feature of character recognition and counted by using equation 7.

$$HS = \sum_{j=1}^{n} IM_{(mr,j)} \tag{7}$$

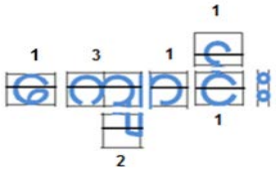


Fig. 8 Horizontal stroke points detection

5.5 Vertical Stroke Point

Vertical stroke is counted the number of binary points in mid column of isolated character and calculated by using equation 8.

$$VS = \sum_{i=1}^{m} IM_{(i,mc)} \tag{8}$$

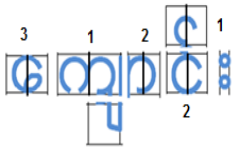


Fig. 9 Vertical stroke points detection

Where, HS is total stokes of horizontal mid row and mr is mid row of isolated character IM . VS is total strokes of mid column and mc is mid column of IM .

5.6 Weight Direction

Weight direction is detection of weight of character that the most number of pixels included in which quadrants. Mathematical form of weight direction detector is equation 9.

$$W\max = \max(Wa, Wb, Wc, Wd) \tag{9}$$

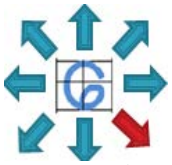


Fig. 10 Weight direction feature detection

6. Trained in Convolutional Neural Network

The six features of a character in the feature extraction stage that is used for classification. The proposed system used convolutional neural network (CNN) for classification. Input character image with various features such as termination, bifurcation, horizontal and vertical stroke points that is trained in CNN. In figure 11 the red

points that are stroke features, blue points are bifurcations and green points are termination features. Aspect ratio and weight direction features are applied in block definition stage [16]. CNN are a special kind of multi-layer neural networks. CNN are designed to recognize visual patterns directly from pixel images with minimal preprocessing. They can recognize patterns with extreme variability (such as handwritten characters).

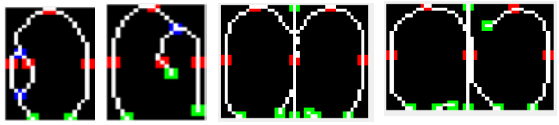


Fig. 11 Sample features image of Myanmar Character 'sa', 'za', 'ta' and 'la' that is trained in CNN

7. Experimental Results

Myanmar old printed documents around 1938 is used as the training dataset and testing data of the system. The experiments have been conducted on a dataset of 60000 Myanmar characters for training. An independent dataset portion of 20000 Myanmar characters in old printed documents that is used to calculate the recognition performance. All input character images are normalized to size 32*32 after computing their bounding rectangles. The accuracy of recognized character with features and without feature are described in table 2 and 3.

Table 2: Recognized Results with Six Features

Number of Document Images	Contained Characters	Miss Recognized Characters	Accuracy
Document 1	172	3	98.2558%
Document 2	212	3	98.5849%
Document 3	182	3	98.3516%
Document 4	193	9	95.3368%
Document 5	152	3	98.0263%
Document 6	165	1	99.3939%
Document 7	184	2	98.913%
Document 8	158	5	96.8354%
Document 9	166	5	96.988%
Document 10	199	3	98.4925%
Document 11	185	6	96.7568%
Document 12	157	3	98.0892%

Table 3: Recognized Results without Features

Number of Document Images	Contained Characters	Miss Recognized Characters	Accuracy
Document 1	172	15	91.279%
Document 2	212	25	88.207%
Document 3	182	20	89.010%
Document 4	193	15	92.227%
Document 5	152	12	92.105%
Document 6	165	15	90.909%
Document 7	184	20	89.130%
Document 8	158	18	88.607%
Document 9	166	15	90.963%
Document 10	199	22	88.944%
Document 11	185	18	90.270%
Document 12	157	15	90.445%

8. Conclusion

There are many feature extraction techniques which are not implemented in case of Myanmar characters recognition. This paper has presented about Myanmar old printed character recognition in general and specifically concentrating on feature extraction. In this system, image preprocessing steps such as binarization, skew correction and segmentation method are suitable for Myanmar old printed documents according to the experimental results. The six structural features extracted from each character is appropriated to use for Myanmar old printed character recognition system to get better accuracy. Therefore, selection of proper feature extraction technique is the key issue in OCR system.

References

[1] R. K. Chadha, N. Mehta, "A Study of Handwritten Character Recognition Techniques", International Journal of Advanced Research in Computer and Communication Engineering, India, Vol. 5, Issue 1, January 2016.
[2] T. R. Zalke, V. N. Bhonge, "An Optical Character Recognition for Handwritten Devanagari Script", Journal of

Engineering Research and Applications, ISSN : 2248-9622, Vol. 5, Issue 1, India, January 2015, pp.120-123.
[3] G.Vamvakas, B.Gatos, N. Stamatopoulos, and S.J.Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents" , The Eighth IAPR Workshop on Document Analysis Systems, 978-0-7695-3337-7/08 $25.00 © 2008 IEEE.
[4] T. Htike and Y. Thein,"Handwritten Character Recognition Using competitive Neural Trees," IACSIT International Journal of Engineering and Technology, Vol. 5, No. 3, University of Computer Studies Yangon (UCSY) Yangon, Myanmar, June 2013.
[5] Y. Thein, S. S. S. Yee, "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, University of Computer Studies Yangon (UCSY) Yangon, Myanmar, November 2010.
[6] Y. Y. Than, D. M. Aung, A. M. Yi and K. T. Win, "Development Of Handwritten Myanmar Alphabet Recognition", International Journal of Video & Image Processing and Network Security IJVIPNS-IJENS Vol:09 No:10, University of Computer Studies, Mandalay, Myanmar, December 2009.
[7] H. P. P. Win and K. N. N. Tun, "Converting Myanmar Printed Document Image into Machine Understandable Text Format", IEEE, University of Computer Studies, Yangon, September 2011.
[8] G.Vamvakas, B.Gatos, N. Stamatopoulos, and S.J.Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents", The Eighth IAPR Workshop on Document Analysis Systems, IEEE, 2008.
[9] S. S. S. Yee and Y. Thein, "High Accuracy Myanmar Handwritten Character Recognition using Hybrid approach through MICR and Neural Network". IJCSI International Journal of Computer Science, Vol.7, Issue 6, November 2010.
[10] W. Bieniecki, S. Grabowski and W. Rozenberg, "Image Preprocessing for Improving OCR Accuracy", MEMSTECH'2007, Lviv-Polyana, UKRAINE, May 23-26, 2007.
[11] W. Abu-Ain, S. N. H. S. Abdullah and K. Omar, "A Simple Iterative Thinning Algorithm For Text and Shape Binary Images", Journal of Theoretical and Applied Information Technology, Vol. 63, No.2, 20th May 2014.
[12] A. Sahlol and C. Suen, "A Novel Method for The Recognition of Isolated Handwritten Arabic Characters".
[13] W. Suparta and K.M. Alhasa, "Modeling of Tropospheric Delays Using ANFIS", SpringerBriefs in Meteorology, DOI 10.1007/978-3-319- 28437-8_2.
[14] M. Billah, S. Waheed and A. Hanifa, "An Optical Character Recognition System from Printed Text and Text Image using Adaptive Neuro Fuzzy Inference System", International Journal of Computer Applications (0975 - 8887), Volume 130 - No.16, November 2015.
[15] A. M. Ardhian, S. P. Hadi, M. I. BS, "Membership Function Comparative Study on Load Forecasting using ANFIS Framework", International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016.
[16] Z.Z.Aung and CMMMaung,"Myanmar Optical Character Recognition using Block Definition and Featured Approach", 3rd International Conference on Science in Information Technology (ICSITech), 978-1-5090-5864-8/17/$31.00 ©2017 IEEE.

First Author I received M.C.Sc (Master of Computer Science) degree in 2010 and now I am PhD (I.T) research student from UTYCC (University of technology Yatanarpon Cyber City). My first paper is "Myanmar Optical Character Recognition using Block Definition and Featured Approach" for ICSITech conference in 2017. My research interests include image processing, adaptive neuro inference system, convolutional neural network and Myanmar character recognition field.

Second Author She received ME (IT) degree from Yangon Technological University in 2006 and PhD (I.T) degree in 2009. She is now professor of UTYCC (University of Technology, Yatanarpon Cyber City). Her first paper is "An efficient communication mechanism for cluster computing" for GMSARM conference in 2008. Her research interests include Image Processing, Computer Architecture and Parallel Processing field.

Third Author She received PhD (IT) degree from University of Technology (Yatanarpon Cyber City) in 2015. She is now Lecturer in UTYCC.