International Journal of Electronics Signals and Systems

Volume 1 | Issue 2 Article 3

September 2011

Handwritten Script Recognition using DCT, Gabor Filter and Wavelet Features at Line Level

G. G. Rajput Dr.
Dept. of Computer Science, Gulbarga University, Gulbarga-585106, Karnataka, India, ggrajput@yahoo.co.in

Anita H. B.
Dept. of Computer Science, Gulbarga University, Gulbarga-585106, Karnataka, India., anitahb@yahoo.com

Available online at www.interscience.in

Handwritten Script Recognition using DCT, Gabor Filter and Wavelet Features at Line Level

Dr. G.G. Rajput, Anita H.B.
Dept. of Computer Science, Gulbarga University, Gulbarga-585106, Karnataka, India.
E-mail: 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 multiple feature based approach is presented to identify the script type of the collection of handwritten documents. Eight popular Indian scripts are considered here. Features are extracted using Gabor filters, Discrete Cosine Transform, and Wavelets of Daubechies family. Experiments are performed to test the recognition accuracy of the proposed system at line level for bilingual scripts and later extended to trilingual scripts. We have obtained 100% recognition accuracy for bi-scripts at line level. The classification is done using k-nearest neighbour classifier.

Keywords *- Handwritten script, Gabor Filter, Discrete Cosine Transform Waelets, K-NN classifier.*

I. 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 multiscript in nature. A document containing text information in more than one script is called a multi-script document. Most of the people use more than one script for communication. Many of the Indian documents contain two scripts, namely, the state's official language (local script) and English. Few other documents contain three scripts, namely, the state's official language (local script), Hindi 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 [2]. 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 runlength histogram using neural networks is described. An automatic scheme to identify text lines of different Indian scripts from a printed document is attempted in [16]. 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 [14,15]. 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 [13].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 [9], 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 [12]. 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[1,8,11]. To the best of our knowledge, script identification at line level for Indian scripts has not been reported in the literature as compared to non Indian scripts[7]. This motivated us to design a robust system for Indian script identification from handwritten documents at line level for bilingual scripts. Later we extend the system for trilingual scripts. Further, the present work is extension to our work presented in [4,5] wherein we proposed script identification techniques for handwritten documents at block level. 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.

In many cases, the most distinguished information is hidden in the frequency content of the signal rather than in the time domain. Hence, in this paper features based upon Gabor filter is presented for identification of script type of eight Indian scripts including English for bi-script documents. Later a multiple feature based approach that combines Gabor with DCT/wavelets is proposed for script type identification from documents consisting of three scripts. The classification is done using k-nearest neighbor (K-NN) classifier.

II. METHOD DESCRIPTION

A. Data collection and Preprocessing

Restrictions were not imposed regarding the content of the text and use of pen. Handwritten documents were written in English, Devnagari, 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 [3]. 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. 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. These lines, representing a small segment of the handwritten document images are then binarized using well known Ostu's global thresholding approach [10]. The binary images are then inverted so that text pixels represent value 1 and background pixels represents value 0. The salt and pepper noise around the boundary is removed using morphological opening. This operation also removes discontinuity at pixel level. 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, with100 lines per scripts. A sample of line images representing different scripts is shown in Figure 1.

B. Feature Extraction

Features are the representative measures of a signal which distinguish it from other signals. The selected features should maximize the distinction between bi-scripts and tri-scripts documents. In our method, features are extracted by using two dimensional Gabor functions by transforming the image in time domain to the image in frequency domain. Gabor filters are formed by modulating a complex sinusoid by a Gaussian function with different frequencies and orientations. The term frequency refers to variation in brightness or color across the image, i.e. It is a function of spatial coordinates, rather than time. The frequency information of image is needed to see information that is not obvious in time-domain. A brief description of the features is given below.

1) *Gabor Filter:* A two dimensional Gabor function consists of a sinusoidal plane wave of some frequency, orientation and modulated by a two dimensional Gaussian.

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

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

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

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

2) *Cosine transforms*: The discrete cosine transform (DCT) concentrates energy into lower order coefficients. The DCT is purely real. The DCT expresses a sequence

of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies that are necessary to preserve the most important features [6]. With an input image, Amn, the DCT coefficients for the transformed output image, Bpq, are computed according to equation shown below. In the equation, A, is the input image having M-by-N pixels, Amn is the intensity of the pixel in row m and column n of the image and Bpq is the DCT coefficient in row p and column q of the DCT matrix.

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos\frac{\pi(2m+1)p}{2M} \cos\frac{\pi(2n+1)p}{2N},$$

$$0 \le p \le M-1,\ 0 \le q \le N-1$$

$$\alpha_p = \begin{cases} 1/\sqrt{M}, & p=0 \\ \sqrt{2/M}, & 1 \le p \le M-1 \end{cases} \qquad \alpha_q = \begin{cases} 1/\sqrt{N}, & q=0 \\ \sqrt{2/N}, & 1 \le q \le N-1 \end{cases}$$

3) *Wavelet Transforms:* The discrete wavelet transform (DWT), which is based on sub-band coding is found to yield fast computation of wavelet transform [6]. It is easy to implement and reduces the computation time and resources required. The wavelet transforms are used to analyze the signal (image) at different frequencies with different resolutions. It represents the same signal, but corresponding to different frequency bands. Wavelets are used for multi resolution analysis, to analyze the signal at different frequencies with different resolutions, to split up the signal into a bunch of signals, representing the same signal, but all corresponding to different frequency bands, and provides what frequency bands exist at what time intervals. Many wavelet families have been developed with different properties. For 2-D images, applying DWT corresponds to processing the image by 2-D filters in each dimension.

In this paper, we employ two dimensional Gabor filters to extract the features from input text line image to identify the script type from 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). The values of these parameters are fixed empirically. From the 18 output images we compute the standard deviation to obtain features of dimension 18. These features are then fed to the K-NN classifier to identify the script. The feature extraction algorithm is given below (Algorithm-1).

Next, we present two novel methods for script identification from tri-script documents. In the first method, we combine Gabor filters and DCT to compute the features of the input image. The algorithm for computing the features is given below (Algorithm-2). In the second method, we combine Gabor and wavelets to obtain the features from the input image. Algorithm for computing the features is given below (Algorithm-3).

Algorithm-1

Input: Image in gray scale at line level.

Output: Feature vector

Method:

1. Apply median filter to remove noise (Figure 2(a)).
2. Binarize the image using Otsu's method and invert the image to yield text representing binary 1 and background binary 0 (Figure 2(b)).
3. Remove small objects around the boundary using morphological opening (Figure 2(d)).
4. Apply thinning operation (Figure 2(e)).
5. Crop the image by placing bounding box over the portion of line (Figure 2(e)).
6. Create Gabor filter bank by considering six different orientations and three different frequencies. We obtain 18 filters.
7. Convolve the input image with the created Gabor filter Bank (Figure 3 and 4).
8. For each output image of step 7 (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
9. Compute the Standard Deviation for 54 convolved images. This forms feature vector of length 54.

The feature extraction algorithm for tri-script is described in following steps.

Algorithm-2 (Gabor combined with DCT)

Input: Image in gray scale at line level.

Output: Feature vector

Method:

1. Perform steps 1 through 7 of algorithm-1 to obtain the preprocessed and cropped mage and Gabor filter bank.
2. Perform following steps.

a. Apply DCT to the cropped image and compute the standard deviation of the DCT image to get one feature.

b. Convolve the cropped image with Gabor filters. For each output image, compute the standard deviation. This gives us 18 features.

3. Concatenate features obtained in step2(a) and (b) to get the feature vector of length 19.

Algorithm-3 (Gabor combined with wavelets)

Input: Image in gray scale at line level.

Output: Feature vector

Method:

1. Perform steps 1 through 7 of algorithm-1 to obtain the convolved images(total 18).
2. Perform Wavelet (Daubechies 9) decomposition for the convolved input images (total 18) to obtain approximation coefficients (cA), vertical coefficients (cV), horizontal coefficients (cH), and diagonal coefficients (cD). Compute the Standard Deviation for each frequency band separately for all images. This forms 4 x 18=72 features.

III. SCRIPT RECOGNITION

K-NN 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 algorithms given in the Feature Extraction section. These features are input to K-NN 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 algorithms 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.

IV. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed multi-script identification system on a dataset of 800 pre-processed images obtained as described in data collection section. The complete dataset is manually processed to generate the ground truth for testing and evaluation of the algorithm. For bi-script documents, we have considered one Indian script and English script. For tri-script evaluation, we consider a local language script, Hindi script and English script, respectively. 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. The proposed method is implemented using Matlab 6.1 software. The recognition results of all the bi-scripts and tri-scripts are tabulated in Table 1,2. The results clearly shows that features extracted by using Gabor function yield good results. The recognition accuracy of 100% is achieved for bi-scripts 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. The results are promising when we applied DCT to the Gabor convolved images as compared to the application of wavelets to Gabor convolved images for tri-scripts.

V. CONCLUSION

In this paper, feature extraction algorithms for script identification from multi script handwritten documents are presented. Gabor filters are used for feature extraction in bi-script identification scheme whereas for tri-script case we have combined Gabor with DCT/wavelets. Experiments are performed at line level for bi-script and tri-scripts. KNN classifier is used in recognition phase that yielded better results for k=1. Recognition rate of 100% is achieved for bi-script as compared to recognition results for tri-scripts. The proposed method is robust and independent of style of hand writing. In future, we extend the proposed method for the remaining Indian scripts and also for script type identification at word level. Furthermore, the methods proposed for script type identification from tri-script documents can be improved to increase the recognition accuracy.

REFERENCES

1. B. V. Dhandra and Mallikarjun Hangarge, “Offline Handwritten Script Identification in Document Images”. International Journal of Computer Applications 4(5):1–5, July 2010.

2. 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.
3. 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, pp 414-417, 2009.
4. 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.
5. G. G. Rajput, Anita H. B., "Kannada, English, and Hindi Handwritten Script Recognition using multiple features", Proc. of National Seminar on Recent Trends in Image Processing and Pattern Recognition, ISBN: 93-80043-74-0, pp 149-152, 2010.
6. Gonzalez and Woods, Digital Image processing, 3/e, Pearson Education, 2008.
7. Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, "Script and language identification for handwritten document images", IJDAR, vol.2, pp. 45-52, 1999.
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. 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.
10. N. Otsu , A Threshold Selection Method from Gray-Level Histogram, IEEE Transaction Systems, Man and Cybernetics, vol 9, no.1, pp.62-66, 1979.
11. 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, ISSN 2151-9617, 2010.
12. 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, 2009.
13. U. Pal and Chaudhuri.B.B, "Automatic identification of English, Chinese, Arabic, Devanagari and Bangla script line", Proc. 6th Intl. Conf: Document Analysis and Recognition (ICDAR'OI), pages 790-794, 2001.
14. U. Pal 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.
15. U. Pal and Chaudhuri B.B., "Script Line Separation from Indian Multi-Script Documents", 5th ICDAR, pp.406- 409,1999.
16. U. Pal, S. Sinha, Chaudhuri B.B., "Multi-Script Line identification from Indian Documents", ICDAR, vol. 2, pp.880, Seventh International Conference on Document Analysis and Recognition, vol 2, 2003.

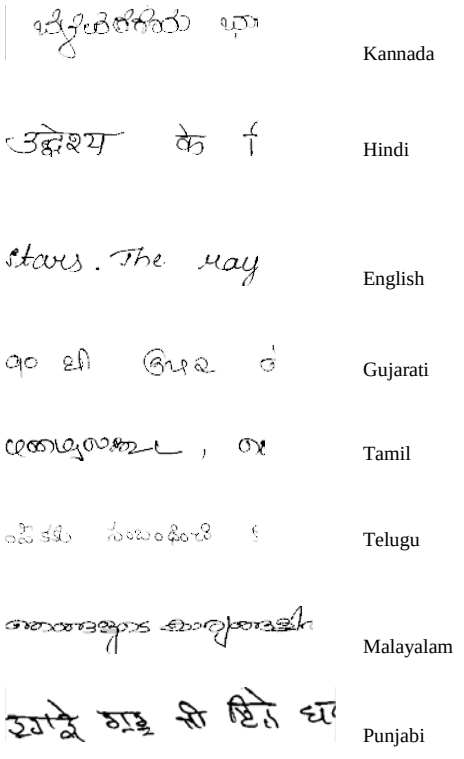
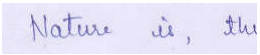
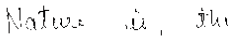


Figure 1: Sample handwritten line images in different scripts.



(a) Gray scale image



(b) binarized image

(c)inverted image

(d) image after noise removal

(e) cropped and thinned image

Figure 2: Pipeline process for feature extraction**.**

(a)

(b)

(c)

Figure 3: Gabor filtered images for zero degree orientation and frequencies a, b, and c

(a)

(b)

(c)

Figure 4: Gabor filtered images for 30 degree orientation and frequencies a, b, and c

Table – I : Recognition results for script type identification for bi-script documents

Bi-scripts	**Recognition %**
Kannada, English	100%
Malayalam, English	100%
Punjabi, English	100%
Tamil, English	100%
Gujarati, English	100%
Telugu, English	100%
Hindi, English	100%

Table-II : Recognition results for script type identification for tri-script documents

Tri-scripts	**DCT applied to Gabor convolved images (19 Features)**	**Wavelets applied to Gabor convolved images (72 Features)**
Kannada, English and Hindi	93%	94%
Malayalam, English and Hindi	95%	90%
Punjabi, English and Hindi	95%	90%
Tamil, English and Hindi	95%	90%
Gujarati, English and Hindi	95%	90%
Telugu, English and Hindi	95%	90%