Bangla/English Script Identification Based on Analysis of Connected Component Profiles

Lijun Zhou[1], Yue Lu [1,2], Chew Lim Tan[3]

[1] Department of Computer Science and Technology
East China Normal University, Shanghai 200062, China
[2] Shanghai Research Institute of Postal Science
China State Post Bureau, Shanghai 200062, China
[3] School of Computing, National University of Singapore
Kent Ridge, Singapore 117543

Abstract. Script identification is required for a multilingual OCR system. In this paper, we present a novel and efficient technique for Bangla/English script identification with applications to the destination address block of Bangladesh envelope images. The proposed approach is based upon the analysis of connected component profiles extracted from the destination address block images, however, it does not place any emphasis on the information provided by individual characters themselves and does not require any character/line segmentation. Experimental results demonstrate that the proposed technique is capable of identifying Bangla/English scripts on the real Bangladesh postal images.

1 Introduction

Language identification acts as an important role in document image processing, especially for multi-lingual OCR systems. Its goal is to automatically classify textual document images, based on analyzing the stroke structure and connections and the fundamentally different writing styles of the different alphabets or character sets.

In past years, many algorithms for script identification have been proposed. According to entities analyzed in the process of script identification, the algorithms proposed in the literature could be typically classified to four categories: (a) the schemes based on analysis of connected components [1-3]. (b) the schemes based on analysis of characters, words and text lines[4-6]. (c) the schemes based on analysis of text blocks[7-9]. (d) the schemes based on analysis of hybrid information of connected components, text lines etc.[10-15]. We discuss briefly the principles, merits and weakness of each approach.

1.1 Connected component analysis

The approaches based on connected component analysis generally use the intrinsic morphological characteristics of the character sets or strokes of each script.

In [1], Hochberg et al. presented a system that automatically identifies the script form using cluster-based templates. It discovers frequent character or word shapes in each script by means of cluster analysis, then looks for instances of these in new documents. The system develops a set of representative Textual symbols (templates), defined as connected components, for each script by clustering textual symbols from a set of training documents and representing each cluster by its centroid. To identify the script used in a new document, the system compares a subset of symbols from the document to each script's templates, screening out rare or unreliable templates, and the script with the best match chosen as the script of the document. In [2], their work has extended to processing thirteen scripts with minimal preprocessing and high accuracy.

In [3], Spitz presented an approach for automatic determination of the script and language content of document images on the basis of character density or the optical distribution. Based on the spatial relationships of features related to the upward concavities in character structures, the method first classifies the script into two broad classes: Han-based and Latin-based. Language identification within the Han script class (Chinese, Japanese, Korean) is performed by analysis of the distribution of optical density in the text images. They handled 23 Latin-based languages using a technique based on character shape codes.

1.2 Character, word or line analysis

Relatively few methods based on the analysis of character, word or line, have been proposed for language identification. Lee and Kim [4] proposed a scheme for multilingual, multi-font, and multi-size large-set character recognition using self-organizing neural network. They determine not the script of the entire document, but the script of individual characters within the document. In order to improve the performance of the proposed scheme, a nonlinear shape normalization based on dot density and three kinds of hierarchical features are introduced. For coarse classification, two kinds of classifiers are proposed. One is a hierarchical tree classifier, and the other is a SOFM/LVQ based classifier which is composed of an adaptive SOFM coarse classifier and LVQ4 language-classifiers. For fine classification, an LVQ4 classifier has been adopted. Actually, their work focused on English, Korean and Chinese.

In [5], John presented Linguini, a vector-space based categorize used for language identification. Linguini uses dictionaries generated from features extracted from training texts, and compares these against feature vectors generated from test inputs. Features used are N-grams and words, and combinations of both. The accuracy of it depends on the size of the input document, the set of languages under consideration, and the features used. They also presented an algorithm for detecting and determining the nature of bilingual documents.

In [6], three efficient techniques for identifying Arabic script and English script were presented and evaluated. These techniques address the language identification problem on the word level and on textline level. The characteristics of horizontal projection profiles as well as run-length histograms for text written in both languages are the basic features underlying these techniques.

1.3 Text block analysis

Since visual appearances of different scripts are often distinctive from each other, a text block in each script class may be considered as a unique texture pattern. Thus, texture classification algorithms may be employed to perform script identification. Such texture based approach is presented in [7, 8]. In [7], G. S. Peake et al. presented a new scheme based on texture analysis for script identification which did not require character segmentation. Via simple processing, a uniform text block on which texture analysis can be performed is obtained from a document image. Multiple channel (Gabor) filters and grey level co-occurrence matrices are used in order to extract texture features. They used the K-NN classifier to classify the test documents. In [8], V. Singhal et al. proposed an approach on script-based classification of hand-written text documents in a multilingual environment. They apply denoising, thinning, pruning, m-connectivity and text size normalization in sequence to produce a unique text block. They also use Multi-channel Gabor fIltering to extract text features.

Wood et al. [9] proposed a scheme for determining the language classification of printed documents. In that algorithm, the characteristics on the horizontal and vertical projections of the document are used to distinguish European languages, Russian, Arabic, Chinese, and Korean. They argued that interference from other parts of the text which did not have the desired features made the profiles less definitive. In order to enhance the features of the profile and remove the interfering effects, they applied a series of filtering methods to the original image containing medial axis transform and morphological erosion and dilation.

1.4 Hybrid analysis

Most comparatively complex methods are based on hybrid feature analysis. These schemes try to combine the different features extracted from global (text block) and local (text line, word, character and connected components) document entities.

In [11], Pal and Chaudhuri used projection profiles, statistical, topological and stroke based features for identifying English, Urdu, Bangla and Devanagari scripts from a document image. Their work was extended to separation of printed roman, Chinese, Arabic, Devnagari, and Bangla text lines from a single document [12]. Shape based features, statistical features and some features obtained from the concept of water reservoir, have been used in this technique.

In [14], Chaudhury et al. proposed three trainable classification schemes for identification of Indian scripts. The first scheme is based upon a frequency domain representation of the horizontal profile of the textual blocks. The other two schemes use connected components extracted from the textual region. They have proposed a novel Gabor filter-based feature extraction scheme for the connected components. They also use frequency distribution of the width-to-height ratio of the connected components for script recognition. It is claimed that the Gabor filter-based scheme provides the most reliable performance.

The methods discussed above is summarized in table 1, from which we can notice that very few works have been down in identification of hand-written document images compared to machine generated document images. Most of the script identification techniques available in the literature so far consider printed text only. These techniques, especially those schemes that are based on the overall visual appearance of the text block, are generally incapable of tackling the variations in the writing style, character style and size, spacing between lines/words, etc.

Table 1. Summarization of the methods on script identification

Method	Language/Script	Nature
Cluster-Based Templates [1,2]	Arabic, Armenian, Burmese, Chinese. Cyrillic, Devanagari, Ethiopic, Greek, Hebrew, Japanese, Korean, Roman, Thai	Printed
Analysis of character density or the optical distribution [3]	Han script class(Chinese, Japanese, Korean), 23 Latin-based languages	Printed
Self-organizing neural network [4]	English, Korean and Chinese	Printed
Linguini [5]	Catalan, Danish, Dutch, English, Finnish French, German, Icelandic, Italian Norwegian, Portuguese, Spanish Swedish	Printed
Horizontal projection profiles and run-length histograms analysis [6]	Arabic, English	Printed
Texture analysis [7]	Chinese, English, Greek, Korean, Malayalam, Persian and Russian.	Printed
Texture classification algorithm [8]	Roman, Devanagari, Bangla and Telugu	Handwritten
Horizontal and vertical projections analysis[9]	European languages, Russian, Arabic, Chinese, and Korean	Printed
Hybrid analysis [11,12,13,14]	Indian scripts, Roman, Chinese, Arabic,	Printed
Morphological analysis combined with geometrical analysis[15]	Arabic and Latin scripts.	Printed and handwritten

China Post is designing and manufacturing automatic letter sorting machine for Bangladesh Post Office. As a multi-language country, Bangladesh envelopes may be handwritten/printed in Bangla or English. To automatically recognize postcodes or address, Bangla/English script identification in the destination address block (DAB) becomes a crucial step. However, we found that the reported approaches are generally not suitable for our purpose. This is because all these methods apply to a wider range of languages while we specifically would like to maximize the discriminating capability between the Bangla and English scripts for this practical application. In this paper, we propose a novel connected component analysis based approach to identifying Bangla/English scripts on both printed and handwritten envelope images.

2 Proposed Technique for Script Identification

At the first step, the grey scale image is captured from the envelope at the resolution of 200DPI, while a letter is passing by the camera on a letter sorting machine. An adaptive threshold approach is utilized to convert the grey scale image to its binary one, as given in Fig.1(a) and (b). The postal stamp block and other graphic parts are detected and deleted. These processing is a basic stage, but however out of the scope of this paper, and so we will not report the corresponding details here. Based on the positional information of the text block, the destination address block is extracted for subsequent processing as showed in Fig.(c) and (d).

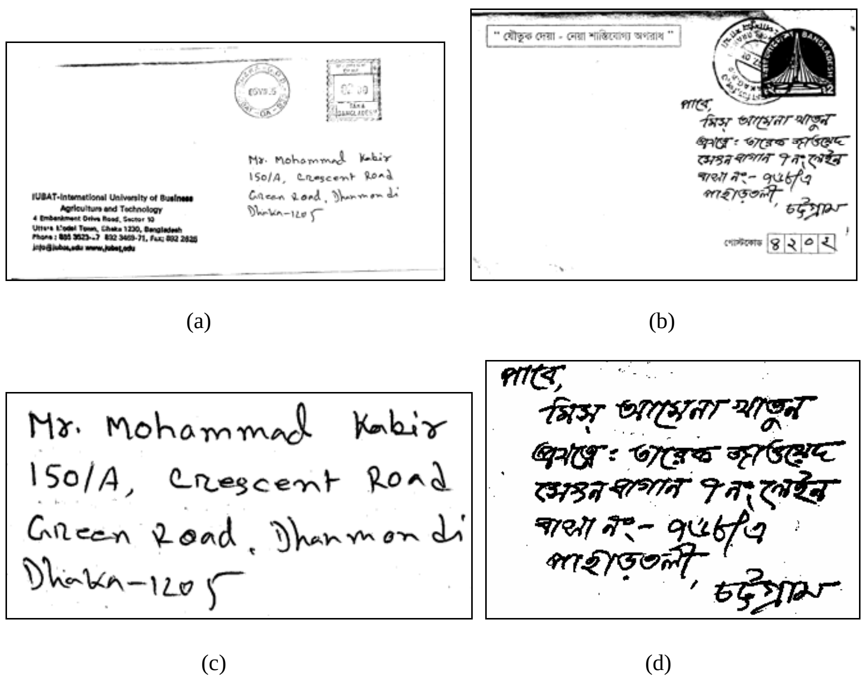
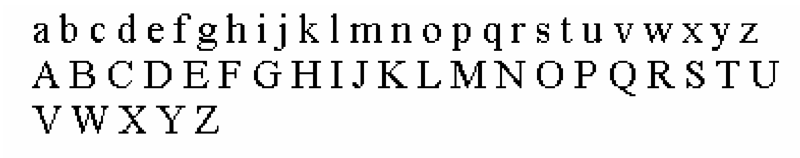
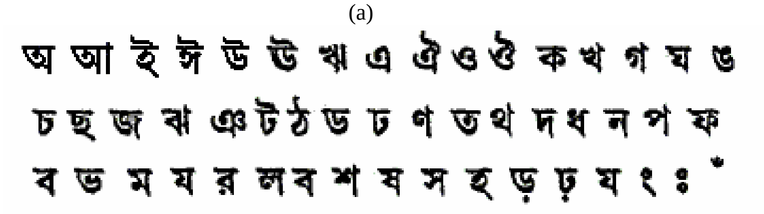


Fig. 1. (a) An example of envelop image written in English. (b) An example of envelop image written in Bangla. (c) Detected destination address block of (a). (d) Detected destination address block of (b).

For language identification, choosing appropriate features is an important or perhaps the most important step. For our purpose, the features used for distinguishing Bangla script and English script are chosen with the following considerations: (a) Easy to detect; (b) Feasible for identification; (c) Independence of fonts, size and style of the text; (b) Robustness.



(a)



(b)

Fig. 2. (a) Basic alphabets of English (b) Basic alphabets of Bangla

There are fifty-two basic alphabets (including 26 upper characters and 26 lower characters) in English script while fifty basic alphabets in Bangla script. And the concept of upper/lower case is absent in Bangla. The basic alphabets are shown in Fig.2, from where we can note that English characters are symmetric and regular in the pixel distribution in the vertical direction whereas the difference in the vertical pixel distribution of Bangla characters is prominent. For the English characters, both the location of the lowest and the topmost pixels of each column of the components vary regularly. In Bangla, it is noted that many characters of these alphabets have a horizontal line at the upper part which is called the head-line. When two or more characters sit side by side to form a word in this language, the head-line portions touch one another and generate a long head-line. Most of the pixels of the head-line are the topmost pixels of vertical columns of the components. This kind of line, however, is absent in the lower part. It results in the distinction between the fluctuations (warps) of the topmost and the lowest pixels in each column (top and bottom profile) of the components. Thus, we can take this characteristic as a feature to distinguish English script and Bangla script. As we observed, such features is weakened on hand-written textual document image, however, it is still sufficient for identification.

2.1 Connected components labelling

To extract features from the text block, the set of connected components in the destination address block image is calculated first. Since Bangla text is cursive i.e. characters are connected within each word, a connected component in Bangla text image may correspond to a word. In contrast, English characters are isolated unless conditions due to low print quality or poor scanning. Thus, a connected component is generally related to a character in printed English text block. In the cases of hand-written text blocks, both English and Bangla, a connected component may correspond to either a character or a word, or several characters within a word owing to different writing styles of different people. However, this doesn't affect the performance of our proposed scheme

2.2 Meaningful connected components selection

In order to minimize the effect of non-script specific markings and reduce the computational time as well, during the analysis of the connected components profiles (both topmost profile and bottommost profile), absolutely very small elements, relatively very small or large elements are eliminated. This ensures that we consider only meaningful connected components and at the same time we can avoid special noise appearing in the text.

Absolutely very small elements deletion: To select meaningful connected components, we firstly deleted those with small area, currently set at less than 9 pixels. This processing can remove noise as well. Then the average area (amount of pixels) of the remaining connected components is computed as

$$avg = \frac{1}{M^{'}} \sum_{i=1}^{M^{'}} pix(i) \tag{1}$$

where $M^{'}$ is the number of the remaining connected components in the destination address block.

Relatively small elements deletion: Based on the considerations that most relatively small connected components correspond to punctuations or broken parts of characters or strokes which will affect the accuracy of the script identification, the components that are smaller than a pre-defined threshold should be excluded from further feature analysis. And the corresponding T_s is defined as

$$T_S = \alpha_1 \times avg \tag{2}$$

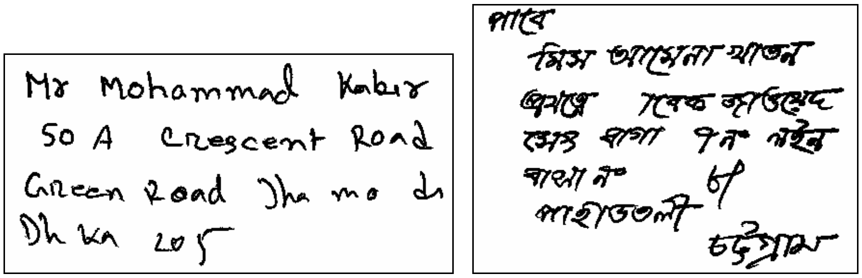
In our experiment, $\alpha_1 = 0.6$ has been proved to be appropriate.

Relatively large elements deletion: As during the step of destination address block extraction, part of the postal stamp may be included because of the overlapping of them. Also, scratched-out words may be sometimes involved in the destination address block. These parts, comparatively large, should be removed too. Here we considered the components which are larger than a threshold T_l as large components. In other words, the component whose area is larger than T_l is eliminated from the profile analysis. T_l is computed as

$$T_l = \alpha_2 \times avg \tag{3}$$

where $\alpha_2 = 5$ has been proved to be a suitable value.

Based on the above processing the results of connected components filtration of Fig.1 (c) and (d) is shown in Fig.3 (a) and (b), respectively.

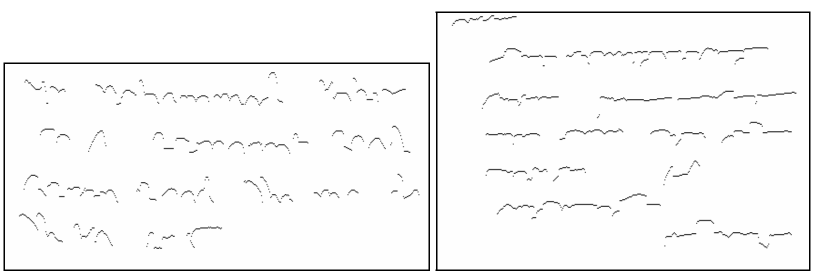


(a) (b)

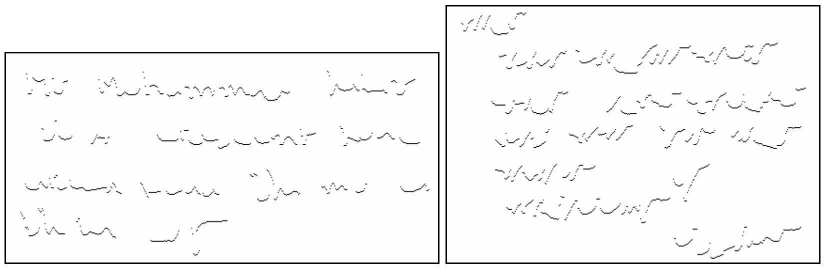
Fig. 3. (a) Connected components filtered of Fig.1(c). (b) Connected components filtered of Fig.1(d).

2.3 Connected Component Profiles analysis

Subsequently, we extract the topmost profile and the bottommost profile of the finally remained connected components respectively, i.e. the topmost pixels and the lowest pixels of vertical columns of the components. To obtain the topmost (bottommost) profile, each vertical column of a particular connected component is scanned from top (bottom) until it reaches a black pixel (p_i). Thus, for a component of width N, we get N such pixels. The topmost profile and bottommost profile of Fig.3.(a) and (b) is showed in Fig.4.



(a) (b)



(c) (d)

Fig. 4. (a) Topmost profile of Fig.3 (a). (b) Topmost profile of Fig.3 (b). (c) Bottommost profile of Fig.3 (a). (d) Bottommost profile of Fig.3 (b).

To measure the discontinuity of topmost (bottommost) contour line of the component, we traverse from p_i to p_{i+1}, and obtain the difference d_i of two adjacent pixels of the components, and is computed as:

$$d_i = \left| y_{p_{i+1}} - y_{p_i} \right|, \quad 0 \le i \le N-1 \tag{4}$$

where y_{p_i} is the Y-coordinate value of the pixel p_i.

And the total distance of the top border of the component is computed as

$$td(j) = \sum_{i=0}^{N-1} d_i \tag{5}$$

On the assumption that the text block has M connected components, its aggregate value of distance of top pixels is produced as

$$ttd = \sum_{j=1}^{M} td(j) \tag{6}$$

The aggregate distance of the bottom pixels, then, is obtained in the similar way and is computed as

$$tbd = \sum_{j=1}^{M} bd(j) \tag{7}$$

where $bd(j)$ is the accumulative difference of bottom pixels of a connected component which is produced like the $td(j)$ term. Text script is inferred from functions ttd and tbd, on the basis that an English text image will have almost equal value of ttd and tbd whereas the difference in ttd and tbd is obviously large in Bangla text image. A normalized measure of this top/bottom difference D_{tb} is defined as

$$D_{tb} = \frac{ttd - tbd}{\min(ttd, tbd)} \tag{8}$$

which is generally indicative of an English text block when positive, and a Bangla text block when appeared to be negative. Exception exists in text blocks when the

D_{tb} term of English text blocks is negative. However, we also find that the absolute value of D_{tb} in both printed and hand- written English text image is generally small. In contrast, the absolute value of D_{tb} in Bangla text image, either printed text or hand-written text, is comparatively large. To investigate we have done an experiment using 100 images (including English text blocks and Bangla text blocks) which are segmented from the real envelope images provided by Bangladesh Post. Through the experiment, two threshold-value are adopted which have been proved to be suitable, the one is thresh1=0.3, another is thresh2=0.1. We identify the script according to the following rules:

Rule 1: if D_{tb} is larger than thresh1, the script of the text block is identified as Bangla.

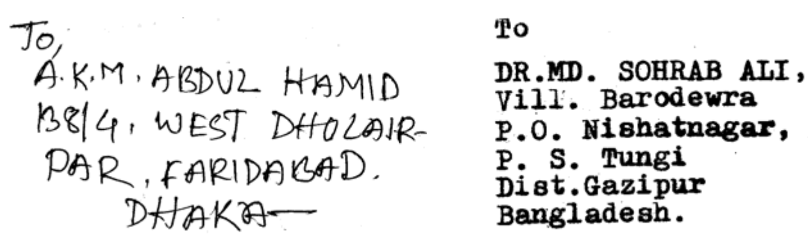
Rule 2: if D_{tb} is smaller than thresh2, the script of the text block is identified as English.

Rule 3: otherwise, the image is rejected from script identification.

The confidence level of the identification increases with increasing difference between D_{tb} and the threshold-value.

3 Experimental Results

1200 images have been used for our experiments. These samples were captured from real Bangladesh postal images. A small amount of page skew was inevitably introduced in practical environment and the character sizes and writing styles were vastly different. Some samples are showed in Fig.5.



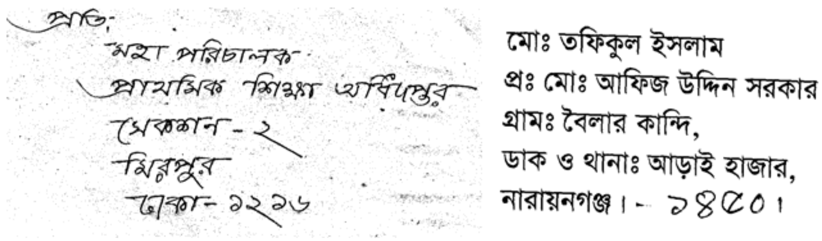


Fig. 5. Examples of destination address block used in the experiment

The experimental results are shown in Table 2 and Table 3. It is observed that the accuracy of script identification is very high for printed text, and for hand-written text, the proposed approach can also achieve a satisfactory accuracy of about 95%.

Table 2. Performance for identifying printed envelop images

Script	Recognized as		Rejected
	English	Bangla	
English	98.00	0.66	1.33
Bangla	0	100.00	0

Table 3. Performance for identifying handwritten envelop images

Script	Recognized as		Rejected
	English	Bangla	
English	94.67	1.33	4.00
Bangla	0	95.33	4.67

From the experiments, we noticed that the main reason of mis-recognition and rejection are poor quality of envelope images. And the erroneous identifications of English script were found to be mostly due to the lower part connection of the characters or words. As the lower part of the components is connected and the upper part of the components is unconnected, the distance of the topmost pixels increases whereas the distance of the bottom pixels decrease. However, this seldom occurs.

4 Conclusions

In this paper, we present a simple but novel technique for script identification with applications to the destination address block of Bangladesh envelope images. The approach is based upon the analysis of connected component profiles, however, it does not place any emphasis on the information provided by individual characters themselves and does not require any character/line segmentation. During the extraction of features characterizing the visual appearance of the destination address block, special connected components that are either too small or too large are deleted prior to feature analysis. Thus, the approach is robust with respect to noise. It is clear that this approach is insensitive to character size, font, writing style and case variation in the destination address block. Also, the approach is immune from text height, inter-line, inter-word spacings and skew. Experimental results have showed that relatively simple technique can reach a high accuracy level for discriminating among English script and Bangla script.

References

1. J. Hochberg, L. Kerns, P. Kelly, T. Thomas, Automatic script identification from images using cluster-based templates, Proceedings of Third International Conference on Document Analysis and Recognition, Vol.1, pp. 378, 1995
2. J. Hochberg, P. Kelly, T. Thomas, L. Kerns, Automatic Script Identification From Document Images Using Cluster-Based Templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 176-181,1997
3. A. L. Spitz, Determination of the Script and Language Content of Document Images, IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 235-245, 1997
4. S. W. Lee, J. S. Kim, Multi-lingual, multi-font and multi-size large-set character recognition using self-organizing neural network, Proceedings of International Conference on Document Analysis and Recognition, Vol.1, pp.28-33, 1995
5. M. P. John, Linguini: Language Identification for Multilingual Documents, Proceedings of 32nd Hawaii International Conference on System Sciences, vol.2, pp. 2035-2045, 1999
6. A. M. Elgammal, M. A. Ismail, Techniques for Language Identification for Hybrid Arabic-English Document Images, IEEE Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp.1100-1104, 2001
7. G. S. Peake, T.N. Tan, Script and Language Identification from Document Images, Proceedings of the Workshop on Document Image Analysis, pp.10-17, 1997
8. V. Singhal, N. Navin, D. Ghosh, Script-based classification of Hand-written Text Document in a Multilingual Environment, Research Issues in Data Engineering, pp.47, 2003
9. S. L. Wood, Xiaozhong Yao, K. Krishnamurthi, L. Dang, Language identification for printed text independent of segmentation, Proceedings of the International Conference on Image Processing, vol.3, pp.3428-3431, 1995
10. J. Ding, L. Lam, Ching Y. Suen, Classification of Oriental and European Scripts by Using Characteristic Features, Proceedings of fourth International Conference Document Analysis and Recognition, pp.1023-1027, 1997
11. U. Pal, B. B. Chaudhuri, Script Line Separation from Indian Multi-Script Documents, Proceedings of fifth Intl. Conf. Document Analysis and Recognition, pp.406-409, 1999
12. U. Pal, B. B. Chaudhuri, Automatic Identification of English, Chinese, Arabic, Devnagari and Bangla Script Line, Intl. Conf. Document Analysis and Recognition, pp.0790-0794, 2001
13. U. Pal, S. Sinha, B. B. Chaudhuri, Multi-Script Line identification from Indian Documents, Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol.2, pp.880-884, 2003
14. S. Chaudhury, R. Sheth, Trainable Script Identification Strategies for Indian Languages, Proceedings of 5th International Conference on Document Analysis and Recognation, pp.657-660, 1999
15. S. Kanoun, A. Ennaji, Y. LeCourtier, A. M. Alimi, Script and Nature Differentiation for Arabic and Latin Text Images, Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition, pp. 309-313, 2002