ISSN 2229-5518

Bilingual OCR System for Myanmar and English Scripts with Simultaneous Recognition

Htwe Pa Pa Win, Phyo Thu Thu Khine, Khin Nwe Ni Tun

Abstract- The increasing amount of development of the digital libraries worldwide raises many new challenges for document image analysis research and development. Storing wide variety of document images in Digital library, for example, for cultural, technical or historical, that are written in many languages, also create many advancement for present day digital image analysis systems. And when the Digital Library is concerned with Science and Technology documents, it needs to advance the OCR system to bilingual nature as most of them are written in Myanmar in combination with English letters. In this paper a bilingual OCR to simultaneously recognize the printed English and Myanmar texts is proposed including segmentation mechanism for the overlapping nature of Myanmar scripts. The effectiveness of the proposed mechanism is proved with the experimental results of segmentation accuracy rates, comparisons of feature extraction methods and overall accuracy rates.

Index Terms— Bilingual OCR, Machine Printed, Myanmar-English Scripts, SVM;

1 INTRODUCTION

There is 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. Then, one of the important tasks in machine learning is 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], [4].

Furthermore, when the Digital Library is concerned with Science and Technology documents, it needs to advance the OCR system to bilingual nature as most of them are written in Myanmar in combination with English letters. Therefore, in this multilingual and multi-script world, OCR systems need to be capable of recognizing characters irrespective of the script in which they are written. In general, recognition of different script characters in a single OCR module is difficult. This is because features necessary for character recognition depend on the structural properties, style and nature of writing which generally differs from one script to another. For example, features used for recognition of English alphabets are in general not good for recognizing Chinese logograms [3].

Many OCR algorithms for English and other developed countries' languages have been developed over the years for the paperless world and these can be available commercially or freely. But these systems can only recognize for specific single scripts and cannot do for Myanmar scripts. OCR system for Myanmar language is in little effort. In addition, there is no system that can recognize the documents that are written in Myanmar and English text. Therefore, a new system is proposed to recognize these documents simultaneously.

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

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. 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 '□' and '□'), 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. There are total of above 1881glyphs and has many similarity scripts in this language (e.g., □ and □, □ and □, and so on). 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 [4, 5]. The authors of [13-15] are discussed for the feature extraction methods. But in [7-9], they stated that the SVM classifier can be used as the effective recognizer. Some of the existing techniques used in OCR for Myanmar scripts are presented in [10, 11]. 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 Fig. 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].

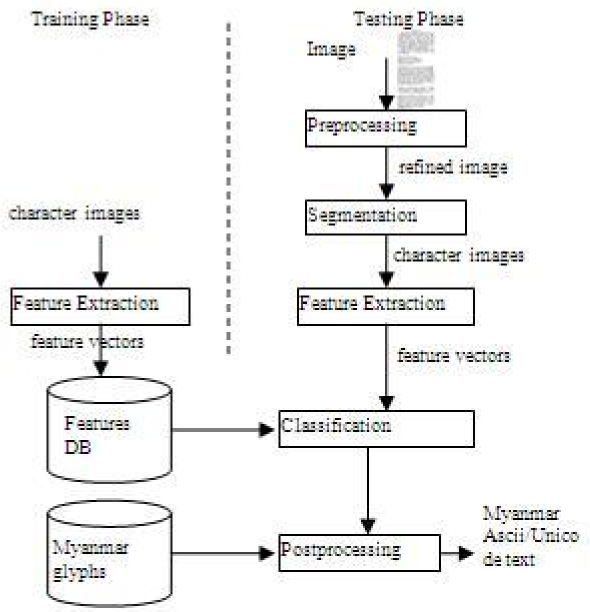

Figure 1. System Design of the Myanmar OCR system

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, for example, 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

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 :

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

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.

4.2.2. 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:

$$H_v(y) = \sum_x f(x,y) \tag{3}$$

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. 2.

Figure 2. Example of wrong segmentation error with projection

And it may also be problem for some connected components. Moreover, the connected components cannot 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

ISSN 2229-5518

```
    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 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 [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 Fig 3. 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.

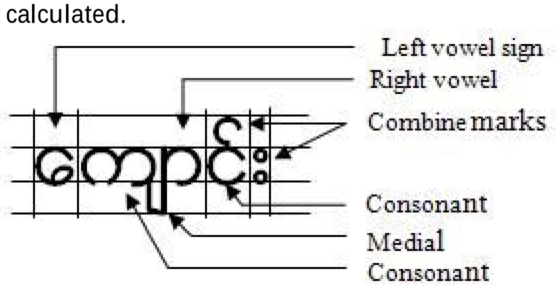

Figure 3. Sample of Myanmar Glyphs

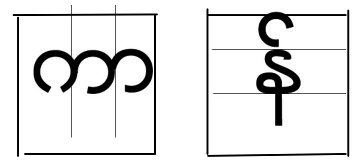

Figure 4. Division of each character depend on writing nature

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:

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

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:

$$y_i = i(h/S_v) - 1, i = 1,\ldots S_v - 1 \qquad (5)$$

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 follow:

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

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:

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

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

$$x_i = i(w/S_h) - 1, i = 1,\ldots S_h - 1 \qquad (8)$$

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:

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

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:

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

Therefore, the total feature for each character image is:

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

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. Because of the existence of a number of characters in any script, 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 [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 and according to width and height ratio. 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 hierarchical group of characters is shown in Figure 5.

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.

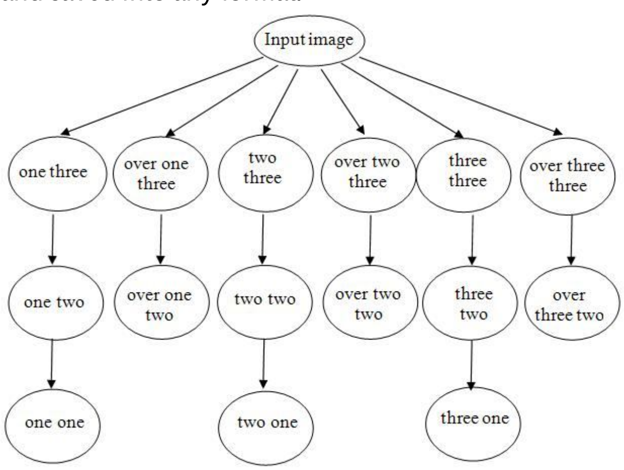

Figure 5: Hierarchical mechanism for Myanmar characters

5 Experimental Results

The implementation is based on Java Environment using open source tool Eclipse and MySql Database. The total of 1881 Myanmar glyphs and 52 of English characters, for small and capital letters are prepared in the training databases. For experiment, 6 Myanmar Printed Documents are used and tested for comparing segmentation accuracy, the effects of feature extraction on the accuracy and recognition accuracy. Table 1 and Table 2 show the segmentation results of the proposed mechanism. Figure 7 compare the effectiveness of hybrid feature extraction method on accuracy rate and Figure 8 reveal the recognition rate of the proposed OCR system. Also 5 different types of bilingual documents are used to test the segmentation accuracy and overall recognition accuracy rates and results are shown in Figure 9.

Table 1. Segmentation Accuracy for Myanmar Printed Document

Document	Contained Characters	Truly Segmented Characters		Accuracy (%)	
		Projection only	Proposed method	Projection only	Proposed method
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

The accuracy of the OCR system is directly proportional with the accuracy of segmentation. The higher the accuracy rate of character segmentation can be obtained, the better the accuracy rate of the OCR system can be getting. The segmentation accuracy rate of bilingual documents is lower than the single language documents because the segmentation scheme is for Myanmar scripts and this cannot be done for English connected component problems.
The character image is normalized into 30x30 and 25 features are used for zoning method and 60 features are for projection profile method.

Table 2. Segmentation Accuracy for Bilingual Documents

Document	Contained Characters	Truly Segmented Characters		Accuracy (%)	
		Projection only	Proposed method	Projection only	Proposed method
1	485	439	477	91.52	98.35
2	443	377	438	85.10	98.87
3	371	338	368	91.11	99.19
4	543	489	533	90.10	98.16
5	495	433	481	87.47	97.17
Average				89.06	98.11

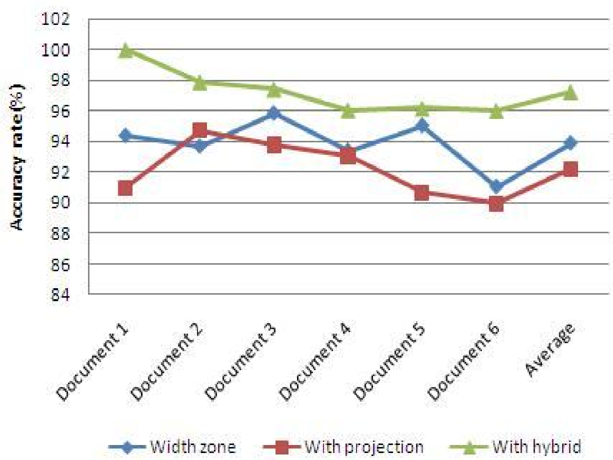

Figure 7: Accuracy Results with various Feature Extraction Methods

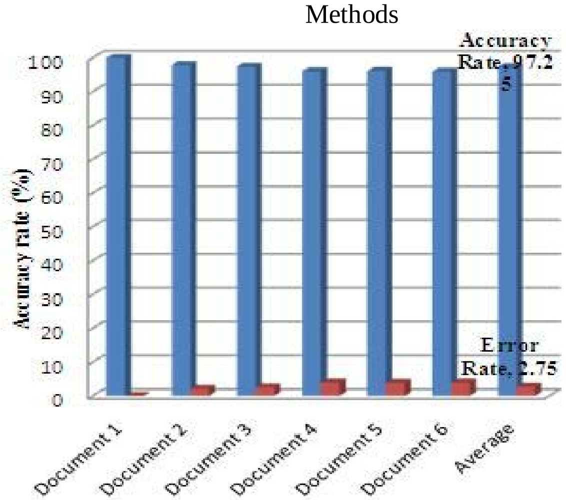

Figure 8. Recognition Accuracy for Myanmar Printed Documents of OCRMPD

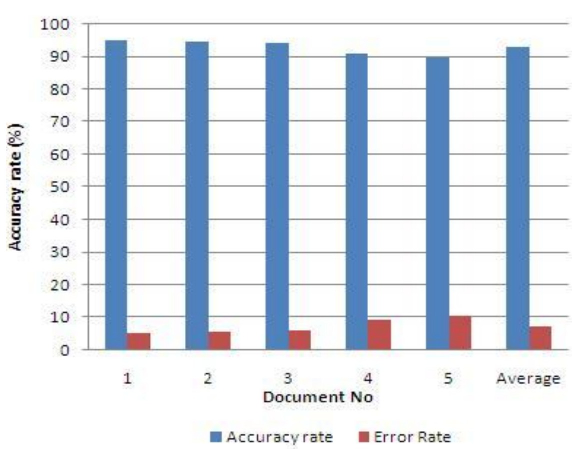

Figure 9. Overall Accuracy rate for bilingual documents

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.

7. References

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

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

[4] N. S. Sarhan and L. Al-Zobaidy, "Recognition of Printed Assyrian Character Based on Neocognitron Artificial Neural Network", The International Arab Journal of Information Technology, Vol 4, No.1, January 2007.

[5] 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.

[6] R. Singh, C. S. Yadav, P. Verma and V. Yadav, "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network", International Journal of Computer Science&CommunicationVol.1, No. 1,January-June 2010, pp. 91-95.

[7] 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.

[8] H. Guo and J. Zhao, "A Chinese Minority Script Recognition Method Based on Wavelet Feature and Modified KNN", Journal of Software, Vol. 5, No. 2, February 2010.

[9] H. A. Al-Muhtaseb, S. A. Mahmoud and R. S. Qahwaji, "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, 2008.

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

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

[12] J. Dong, A. Krzy_ zak and C. Y. Suen, "An improved handwritten Chinese character recognition system using support vector machine", Pattern Recognition Letters, Vol. 26, 2005, pg- 1849–1856.

[13] 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.

[14] M. Meshesha and C. V. Jawahar , "Optical Character Recognition of Amharic Documents", Center for Visual Information Technology, International Institute of Information Technology, Hyderabad - 500 032, India, 2007.

[15] 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.

[16] B. Philip and R. D. Sudhaker Samuel, "Preferred Computational Approaches for the Recognition of different Classes of Printed Malayalam Characters using Hierarchical SVM Classifiers", International Journal of Computer Applications (0975 - 8887) Vol. 1, No. 16, 2010.

[17] G. G. Rajput, R. Horakeri and S. Chandrakant, "Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 05, 2010, pg-1622-1626.

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

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

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

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

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

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

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

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

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

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