An Analytical Study of Issues in Migration of Print Images to Digital Formats

Htwe Pa Pa Win, Phyo Thu Thu Khine, Khin Nwe Ni Tun
University of Computer Studies
Yangon
Myanmar
{hppwucsy, phyothuthukine, knntun}@gmail.com

ABSTRACT: *Migrating large amount of texts and images from the traditional print to digital format involves many issues and challenges despite the technology we have. One major issue is that many languages have different scripts and unique nature and the international protocols fail to comply with the uniqueness of languages. Myanmar is one such special language which has an exclusive script system. We in this paper, propose a method for converting Myanmar print text and images to machine readable formats. In this multi stage method, first task is the correction of noise variants using a unique segmentation way followed by a few more stages. We have deployed the SVM classifier for understanding and reading the character image. Our experimental results show promising outcome.*

Keywords: Noise variant removal, Myanmar Scripts, Text Segmentation, SVM classifiers, OCR, Natural language classification

Received: 18 January 2011, Revised 20 February 2011, Accepted 28 February 2011

1. Introduction

The World is witnessing a considerable transformation from print based-formats to electronic-based formats thanks to advanced computing technology, which has a profound impact on the dissemination of nearly all previous formats of publications into digital formats on computer networks. Text, still and moving images, sound tracks, music, and almost all known formats can be stored and retrieved on computer magnetic disk. Then, one of the important tasks in machine learning becomes the electronic reading of documents. 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 [1], [2].

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 makes 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.

For an OCR system, segmentation phase is an important phase and accuracy of any OCR heavily depends upon segmentation

phase. Due to the different nature of different scripts, the different segmentation mechanism is required for different languages [22]. And Feature extraction is a key step in the process of OCR, which in fact is a deciding factor of the accuracy of the system [23] and can efficient with the use of suitable approach depending on the nature of characters.

Nowadays, support vector machine is a useful technique for data classification and has been found to be successful when used for pattern classification problems. Fundamentally SVMs are binary classification algorithm with a strong theoretical foundation in statistical learning theory. Their ease of use, theoretical appeal, and remarkable performance has made them the system of choice for many learning problems. There are some limitations in most of the non-SVM learning algorithms proposed in the past 20 years due to the fact that they were based, to a large extent, on heuristics or on loose analogies with natural learning systems. The new pattern-recognition SVM algorithms overcome such limitations with a strong underlying mathematical foundation [3].

In this paper, Optical Character Recognition System for Myanmar Printed Document (OCRMPD) 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 SVM 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.

2. Nature Of Myanmar Script

Myanmar (Burmese) script is recognized as Tibeto/Burman 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 'ဥ' and 'ဿ'), 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., က, ဂ, ဃ, ဃ, ဃ and so on). When writing text, space is used after each phrase instead of each word or syllable. The shapes of Myanmar scripts are circular, consist of straight lines horizontally or vertically or slantways, and dots [11], [20].

3. Related Work

Many researchers have proposed several ways to implement various OCR systems. Nazar Saaid Sarhan and Laheeb Al-Zobaidy [4] proposed an OCR system by using well-known Neocognitron Artificial Neural Network for its fast processing time and its good performance for pattern Recognition problems. And in [5] and [6] also used Back-propagation based Neural Network classifier with general features and profile vectors features; they all get in the range of 90% to 95% accuracy rates for their system. Machine printed Kannada [7] and Chinese minority scripts [8] used combination of Tree based classifier and K-Nearest Neighbor algorithm and modified KNN with structural and topological features, and Wavelet Feature, get accuracy 92% and 96% respectively. In [9] and [10] used Hidden Markov Model for Arabic and Nepali texts and showed that sliding window principle technique is good for segmentation, HMM can increase accuracy rate and can be used with open source Tesseract engine. By combining the entire concept from these papers we know that selection of the segmentation strategy, features, and classifier are very important steps in the construction of an OCR. And they also show that degraded document and including small fonts may have low accuracy rate.

Feature Extraction plays the important role in any OCR systems and many methods are proposed. Zone centroid and Image centroid based Distance metric feature extraction system for handwritten numeral OCR of four South Indian scripts are done by [24] and show zone is the simplest and efficient method. G. Vamvakas et. al., [25] and Ngodrup et. al., [26] indicated that sub division method or grid division method is very good Feature Extraction method for Greek historical font and printed Tibetan characters.

Various contributions that report the use of SVM classifier as a good choice for recognition are, J. Dong et al. reported the handwritten Chinese characters recognition by improving nonlinear normalization, feature extraction and tuning kernel parameters of SVM on a large data set with thousands of classes, and contributed to improvement of the overall system performance in [12]; Sachin Rawat and Jawhar [13] described adaptive OCR for Digital Library with human intervention to get feedback for learning; Shivsubramani et al. [3] showed that multiclass Hierarchical SVM has a better accuracy rate than many other classifiers used like Multilayer perceptron, KNN, Naive Bayes, decision tree and other rule based classifiers in implementing OCR for printed Tamil text; OCR for Amharic document [14] used Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for feature extraction and decision directed acyclic graph (DDAG) for multi class classifier; G.Vamvakas et al. [15] proved that hybrid feature extraction method are good and described the techniques of OCR for Greece historic documents and need interference for character clustering; Malayalam character [16] and Kannada numerals [17] OCRs get good accuracy rate between 97% to 98%.

Some of the existing techniques used in OCR for Myanmar scripts are presented here. A system of recognition for printed text in student application form and translated them to English words by using Hopfield Neural Network is proposed in [18] and has 97.56% accuracy rate. MICR based on statistical and semantic approaches for isolated handwritten character [19] is tested on 33 basic characters and 10 digits and gets 81 to 100% accuracy range. To the best of our knowledge, a comprehensive study on the success rate in terms of recognition accuracy for Myanmar printed text OCR system is yet to be reported.

4. Proposed Method

As other traditional OCR systems, the proposed system also includes five processing steps as shown in Figure 1. 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.

4.1 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 denoise 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 deskew the binarized image with generalized Hough Transformed method. The detailed of the preprocessing steps are described in [21].

4.2 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 Devnagari 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 x^{th} row and the y^{th} column of the character matrix is given by a function f:

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

where, a_{xy} takes binary values (i.e., 0 for background white pixels and 1 for black pixels). The horizontal histogram H_h of the character matrix is calculated by the sum of black pixels in each row:

$$H_h(x) = \sum_{y} f(x, y) \tag{2}$$

And cut the lines depend on the $H_h(x)$ values. as shown in Figure 2.

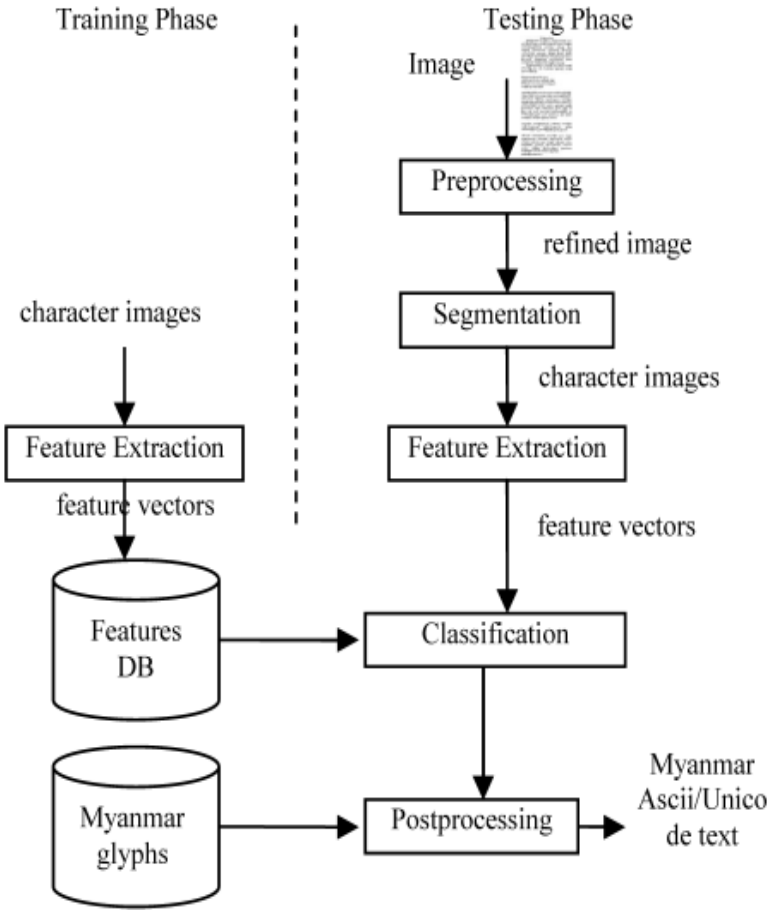

Figure 1. System Design of the Myanmar OCR system

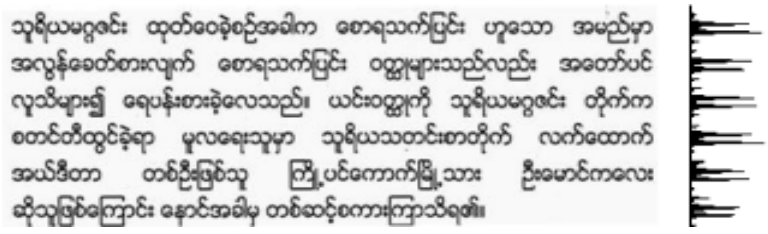

Figure 2. Example of line segmentation using projection

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:

$$H_v(y) = \sum_x f(x,y) \qquad (3)$$

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 Figure 3.

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 (3) as the pre segment points and we need to add the following procedures to check the possible points according to line height:

Figure 3. Example of wrong segmentation error with projection

```
Begin
    CCs ← possible column points of connected components
    mixcharwidth ← the minimum width of the character
    densitythreshold ← the minimun density value for each column
    bottomthreshold ← the threshold distance of the nearest pixel from the bottom
    For each pre segmented point results from (3)
      Begin
       Calculate density of the pixels vertically
       Calculate bottomprojection of each column
       If density < densitythreshold
       Begin
         Store the column point in columnpoints[ ]
         For each column in columnpoints[ ]
         Being
            remaininlength ← width of pre segment point ← column
          If column ∈ CCs
                Begin
                 If (bottomprojection < bottomthreshold &&
                  remaininlength > mixcharwidth)
                  Begin
                       Denote final segment points
                  End
                End
            End
              Else
               Denote pre segment points as the possible points.
        End
      End
  End
```

4.3 Feature Extraction

Before extraction the features we need to normalize the binary character images to have the standard width and height. We normalize all character images height into Nand 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 amount of elements [5]. 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 Figure 4. 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.

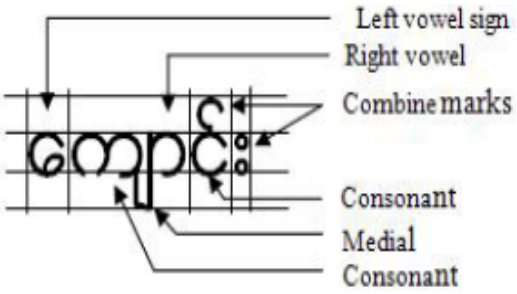

Figure 4. Sample of Myanmar Glyphs

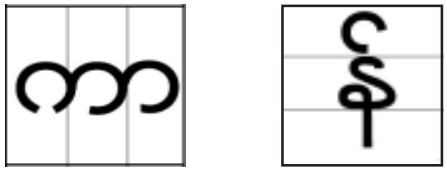

Figure 5. Division of each character depend on writing nature

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:

$$F_z(n) = \sum g(x, y), n=0,...,Z_{max} -1 \quad (4)$$

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

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

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

$$y_i = i(h / S_v) -1, i=1,...,S_v -1 \quad (5)$$

$$\text{follow: } y_s = \begin{cases} y_i - y_p \text{ , for bottom to top} \\ y_p - y_{i-1} \text{, for top to bottom} \end{cases} \quad (6)$$

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:

$$F_v(n) = \sum y_s(x), n = Z_{max},...Z_{max} + F_v - 1 \quad (7)$$

For the horizontal profile projections, the image is split into S_h sections separated by the vertical lines of x and calculated as follow:

$$x_i = i(w/ S_h)-1, i=1,...,S_h -1 \quad (8)$$

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:

$$x_s = \begin{cases} x_i - x_p \text{ , for right to left} \\ x_p - x_{i-1} \text{ , for left to right} \end{cases} \quad (9)$$

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:

$$F_h(n) = \sum x_s(y), n = Z_{max} + F_v, ..., Z_{max} + F_v + F_h - 1 \quad (10)$$

Therefore, the total feature for each character image is:

$$F_{total}(n) = F_z(n) + F_v(n) + F_h(n) \quad (11)$$

4.4 Classification

This process is responsible to match the test features of input images with the train features. SVM [27] is used as the recognizer for this OCR System.

The original form of SVM is the separating of hyperplane between two different classes and is given by the relation

$$f(x) = W^T \Phi(X) + b = 0 \quad (12)$$

where, $\Phi(X) = [\varphi_1(X), \varphi_2(X), ..., \varphi_m(X)]^T$ and $W = [w_1, w_2, ..., w_m]^T$ W and b are weights of the support vector network. The SVM approach consists in finding the optimal hyperplane that maximizes the distance between the closest training sample and the separating hyperplane. Given training vectors $x_i \in R^n, i=1,...,l,$ in two classes and a vector $y_i \in R^t$ such that $y_i \in$ {1, -1}and optimize with the following equation.

$$\min_{w^{ij}, b^{ij}, \xi^{ij}} \frac{1}{2} (w^{ij})^T w^{ij} + C(\sum_t (\xi^{ij})_t) \quad (13)$$

$$\text{subject to } (w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi^{ij}, \text{ if } x_t \text{ in the } i^{th} \text{ class},$$
$$(w^{ij})^T \phi(x_t) + b^{ij} \geq 1 + \xi^{ij}, \text{ if } x_t \text{ in the } j^{th} \text{ class},$$
$$\xi^{ij} \geq 0.$$

Where, C is a parameter that determines the tradeoff between the maximum margin and the minimum classification error. SVM can be used in conjunction with the Kernel and we use the following Gaussian Kernel (RBF).

$$K(X, Y) = \exp\left(\frac{-||X-Y||^2}{\sigma^2}\right) \quad (14)$$

Because of the existence of a number of characters in any script, optical character recognition problem is inherently multi-class in nature. Every character in a language forms a class. The field of binary classification is mature, and provides a variety of approaches to solve the problem of multi-class classification [3], [12], [14].

The Hierarchical mechanism is used for Multi-class SVM classification to reduce search space as there are a large number of characters in Myanmar scripts and there is the similarity between them. Firstly, the similar characters are clustered based on the nature of the writing style of the characters. As a result of this, all characters of 1881 classes can be reduced into 15 classes. And then perform the classification to extract the right class. The example of the hierarchical group of characters is shown in Figure 6. The first level is for the characters for one column, over one column, two columns, over two columns, three columns and over three columns width written in three zones. The second level is for the characters written in two zones with the above column widths and the last level is for the characters for one column, two columns and three columns width that are written in one zone.

4.5 Postprocessing

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.

5. Experimental Results

The implementation is based on Java Environment using open source tool Eclipse and MySql Database. For experiment, 6 Myanmar Printed Documents that are written in Zawgyi-One font with size 12 and scanned on a flatbed scanner at 300 dpi are taken to test the system. These documents have some noise variations. The experiments are carried out for comparing segmentation accuracy, the effects of feature extraction on the accuracy and recognition accuracy. Table 1 shows the segmentation results of the proposed mechanism. Figure 8 compare the effectiveness of hybrid feature extraction method on accuracy rate and

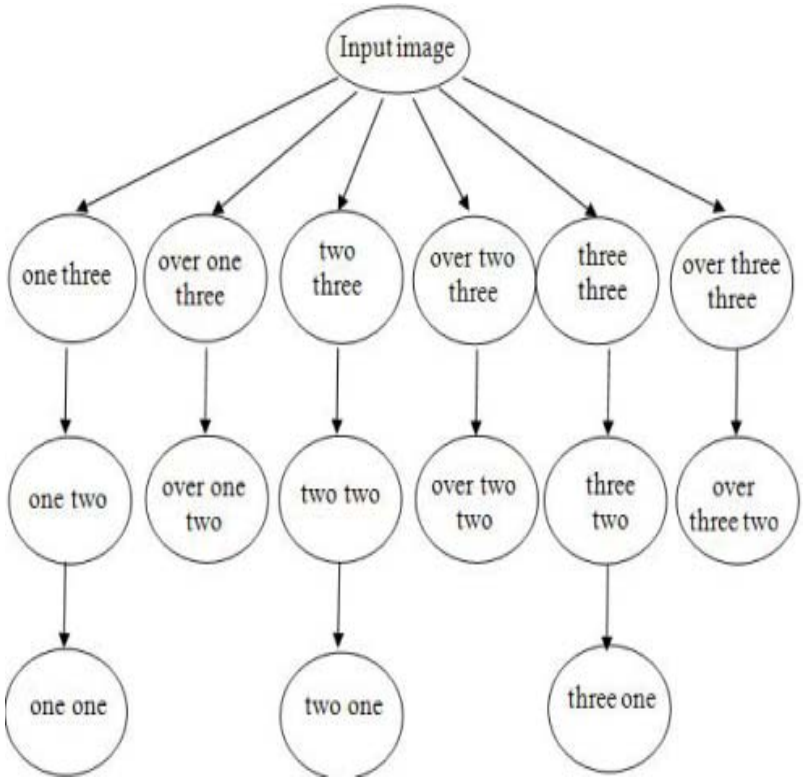

Figure 6. Hierarchical mechanism for Myanmar characters

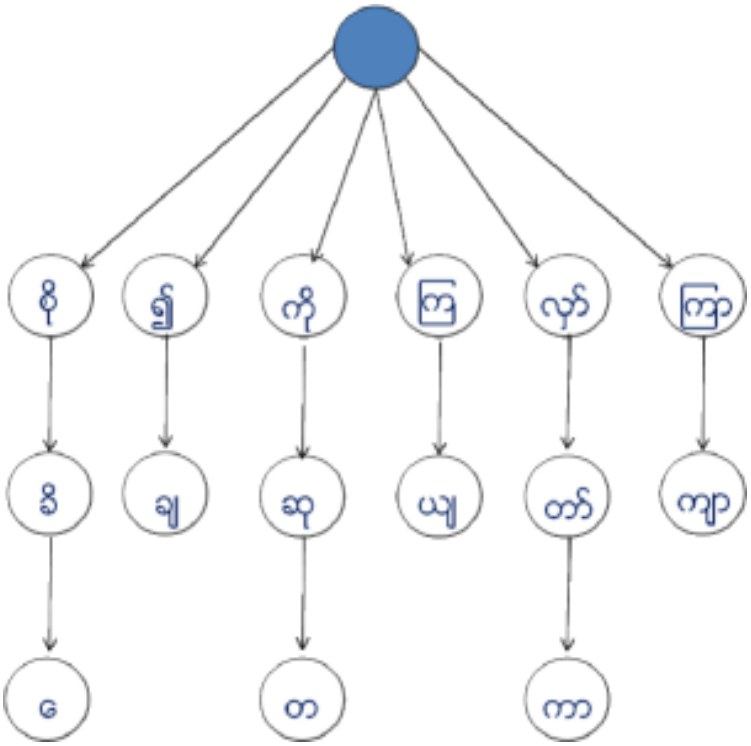

Figure 7. Example of Hierarchical mechanism for Myanmar characters

Document	Contained Characters	Truly Segmented Characters		Accuracy (%)	
		Projection only	OCRMPD	Projection only	OCRMPD
1	89	87	89	97.75	100
2	95	91	92	95.79	96.84
3	193	184	192	95.34	99.48
4	303	285	301	94.06	99.34
5	364	342	359	93.96	98.63
6	1048	1006	1038	95.99	99.05
Average				95.48	98.89

Table 1. Segmentation Accuracy for Printed Document

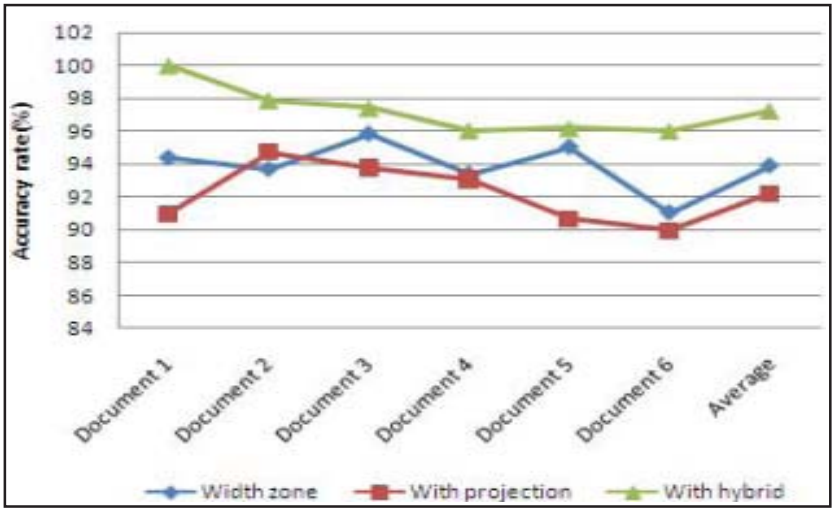

Figure 8. Accuracy Results with various Feature Extraction Methods

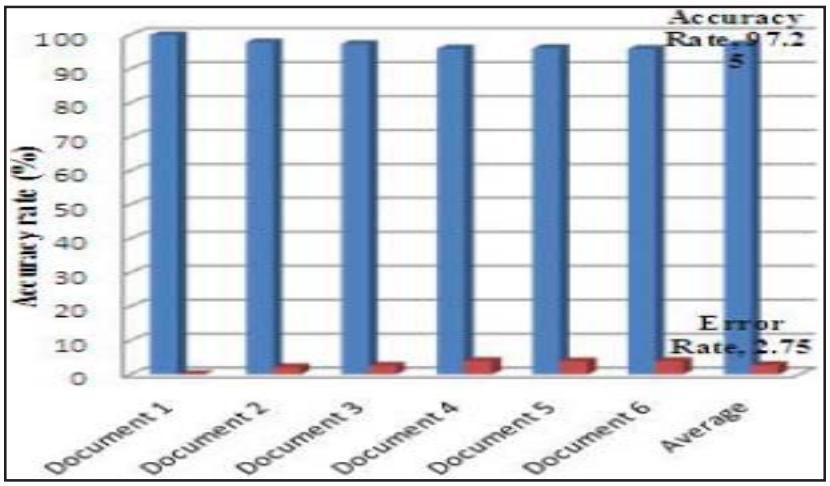

Figure 9. Recognition Accuracy for Myanmar Printed Documents of OCRMPD

Figure 9 reveal the recognition rate of the proposed OCR system.

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

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

6. Conclusion And Future Work

This paper proposes a novel segmentation method to truly separate characters, an efficient feature extraction method and hierarchical classification mechanism for Myanmar Printed document recognition system, OCRMPD, and shows the good result for the system. This result proved the advantages of the innovations. The segmentation scheme can be used for all Myanmar printed documents without user intervention. The combination of feature extraction methods can produce good results but it takes a more time than the normal zoning method. The hierarchical classification scheme can improve accuracy and save the processing time of classifier. The advancement of the system to recognize bilingual documents and historic documents are future works for the Digital Library Requirement.

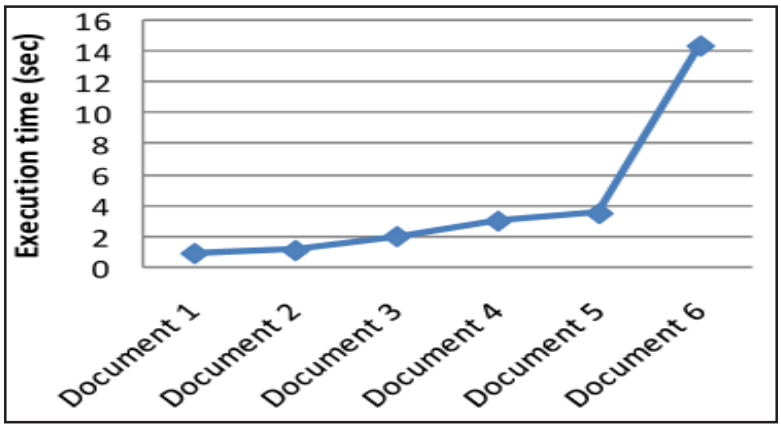

Figure 10. Execution time for each document

References

[1] Govindaraju, V., Setlur, S. (2009). Guide to OCR for Indic Scripts: Document Recognition and Retrieval.

[2] General guidelines for designing bilingual low cost digital library services suitable for special library users in developing countries and the Arabic speaking world, World Library and Information Congress: 75th IFLA General Conference and Counci, 23-27 August (2009), Milan, Italy.

[3] Shivsubramani, K., Loganathan, R., Srinivasan, C. J., Ajay, V. and Soman, K. P. (2007). Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters, Centre for Excellence in Computational Engineering, Amrita Vishwa Vidyapeetham, Tamilnadu, India.

[4] Sarhan, N. S., Al-Zobaidy, L. (2007). Recognition of Printed Assyrian Character Based on Neocognitron Artificial Neural Network, *The International Arab Journal of Information Technology*, 4 (1) January.

[5] Singh, R., Kaur, M. (2010). OCR for Telugu Script Using Back-Propagation Based Classifier, *International Journal of Information Technology and Knowledge Management*, July-December 2010, 2 (2) 639- 643.

[6] Singh, R., Yadav, C. S., Verma, P., Yadav, V. (2010). Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network, *International Journal of Computer Science&Communication*, 1 (1) 91-95.

[7] Achaya, D. Reddy, U. N.V. S., Krishnamoorthi, (2008). Hierarchical Recognition System for Machine Printed Kannada Characters, *IJCSNS International Journal of Computer Science and Network Security*, 8 (11) November.

[8] Guo, H. and Zhao, J. (2010). A Chinese Minority Script Recognition Method Based on Wavelet Feature and Modified KNN, *Journal of Software*, 5 (2) February.

[9] Al-Muhtaseb, H. A., Mahmoud, S. A., Qahwaji, R. S. (2008). Recognition of Off-line Printed Arabic Text Using Hidden Markov Models, Information and Computer Science Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia and Electronic Imaging and media communications department, University of Bradford, Bradford, UK.

[10] Chaulagain, B., Rai, B. B., Raya, S. K. (2009). Final Report on Nepali Optical Character Recognition, NepaliOCR, July 29.

[11] Myanmar Orthography (2003). Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar, June.

[12] Dong, J., Krzyzak, A., Suen, C. Y. (2005). An improved handwritten Chinese character recognition system using support vector machine, *Pattern Recognition Letters*, 26, 1849–1856.

[13] Rawat et al. S. (2006). A Semi-automatic Adaptive OCR for Digital Libraries, Centre for Visual Information Technology, International Institute of Information Technology, Hyderabad, India.

[14] Meshesha, M., Jawahar, C. V. (2007).Optical Character Recognition of Amharic Documents, Center for Visual Information Technology, International Institute of Information Technology, Hyderabad, India.

[15] Vamvakas, G., Gatos, B., Stamatopoulos, N., Perantonis, S. J. (2008). A Complete Optical Character Recognition Methodology for Historical Documents, The Eighth IAPR Workshop on Document Analysis Systems.

[16] Philip, B., Sudhaker Samuel, R. D. (2010). Preferred Computational Approaches for the Recognition of different Classes of Printed Malayalam Characters using Hierarchical SVM Classifiers, *International Journal of Computer Applications* (0975 - 8887) 1 (16).

[17] Rajput, G. G., Horakeri, R., Chandrakant, S. (2010). Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM, (IJCSE) *International Journal on Computer Science and Engineering*, 2 (5) 1622-1626.

[18] Swe, T., Tin, P. (2005). Recognition and Translation of the Myanmar Printed Text Based on Hopfield Neural Network, Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT), p 99-104, Myanmar, November 9-10.

[19] Thein, Y., Sein, M. M. (2006). Myanmar Intelligent Character Recognition for Handwritten, University of Computer Studies, Yangon, Myanmar.

[20] Hussain, S., Durrani, N., Gul, S. (2005). Survey of Language Computing in Asia 2005, Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences.

[21] Win, H. P. P., Tun, K. N. N. (2010). Image Enhancement Processes for Myanmar Printed Documents, The fifth Conerence on Parallel & Soft Computing, University of Computer Studies, Yangon, Myanmar, December 16.

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

[23] Ramanathan et. al., R. (2009). Robust Feature Extraction Technique for Optical Character Recognition, *In*: International Conference on Advances in Computing, Control, and Telecommunication Technologies.

[24] Rajashekararadhya, S. V., Ranjan, P. V. (2008). Efficient Zone Based Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Popular South Indian Scripts, *Journal of Theoretical and Applied Information Technology*.

[25] Vamvakas, G., Gatos, B., Perantonis, S. J. (2009). A Novel Feature Extraction and Classification Methodology for the Recognition of Historical Documents, *In*: 10th International Conference on Document Analysis and Recognition.

[26] Ngodrup et al (2010). Study on Printed Tibetan Character Recognition , International Conference on Artificial Intelligence and Computational Intelligence.

[27] Hsu, C. W., Chang, C. C., Lin, C. J. (2010). A Practical Guide to Support Vector Classification, April 15.