Analyzing Word Error Rate on Optical Character Recognition (OCR) for Myanmar Printed Document Image

[1]Thin Thin Hlaing, [2]May Phyo Oo, [3]Thaint Zarli Myint
[1]*FCST Dept: Computer University,* [2]*Computer University ,* [3]*FCST Dept: Computer University*
(Banmaw) Myanmar

Abstract— *The printed document is used Myanmar language in Myanmar. Sometime, we want to convert this printed document to text document easily. So, this paper describes an effective recognition and calculate error rate for Myanmar printed document image to editing text. Myanmar language contains many words, and most of them are similar, especially for small fonts, the accuracy of the Optical Character Recognition, OCR system for Myanmar may be low. In order to get more accurate system, enhance the input image by removing noise and making some correction on variants. A method for isolation of the character image is proposed by using connected component analysis for wrongly segmented characters produced by projection only. So, this paper proposes a method for obtaining more detail about actual translation errors in the generated output by using word error rate (WER) based the neural network classifier for recognition of the character image. We investigate the use of WER for automatic error analysis using a dynamic programming algorithm like Levenshtein distance over segmentation. This paper gives a better overview of the nature of translation errors. Finally, the proposed algorithms have been tested on a variety of Myanmar printed documents and the results of the experiments indicate that the methods can reduce the segmentation error rate as well as translation rates.*

Index Terms— *Neural Network, OCR, Printed Document, WER*

I. INTRODUCTION

In all over the world, there are different techniques that can be used to recognize characters. Among them, Optical Character Recognition and Intelligent Character Recognition are two basic techniques for character recognition. OCR typically involves the process of translating digitized images of text (usually created by a scanner) into a machine-readable format (such as ASCII or Unicode). Myanmar language is the main language and Myanmar printed document are widely used by over 85 % of 49.5 million populations in Myanmar although there are more than 8 different languages used in Myanmar. Myanmar printed document are mostly used all of the textbook in education. All various fields of the documents, magazines, reports and technical papers can be converted to electronic form using a high performance Optical Character Recognizer (OCR). And optical character recognition is a key enabling technology critical to creating indexed, digital library content, and it is especially valuable for scripts, for which there has been very little digital access.

With the increasing demand for creating a paperless world, many OCR algorithms for English and other developed countries’ languages have been developed over the years and these can be available commercially or freely. But, development of an optical character recognition system for Myanmar languages is in little effort. This is because Myanmar (Burmese) scripts are rich in patterns while the combinations of such patterns make the problem even more complex and hence the motivation to work further in this area. Myanmar scripts, derived from Brahmi scripts, also present some challenges for OCR that are different from those faced with Latin and Oriental scripts. But properly utilized, OCR will help to make Burmese digital archives, practically accessible to local users and lay users alike by creating searchable indexes and machine-readable text repositories. In this system, feed-forward neural network and backpropagation learning algorithm is also used for calculating error rate. The recognition performance of the Back-propagation network will highly depend on the structure of the network and training algorithm.

In this paper, Optical Character Recognition System for Myanmar Printed Document is presented with a variety of proposing techniques, including a novel segmentation method to truly separate Myanmar characters, efficient Feature extraction method using zone and projection profile for isolated character data and the powerful neural network classifier to recognize Myanmar script features. These works are need to exert much effort to come up with better and workable OCR technologies for the local scripts in order to satisfy the need for digitized information processing.

The rest of the paper is organized as follow. Section 2 introduces the nature of Myanmar script. Section 3 presents the previous work as the background theories. Section 4 gives more details on the implementation of recognition system. Results are discussed in Section 5 and Section 6 is the conclusion.

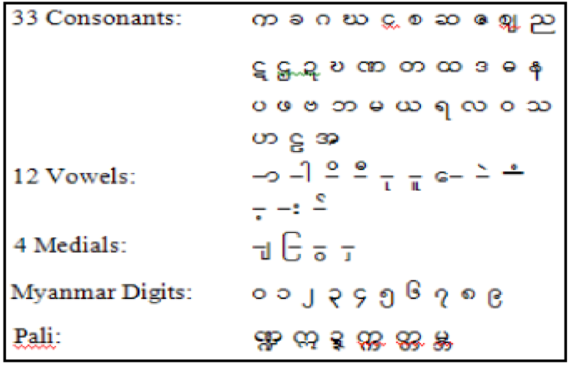
II. BACKGROUND THEORY

A. History of Myanmar Language

The Myanmar language belongs to the Sino-Tibetan family of languages of which the Tibetan-Myanmar (Tibeto-Burman) subfamily forms a part. It has been classified by linguists as a monosyllabic or isolating language with agglutinative features. It is a tonal and analytic language. There are different types of language in Myanmar such as Myanmar, Karen, Rakhine, Chin, Mon, Shan, etc. But, Myanmar language is the mother language in Myanmar. The Myanmar language is the official language of Myanmar and is more than one thousand years old. Texts in the Myanmar language use the Myanmar script, which derives from a Brahmi related script borrowed from South India in about the eight centuries for the Mon language. The first inscription in Burmese dates from the following years and is written in an alphabet almost identical with Mon inscriptions. The earliest Myanmar and Mon language can be seen in MyaZayDi Stone inscription.

B. Myanmar Language Characteristic

Myanmar alphabet consists of 33 consonants, 12 vowels, 4 medial and 10 digits. In Pali alphabet consists of 41 letters: (8) vowels and (33) consonants. The consonants in Pali can be grouped into aspirated and non-aspirated consonants, as shown in Fig.1.



33 Consonants:	က ခ ဂ ဃ င စ ဆ ဇ ဈ ည ဋ ဌ ဍ ဎ ဏ တ ထ ဒ ဓ န ပ ဖ ဗ ဘ မ ယ ရ လ ဝ သ ဟ ဠ အ
12 Vowels:	ာ ါ ိ ီ ု ူ ေ ဲ ံ ့ း ်
4 Medials:	ျ ြ ွ ှ
Myanmar Digits:	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉
Pali:	ဿ ဣ ဍ္ဍ ဏ္ဍ ဏ္ဌ ဋ္ဌ

Fig. 1 Pattern of Myanmar Alphabet

C. Nature of Myanmar Scripts

Myanmar (Burmese) script is recognized as Tibet/Bur man language group, developed from the Mon script and descended from the Brahmi script of ancient South India. It is the official language of Myanmar, where over 35 million people speak it as their first language. Some people in China and India also speak Burmese. The code range is 1000 – 109F according to Unicode Standard, version 3.0, August 2000. The direction of writing is from left to right in horizontally. In Myanmar script, there is no distinction between Upper Case and Lower-Case characters. The character set consists of 35 consonants (including '□'), 8 vowels signs, 7 independent vowels, 5 combining marks, 6 symbols and punctuations, and 10 digits. Each word can be formed by combining consonants, vowels and various signs. It has its own specified composition rules for combining vowels, consonants and modifiers. There are total of above 1881glyphs and has many similarity scripts in this language (e.g., □, □, □, □, □, □ and so on). 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.

D. Levenshtein distance

Levenshtein distance (LD) is a measure of the similarity between two strings, the source string (s) and the target string (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t. The greater the Levenshtein distance, the more different the strings are. In our case, the source string is the input, and the target string is one of the entries in the dictionary. The Levenshtein algorithm (also called Edit-Distance) calculates the least number of edit operations that are necessary to modify one string to obtain another string. The most common way of calculating this is by the dynamic programming approach. A matrix is initialized measuring in the (m, n)-cell the Levenshtein distance between the m character prefix of one with the n-prefix of the other word. The matrix can be filled from the upper left to the lower right corner. Each jump horizontally or vertically corresponds to an insert or a delete, respectively. The cost is normally set to 1 for each of the operations. The diagonal jump can cost either one, if the two characters in the row and column do not match or 0, if they do. Each cell always minimizes the cost locally. This way the number in the lower right corner is the Levenshtein distance between both words. An example of finding LD between the user speech "□□□ " and the recognition output "□□-□□ " is shown in Fig. 2 for WER.

		□	□	-□	□
□	0	1	2	3	4
	1	2,2,0	1,3,2	2,4,3	3,5,4
□□	2	0	1	2	3
		3,1,2	2,2,0	1,3,2	2,4,3
□	3	1	0	1	2
		4,2,3	3,1,2	2,2,1	2,3,1
		2	1	1	1(min distance)

Fig. 2 Example of Levenshtein Distance

E. Neural Network

In this system, feed-forward neural network and back propagation learning algorithm is used. The recognition performance of the Back-propagation network will highly depend on the structure of the network and training algorithm. It consists of three layers forward structure that has hidden layer between input layer and output layer interconnected by links that contains weights. Fig. 3, shows the architecture of network.

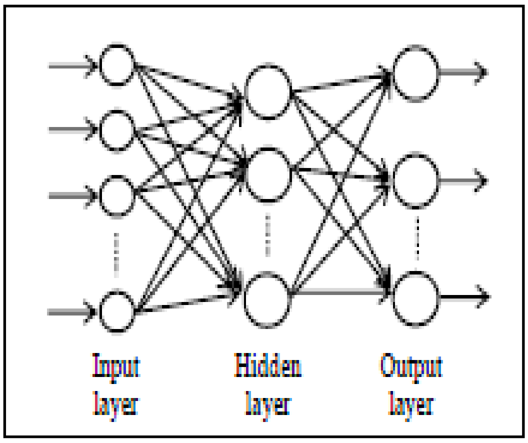


Fig. 3 Neural Network Architecture

Its input has two forms: features extraction of image and pixels of image. Training time will be very long if pixels are used as an input for neural network. In this system, statistical and semantic information of OCR has been used the inputs of neural network to save the training time.

III. PROPOSED METHOD

As other traditional OCR systems, the proposed system also includes five processing steps as shown in Fig. 4. It has 6 different types of documents written in Zawgyi-One font and font size 12 are taken to test the system. These are scanned on a flatbed scanner at 300 dpi for digitization go for the preprocessing steps.

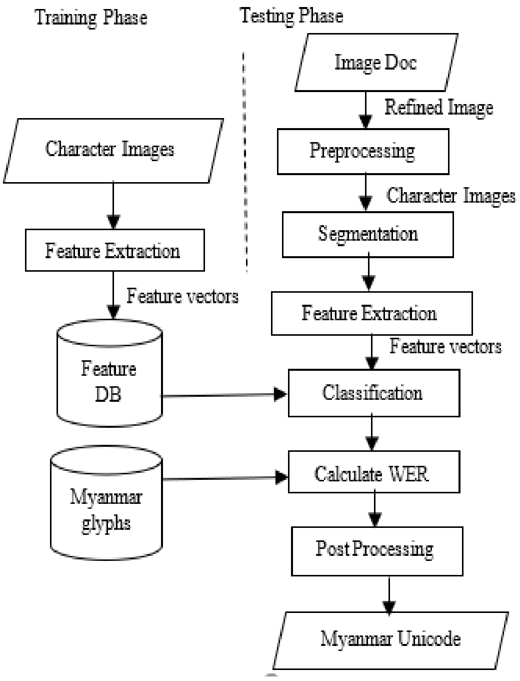


Fig. 4 System Design of the Myanmar OCR system

A. Preprocessing

Preprocessing step is the basic crucial part of the OCR system. The recognition accuracy of OCR systems greatly depends on the quality of the input text image. Firstly, we convert the raw input image into grayscale and then denies it by removing noise using low pass Finite State Impulse Response (FIR) filter. Next, we binarize the clean image to a bi-level image by turning all pixels below some threshold to zero and all pixels about that threshold to one. We find this threshold value using Otsu method. Finally, we desked the binarized image with generalized Hough Transformed method. The detailed of the preprocessing steps are described in [10].

B. Segmentation

Segmentation is the process of the isolation of the individual character images from the refined image. It is considered as the main source of the recognition errors especially for small fonts. This is one of the most difficult pieces of the OCR system [4]. We use the X_Y cut method on the use of histogram or a projection profile technique for segmentation. It has been proven as a classical and more accurate method in vagary scripts such as Bangla and Hindi and some of the South East Asia scripts, English and some Greek OCR [7], [10]. The process of segmentation in our system mainly follows the following pattern:

- Line detection and slicing
- Character Segmentation

1) *Line Detection and slicing:* To detect the lines, assume that the value of the element in the xth row and the y th column of the character matrix is given by a function f:

$$\boldsymbol{f(x,y) = a_{xy}} \quad (1)$$

Where $\boldsymbol{a_{xy}}$ takes binary value (i.e.,) for background white pixels and 1 for black pixels). The horizontal histogram $\boldsymbol{H_h(x)}$ values, as shown in Fig. 5.

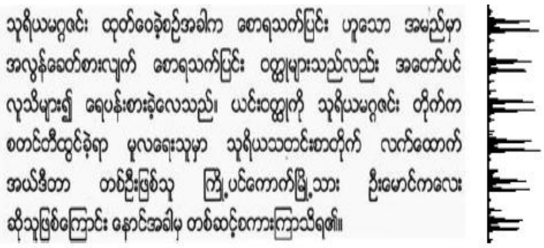


Fig. 5 Example of line segmentation using projection

2) *Character Segmentation:* Similarly, the vertical histogram H_V of the character matrix is calculated by the sum of black pixels in each column of the line segment:

$$H_v(y) = \sum_x f(x,y) \quad (2)$$

Characters are segmented using these histogram values. However, this method alone is not enough for the Myanmar

scripts. As for the small font, some character is not correctly segmented as shown in Fig. 6.

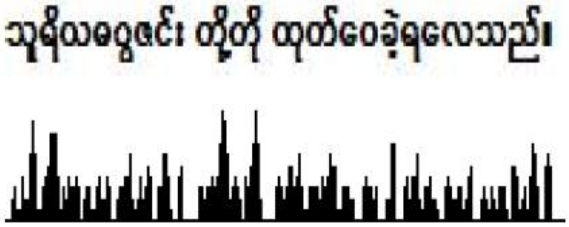


Fig. 6 Wrong Segmentation Error with Projection

And it may also be problem for some connected components. Moreover, the connected components can't extract earlier as other languages because it can appear not only in shorter segments but also in longer segments that of the line height. That's why the nature of Myanmar scripts cause over segmentation and under segmentation problems. To overcome overlaps and wrong segmentation cases, assume the points from (2) as the pre-segment points and we need check the possible points according to line height.

C. Feature Extraction

Before the extraction of features we need to normalize the binary character images to have the standard width and height. We normalize all character images height into N and the equal amount is used for width with respecting the original aspect ratio. Feature extraction involves extracting the attributes that best describe the segmented character image as a feature vectors. This process maximizes the recognition rate with the least number of elements [5].

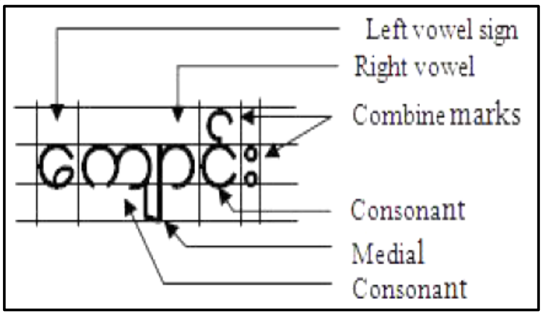


Fig.7 Sample of Myanmar Glyphs

In our approach we employ two types of statistical features. The first one divides the character image into a set of zones and calculates the density of the character pixels in each zone as in [15]. The Myanmar characters are written into three main zones for horizontal and the minimum component for a truly segmented glyph is one and the maximum component may be four as shown in Fig 8. Therefore, we considered for the second type of features, the area that is formed from the projections of the top, middle and bottom as well as of the left, center and right character profiles is calculated.

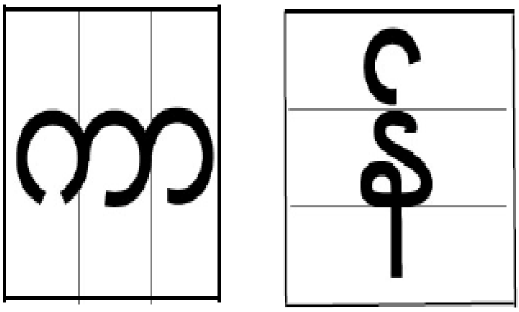


Fig. 8 Division of each character depend on writing nature

Let g(x, y) be the binary image array and w,h be the width and height of the segmented character. In the case of features based on zones, the image is divided into equal zones. For each zones, we calculate the density of the character pixel as follow:

$$\boldsymbol{F_z(n) = \sum g(x,y), n = 0 \ldots Z_{max} - 1} \quad (3)$$

Where, x, y be the pixel point in each zone.

When we consider features based on vertical profile projections, the character image is divided into $S_v$sections separated by the horizontal lines of y and calculated as follow:

$$\boldsymbol{y_i = i(h/S_V) - 1, i = 1, \ldots, S_V - 1} \quad (4)$$

And for each section, we equally divide into blocks and calculate y_t, the distance between the base line and outermost pixel depending on the direction we considered as follows:

$$y_s = \begin{cases} y_i - y_p, & for\ buttom\ to\ top \\ y_p - y_{i-1}, & for\ top\ to\ buttom \end{cases}$$

(5)

Where,$y_p$is the outermost pixel value of 1 and F_v be the total number of blocks to produce the vertical profiles and calculate the feature for each block as follow:

$$F_v(n) = \sum y_s(x), n = Z_{max} , \dots, Z_{max} + F_v - 1$$

(6)

And for each section, we equally divide into blocks and calculate x_s , the distance between the base line and outermost pixel depending on the direction we considered as follow:

$$xy_s = \begin{cases} x_i - x_p, & for\ right\ to\ left \\ x_p - x_{i-1}, & for\ left\ to\ right \end{cases}$$

(7)

Where, x_s is the outermost pixel value of 1 and F_h be the total number of blocks to produce the horizontal profiles and calculate the feature for each block as follow:

$$F_h(n) = \sum x_s(y), n = Z_{max} , \dots, Z_{max} + F_v + F_h - 1$$

(8)

Therefore, the total feature for each character image is

$$F_{total}(n) = F_z(n) + F_v(n) + F_h(n)$$

(9)

D. Classification

This process is responsible to match the test features of input images with the train features. Neural Network Back Propagation algorithm is used as the recognizer for this OCR system.

First, the training sample is fed to the input layer of the network. For unit j in the input layer, its output is equal to its input, that is, $O_j = I_j$ for input unit j. The net input to each unit in the hidden and output layers is computed as a linear combination of its inputs. A unit j in a hidden or output layer, the net input, I_j , to unit j is

$$I_j = \sum w_{ij} O_i$$

(10)

Where w_{ij} is the weight of the connection from unit i in the previous layer to unit j; O_i is the output of unit i from

the previous layer. The net input I_i to unit j, then O_j , the output of unit j, is computed as

$$O_j = \frac{1}{1+e^{-ij}}$$

(11)

For the output layer, the error value is:

$$\delta_j = O_j (1 - O_j)(T_j - O_j)$$

(12)

And for hidden layer is:

$$\delta_j = O_j (1 - O_j) \sum_k \delta_k w_{jk}$$

(13)

Where w_{jk} is the weight of the connection from unit j to a unit k in the next layer, and δ_k is the error of unit j to a unit k in the next higher layer and δ_k is the error of unit k. Weights are updated by the following equations, where Δw_{ij} is the change in weight w_{jk} :

$$\Delta w_{ij} = \beta \delta_j O_i$$

(14)

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

(15)

The variable β is a constant learning rate. The parameters used in the back-propagation neural network experiments are listed in following Table 1.

Table. 1 Parameters used for back propagation neural network

Parameter	Values
Input layer Neurons	165
Hidden Layer Neurons	100
Myanmar glyphs	50000
Output Layer Neurons	6
Back-propagation learning Rate	0.1
Momentum Team	0.9
Minimum Error Exit in the Network	0.01
Initial weights and biased Term Values	Randomly Generated Values Between 0 and 1

E. Standard Work Error Rate

The word error rate (WER) is based on the Levenshtein distance (Levenshtein, 1966) - the minimum number of substitutions, deletions and insertions that have to be performed to convert the recognition text hyp into the reference text ref. A shortcoming of the W ER is the fact that it does not allow reordering of words, whereas the word order of the hypothesis can be different from word order of the reference even though it is correct translation. In order to overcome this problem, the position independent word error rate (PER) compares the words in the two sentences without taking the word order into account. The PER is always lower than or equal to the WER. On the other hand, shortcoming of the PER is the fact that the word order can be important in some cases. Therefore, the best solution is to calculate both word error rates. The WER of the hypothesis (hyp) with respect to the reference (ref) is calculated as

$$WER = \frac{1}{N^*_{ref}} \sum_r min(d_L(ref_{k,r}, hyp_k))$$

(16)

is the Levenshtein distance between the reference sentence $ref_{k,r}$ and the hypothesis sentencehyp_k. The calculation of WER is performed using a dynamic programming algorithm.

F. Post Processing

This process is to produce the relevant text from the recognition results. This stage is also called the converting process because it converts the recognized character image or classified character image into related ASCII or Unicode text. The final result of this system, the output text can be modified and saved into any format.

IV. EXPERIMENTAL RESULT

The implementation is based on Java Environment using open source tool Eclipse and MySQL Database. Six types of documents which are written in Zawgyi-One font and 12 font sizes is used for comparison of recognition rates in the experimental results. These documents are scanned at 300 dpi for digitization. Table 2 shows the segmentation results of the proposed mechanism. Fig. 9 reveal the recognition error rate of the proposed OCR system on Projection and OCRMPD.

Table. 2 Segmentation Accuracy for Printed Document

Document	Contained Words	Truly Segmented Word		Word Error Rate(WER)	
		Projection only	OCRMPD	Projection on Only	OCRMPD
1	89	87	89	2.25	0
2	95	91	92	4.21	3.16
3	193	184	192	4.66	0.52
4	303	285	301	5.04	0.66
5	364	342	359	6.34	1.37
6	1048	1006	1038	4.01	0.95
	Average			4.41	1.11

The WER of the OCR system is directly proportional with the accuracy of segmentation. The lower the error rate of character segmentation can be obtained, the better the accuracy rate of the OCR system can be got.

The character image is normalized into 30x30 and 25 features are used for zoning method and 60 features are for projection profile method.

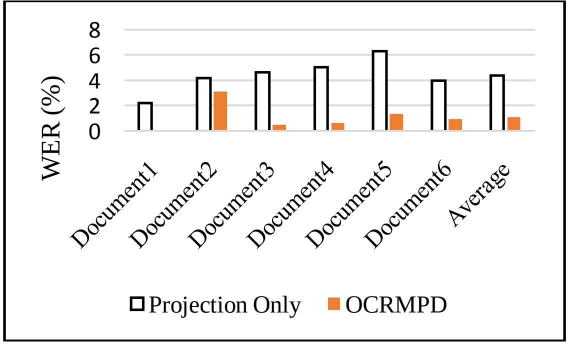


Fig. 9 Compare Recognition WER on Projection and OCRMPD

Fig. 10 shows the average execution time per each document. These times are computed on a PC with 2.4GHz CPU and 2GB of RAM.

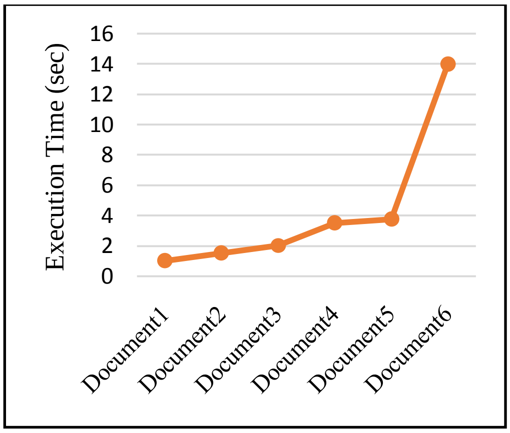


Figure 10. Execution Time for Each Document

V. CONCLUSION

This paper presented neural network classification mechanism for Myanmar Printed document recognition system, OCRMPD, and shows the good result for the system. The experimental results show promising recognition rates. The segmentation scheme can be used for all Myanmar printed documents without user intervention. According the result word error rate, the neural network classification scheme can improve accuracy and save the processing time of classifier. We will show recognition results using different fonts and font sizes in order to prove generality of the proposed OCRMPD (OCR Myanmar Printed Document) in the future. The advancement of the system to recognize bilingual documents and historic documents are future works for the Digital Library Requirement.

ACKNOWLEDGMENT

The authors would like to thank all participants in OCRMPD (Optical Character Recognition (OCR) for Myanmar Printed Document) research field from the University of Computer Studies in Yangon. We would like to thank the reviewers and editors of IJEIT. Finally, we appreciate all the readers who spend their time.

REFERENCES

[1] H. P. P. Win and K. N. N. Tun, “Image Enhancement Processes for Myanmar Printed Documents”, the fifth Conference on Parallel & Soft Computing, University of Computer Studies, Yangon, Myanmar, December 16, 2010.

[2] D. Achaya U, N. V. S. Reddy and Krishnamoorthi, “Hierarchical Recognition System for Machine Printed Kannada Characters”, IJCSNS International Journal of Computer Science and Network Security, Vol. 8 No.11, November 2008.

[3] T.Z.N. Myint "Analyzing Word Error Rate Using Semantic Oriented Approach on Bing Search Engine", IJERT Internal Journal of Engineering Research and Technologies, Vol 2, Issue 11, November 2014, pp. 1094-1102.

[4] R. Singh and M. Kaur, "OCR for Telugu Script Using Back-Propagation Based Classifier", International Journal of Information Technology and Knowledge Management, July-December 2010, Vol. 2, No. 2, pp. 639-643.

[5] 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, 2008.

[6] M. Jaderberg,K. Simonyan, A. Vedaldi, and A. Zisserman," Reading text in the world convolutional neural networks,"Int J.comput. Vis., vol. 116, no. 1,pp. 1-20,2016, http//dx.doi.org/10.107/s11263-015-0823-z

[7] S. loffe and C. Szegedy, "Bath normalization: Accelerating deep network training by reducing internal covariate shift" in Proc. Int. Conf. Mach. Learn, 2015, pp. 448-456.

[8] B. Shi and X. Bai, "An end to end Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition", IEEE Transaction on Pattern Analysis and Machine Intelligence Vol. 39, No. 11, November 2017.

[9] J. Almazan, A Gordo, A. Fornes, and E. Valveny, " Word Spotting and recognition with embedded attributes", IEEE Trans. Pattern Anal. Mach. Intell, vol.36, no.12,pp. 2552-2566, Dec 2014.

[10] Z.Zuo, et al. "Convolution recurrent neural network Learning spatial dependences for image Representation" in Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops,2015,pp. 18-26.

[11] Y. Zhu, C. Yao, and X. Bai, " Scene text detection and recognition: Recent advances and future trends", Frontiers Comput. Sci., vol.10, no1, pp19-36, 2016.

[12] C. Lee, A Bhardwai, W.Di, V.Jagadeesh and R. Piramuthu," Region based discriminative feature pooling for scene text recognition," in Proc. IEEE Conf. Comput. Vision Pattern Recog., 2014, pp.4050-4057.

[13] B. Chaulagain, B. B. Rai and S. K. Raya, "Final Report on Nepali Optical Character Recognition, NepaliOCR", July 29, 2009.

[14] "Myanmar Orthography". Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar, June, 2003.

[15] Y. Thein and M. M. Sein, "Myanmar Intelligent Character Recognition for Handwritten", University of Computer Studies, Yangon, Myanmar, 2006.

[16] H. P. P. Win and K. N. N. Tun, "Image Enhancement Processes for Myanmar Printed Documents", the fifth Conference on Parallel & Soft Computing, University of Computer Studies, Yangon, Myanmar, December 16, 2010.

[17] M. Agrawal and D. Doermann, "Re-targetable OCR with Intelligent Character Segmentation", The Eight IAPR Workshop on Document Analysis Systems, 2008.

[18] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

[19] J. U. Buncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[20] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.