Handwritten Script Identification from a Bi-Script Document at Line Level using Gabor Filters

G.G. Rajput[1] and Anita H.B.[2]

[1] Department of Computer Science, Gulbarga University,
Gulbarga 585106, Karnataka, India
ggrajput@yahoo.co.in, anitahb@yahoo.com

Abstract. In a country like India where more number of scripts are in use, automatic identification of printed and handwritten script facilitates many important applications including sorting of document images and searching online archives of document images. In this paper, a Gabor feature based approach is presented to identify different Indian scripts from handwritten document images. Eight popular Indian scripts are considered here. Features are extracted from pre-processed images, consisting of portion of a line extracted manually from a handwritten document, using Gabor filters. Script classification performance is analyzed using the k-nearest neighbor classifier (KNN). Experiments are performed using five-fold cross validation method. Excellent recognition rate of 100% is achieved for data set size of 100 images for each script.

Keywords: handwritten script, multilingual documents, Gabor filters, KNN classifier.

1 Introduction

In present information technology era, document processing has become an inherent part of office automation process. Many of the documents in Indian environment are multi-script in nature. A document containing text information in more than one script is called a multi-script document. Many of the Indian documents contain two scripts, namely, the state's official language (local script) and English. An automatic script identification technique is useful to sort document images, select appropriate script-specific OCRs and search online archives of document images for those containing a particular script. Handwritten script identification is a complex task due to following reasons; complexity in pre-processing, complexity in feature extraction and classification, sensitivity of the scheme to the variation in handwritten text in document (font style, font size and document skew) and performance of the scheme. Existing script identification techniques mainly depend on various features extracted from document images at block, line or word level. Block level script identification identifies the script of the given document in a mixture of various script documents. In line based Script identification, a document image can contain more than one script but it requires the same script on a single line. Word level script identification allows the document to contain more than one script and the script of every word is

identified. A brief description of the existing pieces of work at line level is given below.

To discriminate between printed text lines in Arabic and English, three techniques are presented in [1]. Firstly, an approach based on detecting the peaks in the horizontal projection profile is considered. Secondly, another approach based on the moments of the profiles using neural networks for classification is presented. Finally, approach based on classifying run length histogram using neural networks is described. An automatic scheme to identify text lines of different Indian scripts from a printed document is attempted in [2]. Features based on water reservoir principle, contour tracing, profile etc. are employed to identify the scripts. Twelve Indian scripts have been explored to develop an automatic script recognizer at text line level in [3,4]. Script recognizer has been designed to classify using the characteristics and shape based features of the script. Devanagari was discriminated through the headline feature and structural shapes were designed to discriminate English from the other Indian script. Further this has been extended with Water Reservoirs to accommodate more scripts rather than triplets. Using the combination of shape, statistical and Water Reservoirs, an automatic line-wise script identification scheme from printed documents containing five most popular scripts in the world, namely Roman, Chinese, Arabic, Devnagari and Bangla has been introduced [5]. This has been further extended to accommodate 12 different Indian scripts in the same document instead of assuming the document to contain three scripts (triplets). Here various structural features, horizontal projection profiles, Water reservoirs (top, bottom, left and right reservoirs), Contour tracing (left and right profiles) were employed as features with a decision tree classifier for script identification. In [6], a model to identify the script type of a trilingual document printed in Kannada, Hindi and English scripts is proposed. The distinct characteristic features of these scripts are thoroughly studied from the nature of the top and bottom profiles and the model is trained to learn thoroughly the distinct features of each script. Some background information about the past researches on both global based approach as well as local based approach for script identification in document images is reported in [7]. Thus, all the reported studies, accomplishing script recognition at the line level, work for printed documents. Script identification from handwritten documents is a challenging task due to large variation in handwriting as compared to printed documents. Some pieces of work of handwritten script identification of Indian scripts at block and word level can be found in the literature [8-11]. To the best of our knowledge, script identification at line level for Indian handwritten scripts has not been reported in the literature as compared to non Indian scripts [12]. This motivated us to design a robust system for Indian script identification from handwritten documents at line level for bilingual scripts. The method proposed in this paper employs analysis of portion of a line comprising at least two words, for script identification, extracted manually from the scanned document images. Consequently, the script classification task is simplified and performed faster as compared to the analysis of the entire line extracted from the handwritten document. Gabor filter bank is used for feature extraction and classification is done using KNN classifier.

2 Properties of Scripts

A brief description of the properties of the scripts considered in our study is given below. All these scripts are written from left to right.

English Script. The modern English (Roman) alphabet is a Latin-based alphabet consisting of 26 letters each of upper and lower case characters. In addition, there are some special symbols and numerals. The letters A, E, I, O, U are considered vowel letters and the remaining letters are considered consonant letters. The structure of the English alphabet contains more vertical and slant strokes.

Indian Scripts. The scripts considered in this paper are Devanagari, Kannada, Tamil, Bangla, Telugu, Punjabi, and Malayalam. All the Indian languages do not have the unique scripts. Some of them use the same script. Devanagari script is used to write the languages Hindi, Bhojpuri, Marathi, Mundari, Nepali, Pali, Sanskrit, Sindhi and many more. Devanagari is recognizable by a distinctive horizontal line running along the tops of the letters that links them together. Assamese and Bangla languages are written using the Bangla script; Urdu and Kashmiri are written using the same script and Telugu and Kannada use the same script. Like Kannada and Telagu, Tamil and Malayalam belong to southern group of Dravidian languages. The Gujarati script is derived from the Devanagari script. The major difference between Gujarati and Devanagari is the lack of the top horizontal bar in Gujarati. The Gurmukhi (Punjabi) alphabet is modeled on the Landa alphabet. Similar to Devanagari script, in Gurumukhi most of the characters have a horizontal lines at the upper part called headline and primarily the characters of words in these scripts are connected by a these headlines. The image blocks of these scripts are shown in Fig.1. The details about these scripts can be found elsewhere [http://en.wikipedia.org/wiki/Languages_of_India].

3 The Proposed Method

The proposed method is inspired by the observation that in Indian context, handwritten script identification from multilingual/multi-script documents images is very promising and is still in emerging status.

3.1 Data collection and Preprocessing

At present, standard database of handwritten Indian scripts are not available. Hence, we created our own database of handwritten documents. The document pages for the database have been collected by different persons on request under our supervision. The writers were asked to write few text lines inside A-4 size pages. Restrictions were not imposed regarding the content of the text and use of pen. Handwritten documents were written in English, Devanagari, Kannada, Tamil, Bangla, Telugu, Punjabi, and Malayalam scripts by persons belonging to different professions. The document pages were scanned at 300 dpi resolution and stored as gray scale images. The scanned

image is then deskewed using the method defined in [13]. Noise is removed by applying median filter. The portion of lines of width 512 pixels and height equal to that of the height of the largest character appearing in that line were then manually cropped out from different areas of the document image, and stored as data set (Fig. 1). It should be noted that the handwritten text line (actually, portion of the line arbitrarily chosen) may contain two or more words with variable spaces between words and characters. Numerals that may appear in the text were not considered. It is ensured that at least 50% of the cropped text line contains text. A sample of line images representing different scripts is shown in Fig. 1. These lines, representing a small segment of the handwritten document images are then binarized using well known Otsu's global thresholding approach [14] (Fig. 2(b)). The binary images are then inverted so that text pixels represent value 1 and background pixels represents value 0(Fig. 2(c)). The salt and pepper noise around the boundary is removed using morphological opening. This operation also removes discontinuity at pixel level (Fig. 2(d)). However, we do not try to eliminate dots and punctuation marks appearing in the text line, since these contribute to the features of respective scripts. A total of 800 handwritten line images containing text are created, with 100 lines per scripts.

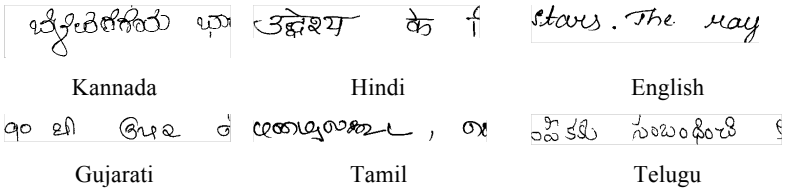

Fig. 1. Sample handwritten line images in different scripts (display in binary form).

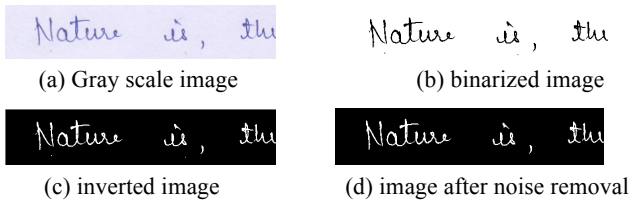

Fig. 2. Pipeline process of pre-processing

3.2 Feature Extraction

Features are the representative measures of a signal which distinguish it from other signals. A bank of Gabor filters are chosen for the task under consideration. The features are extracted by using two dimensional Gabor functions by transforming the image in time domain to the image in frequency domain. The frequency information of image is needed to see information that is not obvious in time-domain. Inherent advantages offered by Gabor function include (i) it is the only function for which the lower bound of space bandwidth product is achieved, (ii) the shapes of Gabor filters resemble the receptive field profiles of the simple cells in the visual pathway, and (iii) they are direction specific band-pass filters. Gabor filters are formed by modulating a

complex sinusoid by a Gaussian function with different frequencies and orientations. A two dimensional Gabor function consists of a sinusoidal plane wave of some frequency

$$g(x,y)=\left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left(-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2}\right)\right)\exp\left(2\pi jWx'\right) \quad \textbf{(1)}$$

$$x' = \quad x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta$$

where σ_x^2 and σ_y^2 control the spatial extent of the filter, θ is the orientation of the filter and w is the frequency of the sinusoid.

We employ two dimensional Gabor filters to extract the features from input text line image to identify the script type from a bi-script document. The preprocessed input binary image is convolved with Gabor filters considering six different orientations (0º, 30º, 60º, 90º, 120º, and 150º) and three different frequencies (a=0.125, b=0.25, c=0.5) with $\sigma_x = 2$ and $\sigma_y = 4$. The values of these parameters are fixed empirically. From the 18 output images we compute the features of dimension 54. These features are then fed to the K-NN classifier to identify the script. The feature extraction algorithm is given below (Algorithm-1). An example of Gabor filtered images for 0^0 and 30^0 degree orientations and frequencies a, b, and c is shown in Fig. 3.

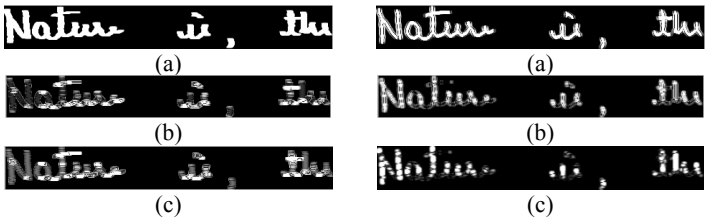

Fig. 3. Gabor filtered images for 0^0 and 30^0 degree orientations and frequencies a, b, and c, respectively

Algorithm-1

Input: Image in gray scale at line level.

Output: Feature vector

Method:

1. Apply median filter to remove noise (Fig. 2(a)).
2. Binarize the image using Otsu's method and invert the image to yield text representing binary 1 and background binary 0 (Fig. 2(b)).
3. Remove small objects around the boundary using morphological opening (Fig. 2(c)).
4. Crop the image by placing bounding box over the portion of line. And apply thinning operation (Fig. 2(d))
5. Create Gabor filter bank by considering six different orientations and three different frequencies to obtain 18 filters.
6. Convolve the input image with the created Gabor filter Bank (Fig. 3).

7. For each output image of step 6 (out of total 18), perform following steps.
   a. Extract cosine part and compute the standard deviation (18 features).
   b Extract sine part and compute the standard deviation(18 features).
   c. Compute the standard deviation of the entire output image (18 features).
   This forms feature vector of length 54.

3.3 Script Recognition

The KNN classifier is adopted for recognition purpose. This method is well-known non-parametric classifier, where posterior probability is estimated from the frequency of nearest neighbors of the unknown pattern. The key idea behind k-nearest neighbor classification is that similar observations belong to similar classes. The test image feature vector is classified to a class, to which its k-nearest neighbor belongs to. Feature vectors stored priori are used to decide the nearest neighbor of the given feature vector. The recognition process is described below.

During the training phase, features are extracted from the training set by performing feature extraction algorithm given in the feature extraction section. These features are input to KNN classifier to form a knowledge base that is subsequently used to classify the test images. During test phase, the test image, which is to be recognized, is processed in a similar way and features are computed as per the algorithm described in feature extraction section. The classifier computes the Euclidean distances between the test feature vector with that of the stored features and identifies the k-nearest neighbor. Finally, the classifier assigns the test image to a class that has the minimum distance with voting majority. The corresponding script is declared as recognized script.

3.4 Experimental Results

We evaluated the performance of the proposed bi-script identification system on a dataset of 800 pre-processed images obtained as described in section 3.1. Each bi-script document contains one Indian script and an English script. Further, we have assumed that the bi-script document contains uniscript text in the portion of line extracted for experimentation. Samples of one script are input to our system and performance is noted in terms of recognition accuracy. For each data set of 100 line images of a particular script, 60 images are used for training and remaining 40 images are used for testing. Identification of the test script is done using KNN classifier. The results were found to be optimal for k=1 as compared to other values of k. To test the robustness of the proposed method k-fold cross validation was carried out with k=5. The proposed method is implemented using Matlab 6.1 software. The recognition results of all the scripts are tabulated in Table 1 and Table 2. The results clearly shows that features extracted by using Gabor function yield very good results. The recognition accuracy of 100% (nearly) is achieved demonstrating the fact that Gabor filters provide good features for the text images at line level as compared to other methods found in the literature. However, the results obtained have certain limitation as explained below. Firstly, the process of extracting the portion of a line, ensuring that it consists of at least two words, is manual. Secondly, we have assumed that the extracted portion of the line is uniscript in text. Thirdly, as with the many other

researchers, we have assumed that the documents are text only. Lastly, we need to validate our proposed system on a larger database. Experimentation is underway to take care of these limitations and propose the system in general to recognize the script at word level.

4 Conclusion

In this paper, a robust algorithm for script identification from multi script handwritten documents is presented. Gabor filters are used for feature extraction. Experiments are performed at line level by considering only a portion of the line. KNN classifier is used in recognition phase. Recognition rate of 100% is achieved. The proposed method is independent of style of hand writing. The novelty of the proposed method lies in the use of Gabor features on a portion of the line for script recognition, instead of entire line. We have assumed that such a portion of line contains uniscript text. Though this assumption is valid for many of the multi-lingual documents, in a general case we need to recognize the script at word level. Hence, our further study involves extending the proposed method for the remaining Indian scripts and also for script type identification at word level.

Table 1. Handwritten script recognition performance of Gabor filter based technique on bi-script documents

Script	% of recognition
Kannada	100%
Malayalam	100%
Punjabi	100%
Tamil	100%
Gujarati	99.92%
Telugu	100%
Hindi	99.98%

Table 2. The average recognition results in the form of confusion matrix with k-fold (k=5) cross validation for Hindi-English and Gujarati-English

Script	Hindi	English
Hindi	99.98%	0.02%
English	0%	100%

Script	Gujarati	English
Gujarati	99.92%	0.08%
English	0%	100%

Acknowledgement. The authors thank the reviewers for their helpful comments. Also, the authors are very grateful to Dr. P. S. Hiremath, Professor, Department of Computer Science, Gulbarga University, Gulbarga and Dr. Peeta Basa Pati, Bangalore, for their valuable suggestions during this work.

5 References

1. Elgammmal. A. M and Ismail. M.A, “Techniques for Language Identification for Hybrid Arabic-English Document Images”, Proc. Sixth Int’l Conf. Document Analysis and Recognition, pp. 1100-1104, (2001).
2. U. Pal, S. Sinha, B. B. Chaudhuri, "Multi-Script Line identification from Indian Documents", ICDAR, Seventh International Conference on Document Analysis and Recognition (ICDAR'03) – vol. 2, pp.880 (2003).
3. Pal. U and Chaudhury.B.B, “Identification of Different Script Lines from Multi-Script Documents”, Image and Vision Computing, vol. 20, no. 13-14, pp. 945-954 (2002).
4. Pal. U and B.B. Chaudhuri, “Script Line Separation From Indian Multi-Script Documents,” 5th ICDAR, pp.406- 409(1999).
5. Pal U. and Chaudhuri. B. B, “Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line”, Proc. 6th Intl. Conf: Document Analysis and Recognition (ICDAR'OI), pp 790-794(2001).
6. M. C. Padma and P. A. Vijaya, Script Identification From Trilingual Documents Using Profile Based Features, International Journal of Computer Science and Applications, Technomathematics Research Foundation, Vol. 7 No. 4, pp. 16 - 33 (2010).
7. S. Abirami, Dr. D. Manjula, “A Survey of Script Identification techniques for Multi-Script Document Images”, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
8. K. Roy, A. Banerjee and U. Pal, “A System for Wordwise Handwritten Script Identification for Indian Postal Automation”, In Proc. IEEE India Annual Conference 2004,(INDICON-04), pp. 266-271 (2004).
9. Ram Sarkar, Nibaran Das, Subhadip Basu, Mahantapas Kundu,Mita Nasipuri and Dipak Kumar Basu, “Word level Script Identification from Bangla and Devanagri Handwritten Texts mixed with Roman Script, Journal of Computing”, volume 2, Issue 2, February 2010, ISSN 2151-9617 (2010).
10. B. V. Dhandra and Mallikarjun Hangarge. Article: “Offline Handwritten Script Identification in Document Images. International Journal of Computer Applications”, 4(5): 1–5, July 2010.
11. G. G. Rajput and Anita H B., “Handwritten Script Recognition using DCT and Wavelet Features at Block Level”, IJCA, Special Issue on RTIPPR (3):158–163 (2010).
12. Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, “Script and language identification for handwritten document images,” *IJDAR,* vol.2, pp. 45-52 (1999).
13. G. G. Rajput, Anita H. B., “A Two Step Approach for Deskewing Handwritten and Machine Printed Document Images using Histograms and Geometric features”, Proc. of Second Intl. Conf. on Signal and Image Processing(ICSIP-2009), Editors: D. S. Guru and T. Vasudev, pp 414-417(2009).
14. N. Otsu, “A Threshold Selection Method from Gray-Level Histogram”, IEEE Transaction Systems, Man and Cybernetics, Vol 9, no.1, pp.62-66 (1979).