Optical Character Recognition System For Myanmar Printed Documents

Htwe Pa Pa Win, Khin Nwe Ni Tun
University of Computer Studies, Yangon, Myanmar
hppwucsy@gmail.com, knntun@gmail.com

Abstract

Automatic machine-printed Optical Characters or texts Recognizers (OCR) are highly desirable for a multitude of modern IT applications, including Digital Library software. However, the state of the art OCR systems can't do for Myanmar scripts as our language pose many challenges for document understanding. Therefore, we design an Optical Character Recognition System for Myanmar Printed Document (OCRMPD), with several propose techniques that can automatically recognize Myanmar printed text from document image. In order to get more accuracy system, we propose the method for isolation of the character image by using not only the projection methods but also structural analysis for wrongly segmented characters. To reveal the effectiveness of our segmentation technique, we follow a new hybrid feature extraction method and choose the SVM classifier for recognition of the character image. The proposed algorithms have been tested on a variety of Myanmar printed documents and the results of the experiments indicate that the methods can increase the segmentation accuracy as well as recognition rates.

1. Introduction

Optical Character Recognition is one of the most fascinating and challenging areas of pattern recognition with various practical applications. An OCR system converts a document image in optically scanned and digitized pages of text into text format for easy editing, storage, and transmission, searching, indexing and integrating into other applications. Furthermore, the rapid growth of digital library worldwide poses many new challenges for document image understanding [7], [15], [16], [17],

Extensive research have been done on OCR in the last half century and progressed to a level, sufficient to produce technology driven applications [18]. Currently there are many OCR systems that are commercially or freely available for many English and European languages as well as some of the Asian languages such as Japanese, Chinese, etc. As automatic machine-printed Optical Character Recognizers (OCR) are highly desirable for a multitude of modern IT applications, including Digital Library software, efficient OCR systems for Myanmar text are one of the present day requirements. Therefore, we need to concern with printed characters, since handwritten characters become less and less used and only found in signatures because of computerization everywhere.

For an OCR system, segmentation phase is an important phase and accuracy of any OCR heavily depends upon segmentation phase. Character segmentation is an operation that seeks to decompose an image of a sequence of characters into sub images of individual symbols. It is one of the decision processes in a system for optical character recognition (OCR). The

demand for greater than 99% accuracy for printed OCR mandates that the error budget for segmentation be very small, which is indeed a significant challenge for the complex scripts such as those in the Brahmi family. Even in good quality documents, some adjacent characters touch each other due to inappropriate scanning resolution. While there are several scripts for which the process of character segmentation is well researched, and for which very good solutions do exist, there are many more scripts for which the segmentation error rate is high enough to make those OCRs impractical to use. And South-East Asian scripts, syllabic scripts which are in turn a complex combination of one or more characters require different procedures for character alignment and segmentation [1], [2], [4], [9].

Therefore, in this paper we propose the Optical Character Recognition System for Myanmar Printed Document (OCRMPD), for our script and address the segmentation of overlapped characters. The proposed algorithm is based on projection profiles and connected component analysis depending on the nature and structure of our script.

The paper of the later sections is organized as follows. In Section 2, we introduce the nature of Myanmar script. In section 3, we present the previous work as the background theories. In section 4 gives more details on our implementation of recognition system. Results are discussed in Section 5 and are followed by our conclusions.

2. Nature of Myanmar Script

In Myanmar script, there is no distinction between Upper Case and Lower Case characters. The direction of writing is from left to right in horizontally. 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., ဃ, ဓ 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 [19], [20], [21].

From the segmentation point of view, the longest component to form a glyph is 8. But the maximum number of connected components ranging from 1 to 4 and can't be greater than 4. The style of writing is done in upper, middle and lower zones. The sample of Myanmar glyph is as show in figure 1.

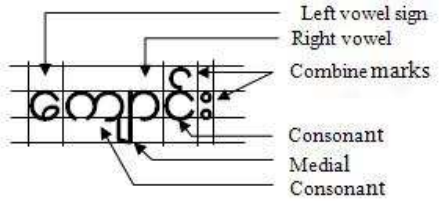

Figure 1. Sample of Myanmar Glyphs

3. Theory and Related Work

The methods for character segmentation can be roughly classified into three categories: straight segmentation method, recognition-based segmentation method, and cut classification method [3]. In the first category, each word is segmented into several characters, and the character recognition techniques are applied to each segment. In spite of the simplicity in implementing this method, its limit comes from the fact that it should depend on high accuracy of the segmentation points found. However, such accurate segmentation technique is not yet

available yet. Consequently, word segmentation and character recognition are needed to be combined. In the second category, a number of potential segmentation points are found in the touched characters. And the candidates are confirmed by using recognition results. This method is more reasonable than the first one, but it depends on the performance of the recognizer. The third category is cut classification method for segmentation. This method is based on a classifier deciding whether it represents a cut hypothesis or not, for each column of the character image. In this method, the decision rules are created automatically rather than being man-made heuristics. But, it is difficult to train for every pair of touching characters when the number of characters to be recognized increases.

In fact, researchers have been aware of the limitations of the classical approach for many years.

Researchers observed that segmentation caused more errors than shape distortions in reading unconstrained characters, whether hand or machine-printed. The problem was often masked in experimental work by the use of databases of well-segmented patterns, or by scanning character strings printed with extra spacing [1]. This phenomenon is the most likely occurrence in Myanmar OCR world.

Most of the previous segmentation works in others' languages OCR system are done with straight segmentation method based on profile as follow.

Authors of [5] propose segmentation of Arabic scripts using high profile vector on binary image of printed documents. They determined the number and locations of possible segmentation points in the sub-word but in some cases, over segmentation and under segmentation occur depend on their nature of scripts. The method for Urdu script in [6] uses structural analysis based segmentation method for their OCR system and overcome different segmentation problems by processing with 4 stages and tune accuracy with horizontal and vertical strokes. They achieve the wonderful segmentation accuracy rate. Some inventions on segmentation of Gurmukhi [8] and Devanagari and Gurmukhi [9] are based on characteristic of the scripts and pre separate upper and lower modifiers and then follow the main strips dissection. This method need to combine the core character and upper/ lowers modifiers and this can create wrong joining problem before classification results. But they were remarked as efficient mechanism for their script nature. Only line and word segmentation technique for Indian printed documents is done by [10].

The works for the group 2 segmentation methods for Khmer can be seen in [4] and it can be used for English language using font model. Urdu segmentation is discussed in [11] by utilizing the knowledge of collocation of ligatures and words in the corpus.

Alternative group for category 3 are for English, Bangla, Farsi Arabic and Telgu of printed documents in [12], [2], [13] and [14] respectively and show the best result of their creativities.

From the evidence of the literature, we have no found any OCR system that can recognize overlapped characters for Myanmar script. And to the knowledge of the authors, we see the simplest and less complexity method and need to modify to find high accuracy of the segmentation points is the dissection method. But we no need to separate the upper or lower modifiers like other our language family because the nature of our script can create wrong glyph for that splitting.

4. Implementation Model

As other traditional systems, our OCR system also includes five processing steps as shown in

Fig. 2. We take 6 different types of documents written in Zawgyi-One font and font size 12 to test our system. These are scanned on a flatbed scanner at 300 dpi for digitization go for the preprocessing steps.

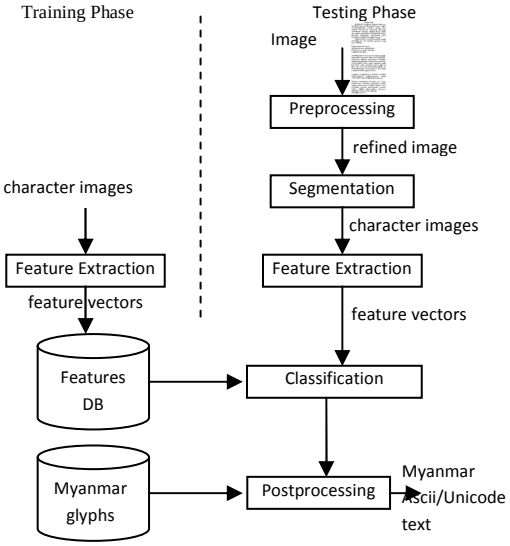

Figure 2. System Design of the Myanmar OCR system

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 [22].

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 [7]. 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. The process of segmentation in our system mainly follows the following pattern:

- Line Detection and slicing
- Word Segmentation
- Character Segmentation

4.2.1. 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 :

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

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:

$$H_h(x) = \sum_y f(x, y) \quad (2)$$

And cut the lines depend on the $H_h(x)$ values as shown in Fig: 3.

4.2.2. Word Segmentation

In Myanmar documents words are generally well spaced. To segment the text line image into words, compute vertical projection profiles as in eq (3). In the profile, the zero valley peaks may represent the character or word space. To differentiate the whether it is character or word spacing, find the maximum character space cluster and use it for separating the words.

4.2.3. Character Segmentation

Similarly, the vertical histogram H_v of the character matrix character matrix is calculated by

the sum of black pixels in each column of the line segment:

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

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 Fig. 5.

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:

```
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
       Remaininglength ←— 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 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 amount of elements [23]. 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. 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. 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.

4.4. Classification

This process responsible to determine each character image into their corresponding text according to features extracted in previous section. We choose SVM as the recognizer for our OCR. The original form of SVM implements two-class classifications. However, because of

the existence of a number of characters in any script [24], [25], [26], optical character recognition problem is inherently multi-class in nature. The field of binary classification is mature, and provides a variety of approaches to solve the problem of multi-class classification [27]. We use the Hierarchical Multi-class SVM for classification to reduce search space as there are a large number of characters in Myanmar scripts and there is the similarity between them. Every character in a language forms a class.

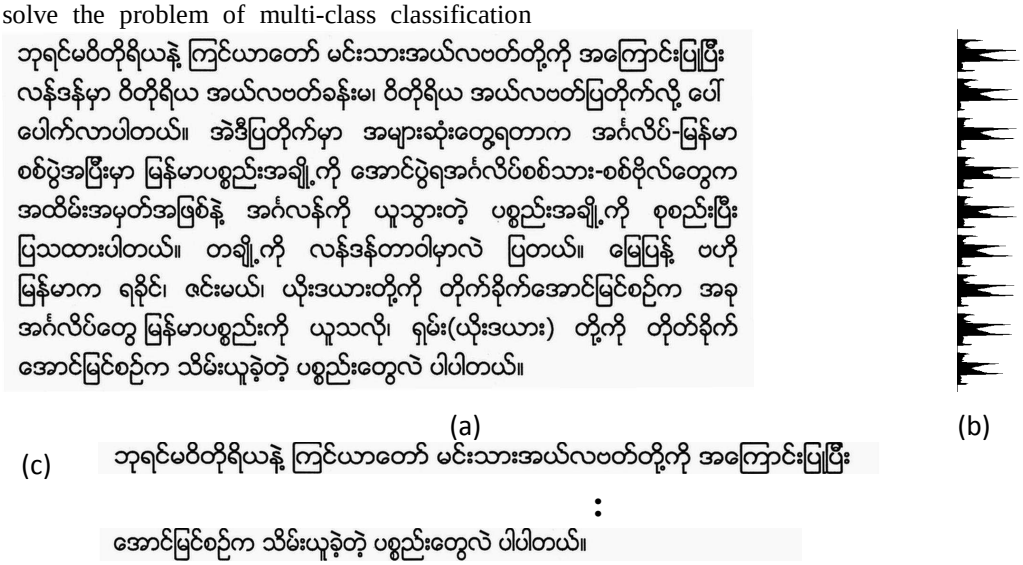

Figure 3. Line Segmentation, (a) Sample input document image, (b) Horizontal Profile of the sample image, (c) Results text line images

Figure 4. Word Segmentation, (a) Extracted Text Line image, (b) Profile of the text line, (c) Results word images extracted from text line image

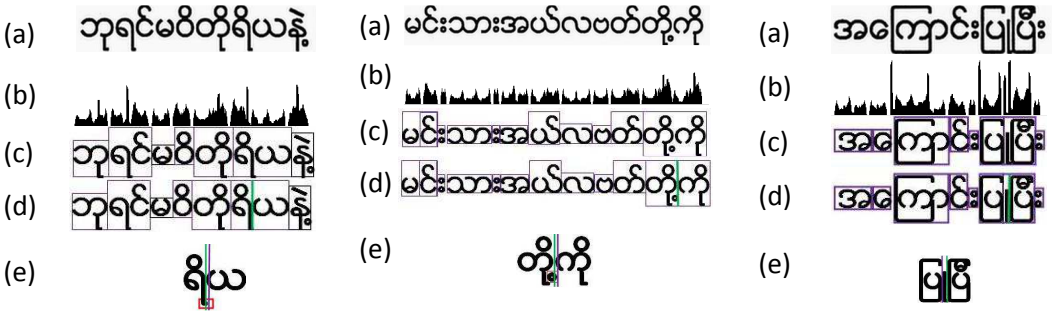

Figure 5. Character Segmentation, (a) Some word images that can create wrong character segmentation, (b) Profile of the word images, (c) Result of the character from vertical profile, (d) Result of the character with our approach, (e) Extracted wrong portion

4.5. Postprocessing

The final process of any OCR system is the outputting of the relevant text from recognition results. Some systems add in this process to increase accuracy by using natural language processing steps, dictionary, lexicons and corpus. Instead of we emphasize for this process, we contribute to others to get the more rate of accuracy.

5. Experimental Result

We implement OCRMPD in Java Environment using open source tool Eclipse and MySql Database. We use 6 Myanmar Printed Documents and tested for our segmentation results and recognition results. Table 1 show the segmentation results of our propose mechanism and Figure 6 show the recognition rate of our proposed OCR system.

Table 1. Segmentation Accuracy for Printed Document

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

The accuracy of the OCR system is in connection with the character segmentation accuracy. The higher the correct segmentation result we can obtain, the more accuracy of the OCR system we get.

6. Conclusion and Future Work

We proposed the mechanism for Myanmar Printed document recognition system, OCRMPD, and experimental results reveal that the proposed methodology is effective for the segmentation and recognition of touched or overlapped characters. We are the first group to propose the SVM recognizer for our scripts. This result proved the advantages of our techniques. The segmentation scheme can be used for all Myanmar printed documents without user intervention. We will try to improve accuracy. And also we need to add bilingual feature and the layout analysis process for historic documents for Digital Library Requirement.

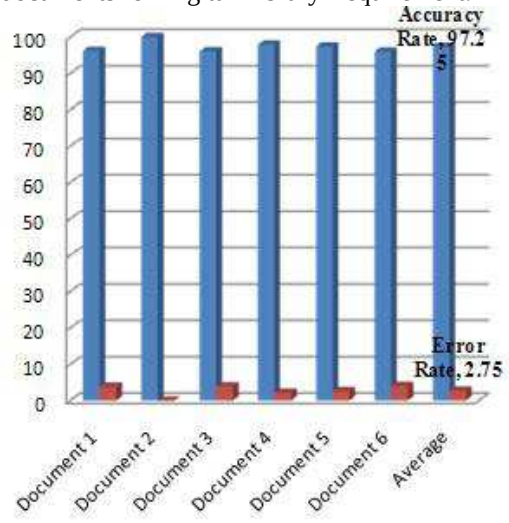

Figure 6. Figure 6: Recognition Accuracy for Myanmar Printed Documents

References

[1] Richard G. Casey and E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 18 No. 7, July, 1996.

[2] Md. A. Hasnat and M. Khan, Rule Based Segmentation of Lower Modifiers in Complex Bangla Scripts, Proceedings of the Conference on Language & Technology 2009.

[3] S. W. Lee, D. J. Lee and H. S. Park, "A New Methodology for Gray-Scale Character Segmentation and Recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.18, No.10, October, 1996.

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

[5] N. A. Shaikh, G. A. Mallah, Z. A. Shaikh, "Character Segmentation of Sindhi, an Arabic Style

Scripting Language, using Height Profile Vector", Australian Journal of Basic and Applied Sciences, 3(4): 4160-4169, ISSN 1991-8178, 2009 .
[6] H. Malik, M. Abuzar Fahiem, "Segmentation of Printed Urdu Scripts Using Structural Features", Second International Conference on Visualization, 2009.
[7] S. Rawat et al., "A Semi-automatic Adaptive OCR for Digital Libraries", Centre for Visual Information Technology, International Institute of Information Technology, Hyderabad - 500032, India, 2006.
[8] M. K. Jindal, R. K. Sharma, G. S. Lehal, "Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script", Compute 2009, Jan 9, 10, Bangalore, Karnataka, India.
[9] V. Kumar and P. K. Sengar, "Segmentation of Printed Text in Devanagari Script and Gurmukhi Script", International Journal of Computer Applications (0975 – 8887), Vol. 3 No.8, June 2010.
[10] N. Priyanka, S. Pal and R. Mandal, "Line and Word Segmentation Approach for Printed Documents", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition, RTIPPR", 2010.
[11] M. Akram and S. Hussain, "Word Segmentation for Urdu OCR System", National University of Computer and Emerging Sciences, B Block, Faisal Town, Lahore, Pakistan, 2008.
[12] P. P. Roy, U. Pal, J. Llados and M. Delalandre, "Multi-Oriented and Multi-Sized Touching Character Segmentation using Dynamic Programming", 10th International Conference on Document Analysis and Recognition, 2009.
[13] A. Broumandnia, J. Shanbehzadeh, M. Nourani, "Segmentation of Printed Farsi/Arabic Words", Islamic Azad University-Tehran South Branch, Tarbiat Moalem University, Tehran University, I. R. Iran, 2007.
[14] M. S. Das et al., "Segmentation Of Overlapping Text Lines, Characters In Printed Telugu Text Document Images", International Journal of Engineering Science and Technology, Vol. 2(11), pg-6606-6610, 2010.
[15] Mantas, J., 1986. An overview of character recognition methodologies, Pattern recognition, 19 (6): 425-430.
[16] Govindan, V.K. and A.P. Shivaprasad, 1990. Character Recognition-A Review, Pattern Recognition, 23 (7): 671-683.
[17] K. Shah and A Sharma, "Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching", Gujarat. IE(I) Journal−ET, 2006.
[18] R. Jagadeesh Kannan and R. Prabhakar, "A Comparative Study of Optical Character Recognition for Tamil Script", European Journal of Scientific Research, ISSN 1450-216X Vol.35 No.4, pp.570-582, 2009.
[19] S. Hussain, N. Durrani and S. Gul, "Survey of Language Computing in Asia", National University of Computer and Emerging Sciences, 2005.
[20] A. Maw, "Encoding of Myanmar Character Set and Implementation", Member, Myanmar IT Standardization Committee, 14 July 2001.
[21] "Myanmar Orthography". Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar, June, 2003.
[22] H. P. P. Win and K. N. N. Tun, "Image Enhancement Processes for Myanmar Printed Documents", PSC, Yangon, Myanmar, December 16, 2010.
[23] R. Singh and M. Kaur, "OCR for Telugu Script Using Back-Propagation Based Classifier", International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 639-643, July-December 2010.
[24] Shivsubramani K, Loganathan R, Srinivasan CJ, Ajay V, Soman KP, Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters", Centre for Excellence in Computational Engineering, Amrita Vishwa Vidyapeetham, Tamilnadu, India, 2007.
[25] J. Dong, A. Krzyzak, C. Y. Suen, "An improved handwritten Chinese character recognition system using support vector machine", Pattern Recognition Letters 26 (2005) 1849–1856.
[26] M. Meshesha and C. V. Jawahar , "Optical Character Recognition of Amharic Documents", International Institute of Information Technology, Hyderabad - 500 032, India, 2007.
[27] Allwein, E. L., Schapire, R. E. and Singer, Y., "Reducing multiclass to binary: A unifying approach for margin classification," In Proceeding of 17th International Conf. on Machine Learning, 2000, pp. 9-16.