INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 4)
Available online at: www.ijariit.com

Literature review of Myanmar Optical Character Recognition techniques

Zu Zu Aung
zuzuaungucsm@gmail.com
University of Technology, Yadanabon Cyber City, Myanmar

Cho Me Me Maung
chomememaung@gmail.com
University of Technology, Yadanabon Cyber City, Myanmar

Yadana Htun
yadanahtunutycc@gmail.com
University of Technology, Yadanabon Cyber City, Myanmar

ABSTRACT

Optical Character Recognition (OCR) is an active research area in the field of pattern recognition due to its application. OCR system converts document images into machine-encoded text so that it can be easily accessed and preserved. This system is applied in data processing such as bank cheque, postal address reading, examination question papers, text-to-speech, reading aid for the blind, etc. Myanmar OCR system is essential to convert numerous published books, newspapers and journals of Myanmar into editable computer text files. It is a challenge for recognizing Myanmar characters. During the last decades, a lot of research has been done in the field of Myanmar optical character recognition (OCR). Myanmar characters recognition is still an open problem for the research community. This paper presents the review of Myanmar OCR system such as pre-processing, feature extraction, classification, post-processing etc. This paper may act as a support material for those who wish to know about Myanmar OCR.

Keywords— *Myanmar OCR, Pre-processing, Feature extraction, Classification, Post-processing*

1. INTRODUCTION

Character recognition is a process of recognizing characters from input text images and converts it into ASCII or machine editable form. The text may be in the form of scanned handwritten documents or machine-printed documents. OCR system improves the interface between man and machine in many applications. 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, other extended characters and a lot of Pali characters as shown in figure (1). Myanmar script is round script is written from left to right.

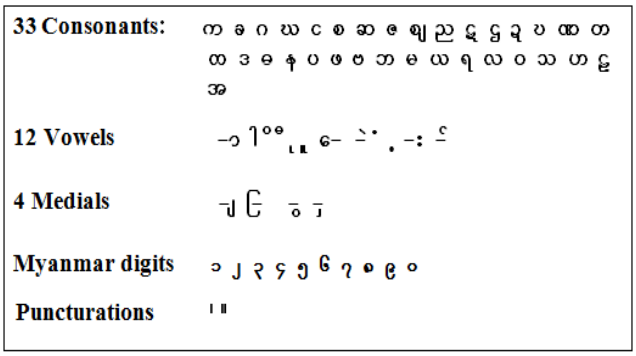

Fig. 1: Myanmar character set

2. MAJOR STEPS IN OCR

The basic OCR system consists of the following steps:
1. Pre-processing
2. Segmentation
3. Feature extraction
4. Recognition
5. Post-processing

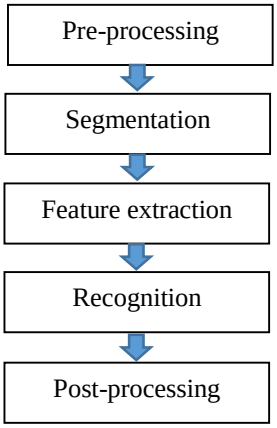

Fig. 2: Block diagram of OCR system

3. LITERATURE SURVEY

In [1], they have performed Handwritten Myanmar Alphabet Recognition. In this system consists of seven preprocessing steps (1. Scanning 2. Binarization 3. Segmentation 4. Noise Removing 5. Delete Extra Space 6. Normalization 7. Thinning). In binarization, global thresholding method is used then images were converted to binary with threshold value 0.85. Histogram-based segmentation method is applied to segment line and character. In deleting extra space, the image is cropped to get the smaller image by using horizontal and vertical projection profile. Image size is assigned (30*64) in the normalization step. An object in a digital image is reduced by applying a thinning method. In feature extraction, zoning method is used. A normalized image is partitioned into 4*3 zones. The feature vector is formed by the sub-block images. Each block is determined by using a structuring element whether there is data or not. In each zone, when the total number of pixels is less than 7 the system will fill with value 0. In this paper, five Myanmar alphabets such as Kagyi, Kakway, Gange, Gagyi, and Nga are used as the input image. The rule-based method is used for recognition. Five rules are applied to recognize. This paper tested with handwritten Myanmar alphabet recognition (HMAR) for the result of 98.8% recognition accuracy rate on 375 training data and 125 testing data.

In [2] have proposed Myanmar Printed Document Image recognition, they worked low pass Finite State Impulse Response (FIR) filter is used for noise removing. By applying the Otsu method, the clean image is binarized. For deskewing, the Hough transform method is utilized. They used X_Y cut method on the use of histogram or a projection profile technique for segmentation. Before the feature extraction stage, the binary character image is normalized into (30*30) character image. They extracted two types of statistical features. The first one applied zoning method and calculates the density of the character pixels in each zone. The second features used projection method which divides top, middle and bottom as well as left, right and center character. 25 features are used in zoning method and 60 features are used in projection profile method. SVM is used as a recognizer. The hierarchical mechanism is constructed to overcome the similar problems of the Myanmar script. In a post-processing step, perceived characters are changed over to ASCII design. In the experiment, 98.89% segmentation accuracy rate and 97.25% recognition accuracy rate were obtained.

Myanmar handwritten character recognition based on Competitive Neural Trees (CNet) proposed in [3]. They worked the steps of binarization, thinning and resizing in preprocessing. Region-based methods are used to extract features. Eighteen features are extracted. In the recognition stage, Competitive Neural Trees (CNeT) is used. CNeT consists of five steps: CNeT learning phase, the life cycle of CNet, the training procedure of CNet, recall procedures and global search method. In this paper, 33 alphabets of handwritten Myanmar character are used as input and 330 alphabets are tested. Recognition accuracy rate of 97% was obtained.

In [4], this paper presents an automatic data entry system of passport for a security system. Skewing, resizing, filtering, grayscale converting, normalization steps are done in pre-processing steps. For segmentation, a region-based segmentation method is used. Thinning and skeleton methods are extended for feature extraction. In the experiment, fifty Myanmar passports are used. Handwritten characters are recognized with high data accuracy rate.

Recognition on User-Entered Data from Myanmar Bank Cheque is implemented in [5]. In this paper, they have employed a model to recognize the payee's name and legal amount, the courtesy amount on a bank cheque. For preprocessing, the noise removing, cropping, skew detection, thinning and normalization method are applied. In the feature extraction step used MWR algorithm. They classified Myanmar character into many groups depend on the nature of its writing style. Hidden Markov Model is trained using these feature an experiment is carried out.

In paper [6] and [7] implemented Intelligent Character Recognition (ICR) and Optical Character Recognition (OCR) technology through Myanmar Intelligent Character Recognition (MICR) and some neural networks. Grayscale converting, noise filtering, binarization, row column extraction, resizing and normalization are involved in preprocessing. In this paper, MICR is developed from the ICR system and recognizes both on-line and off-line characters. Statistical and semantic information is used for features extraction. The statistical approach looks for the pixel values for each character such as the ratio of the black pixels to white pixels, width and height ratio, pixel density, histogram, etc. And semantic approach consists of black stroke count, endpoints, and connection points. MICR is composed of statistical and semantic information from each character. These features are used to train the neural network. The final decision is made by the voting system. The input images of English and Myanmar character are tested in ICR and OCR. Backpropagation neural network is used for recognition. The input layer has 165 neurons equal to a number of pixels, the hidden layer has 100 neurons and the output layer has 6 neurons. In post-processing, recognized characters are converted to UNICODE or ASCII format. The proposed system achieved the best recognition and accuracy rates.

Table 1: Myanmar OCR with various Methods

Title	Methods	Shortcomings	Author
Handwritten Character Recognition Using Competitive Neural Trees	▪ Global search method. ▪ CNet	They can only show 97% accuracy for 33 consonants of Myanmar handwritten characters are recognized and not yet done for all Myanmar alphabets and compound words.	T. Htike and Y. Thein 2013 [3]
Hybrid of ICR/OCR technology through MICR and Neural Network	▪ Statistical and semantic information of MICR used to extract as the features. ▪ Back-propagation neural network for recognition.	They can show 94% accuracy for Myanmar handwritten characters but experiment Myanmar compound handwritten characters are a little different with normal Myanmar handwritten characters in the real world.	Z. L. Phyu, Y. M. Aung, E. P. Min, Y. Thein 2010 [6]
Development Of Handwritten Myanmar Alphabet Recognition	▪ Feature extraction methods based on zoning method. ▪ The rule-based recognition system is used for Myanmar alphabet.	They can only show the 98.8% accuracy for first five Myanmar alphabets and not yet done for all Myanmar alphabets and compound words.	Y. Y. Then, D. M. Aung, A. M. Yi and K. T. Win 2009 [1]
Converting Myanmar Printed Document Image into Machine Understandable Text Format	▪ X_Y cut method on the use of histogram. ▪ A projection profile technique for segmentation/feature extraction. ▪ SVM is used to recognize.	They can show the 97% accuracy for Myanmar printed documents and the system needs to recognize for historical documents.	H. P. P. Win and K. N. N. Tun 2011 [2]
Hand Written Recognition System for Automatic Data Entry of Passport	▪ Region-based segmentation method. ▪ Thinning and skeleton methods are extended for feature extraction. ▪ Gaussian Elimination method for recognizing	They can only recognize English character A to Z and 0-9 digits.	M. M. Thinn, M. M. Sein [4]
Recognition on User-Entered Data from Myanmar Bank Cheque	▪ In the feature extraction used MWR algorithm ▪ Hidden Markov Model (HMM) is applied for recognition.	They can only implement the payee's name and legal amount, the courtesy amount on bank cheque and not yet done for all Myanmar handwritten characters in bank cheque.	N. A. A. Htwe, S. S. Mon, M. M. Sein [5]

4. CONCLUSION

This paper presents a review of Myanmar OCR systems such as pre-processing, segmentation, feature extraction, and recognition. Each and every step contributes directly to the accuracy of the system. Each technique has its own uniqueness and level of accuracy, but still some modifications have to be done for characters of different size and fonts. Myanmar handwritten character recognition and Myanmar printed character recognition is still a research area of pattern recognition.

5. REFERENCES

[1] Y. Y. Then, D. M. Aung, A. M. Yi and K. T.Win, "Development of Handwritten Myanmar Alphabet Recognition", University of Computer Studies, Mandalay, Myanmar, International Journal of Video & Image Processing and Network Security IJVIPNS-IJENS Vol:09 No:10.

[2] H. P. P. Win and K. N. N. Tun, "Converting Myanmar Printed Document Image into Machine Understandable Text Format", University of Computer Studies, Yangon, Myanmar, https://www.researchgate.net/publication/221254210.

[3] T. H. and Y. Thein "Handwritten Character Recognition Using Competitive Neural Trees", *IACSIT International Journal of* Engineering and Technology, Vol. 5, No. 3, June 2013.

[4] M. M. Thinn and M. M. Sein, "Hand Written Recognition System for Automatic Data Entry of Passport", Mandalay Technological University, Myanmar.

[5] N. A. A. Htwe, S. S. Mon and M. M. Sein "Recognition on User-Entered Data from Myanmar Bank Cheque", Mandalay Technological University, University of Computer Studies, Yangon, Myanmar.

[6] Z. L. Phyu, Y. M Aung, E. P. Min, and Y Thein, "Hybrid of ICR/OCR technology through MICR and Neural Network", University of Computer Studies, Yangon, Myanmar.

[7] Dr. Y. Thein and San 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, November 2010.
[8] R. K. Chadha and 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.
[9] T. R. Zalke and 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.
[10] 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.

BIOGRAPHY

Zu Zu Aung
Assistant lecturer
Ph.D. (I.T.) Research student, University of technology Yatanarpon Cyber City, Myanmar

Cho Me Me Maung
Professor, University of Technology, Yadanabon Cyber City, Myanmar
Ph.D. (I.T.) 2009

