ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 4, October 2013

Handwritten Script Recognition at Line Level – A Multiple Feature Based Approach

Dr. G.G. Rajput, Anita H.B.
Department of Computer Science, Gulbarga University, Gulbarga-585106, Karnataka, India

Abstract*— Automatic recognition of scripts present in a multiscript document has a variety of practical and commercial applications in banks, post offices, reservation counters, libraries, etc. In a country like India, a printed/handwritten document consisting of English script and a regional script is quite common. Such a document is termed as bi-script document. In this paper, a multiple feature based approach is proposed to identify the script type from a bi-script document. Features are extracted using Gabor filters, Discrete Cosine Transform, and Wavelets of Daubechies family. The classification is done using k-nearest neighbor classifier and SVM classifier. Experiments are performed on nine popular Indian scripts along with English script. An average recognition accuracy of 96.7% is obtained using SVM classifier.

Index Terms—**Discrete Cosine Transform, Handwritten script, Gabor Filter, Wavelets.**

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. An automatic script identification technique is 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 overview of the state of the art in handwritten script recognition is given below. 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 [2]. This motivated us to design a robust system for script identification from handwritten bi-script documents at line level. To discriminate between printed text lines in Arabic and English, three techniques are presented in [4]. 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. 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 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. Twelve Indian scripts have been explored to develop an automatic script recognizer at text line level in [3, 6]. 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. An automatic scheme to identify text lines of different Indian scripts from a printed document is attempted in [7]. Features based on water reservoir principle, contour tracing, profile etc. are employed to identify the scripts. Most of the methods proposed in the literature accomplishing script recognition work for printed documents. Script identification from handwritten documents is a challenging task due to large variation in handwriting as compared to printed documents. From literature survey it is observed that relatively less amount of work is done in the area of handwritten script recognition [8, 14, 16]. 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 [11]. Later, we extend the proposed for tri-script documents. The present work is extension to our work presented in [12, 13], where 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

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 4, October 2013

at least two words, for script identification, extracted manually from the scanned document images. In [15], 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. 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 filters combined with DCT/wavelets are extracted for identification of script type from a bi-script document. The classification is done using k-nearest neighbor (K-NN) and SVM classifiers. The details of the proposed methods are described in the following sections.

II. METHOD DESCRIPTION

A. Data Collection and Preprocessing

Persons belonging to different professions were identified and were asked to write few lines of text in their regional language along with English. Restrictions were not imposed on the writers regarding the content of the text and use of pen. Writers included both native-writers and non native writers. The document pages collected from writers written in English, Devnagari, Kannada, Tamil, Bangla, Telugu, Punjabi, Odiya or Malayalam scripts were scanned at 300 dpi resolutions and stored as gray scale images. Noise present in the image is removed by applying median filter. From the preprocessed documents, individual lines are extracted manually and stored as data set. It was observed that the line consisted of ten to twelve on an average. These lines are then binarized using well known Ostu's global thresholding approach [1]. 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, dots and punctuation marks appearing in the text line were retained since these contribute to the features of respective scripts. A total of 900 handwritten line images containing text are created, with 100 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. In this paper, for script identification features are extracted by using two dimensional Gabor functions by transforming the image into 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.

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

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

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

where σx2 and σy2 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 [9]. 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.

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

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

3) Wavelet Transforms: The discrete wavelet transform (DWT), which is based on sub-band coding is found to yield fast computation of wavelet transform [9]. 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 with DCT/Wavelet filters to extract the features from input text word image for identification for script type. 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 σx

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 4, October 2013

= 2 and σy = 4. The values of these parameters are fixed empirically. From the 18 output images we compute the standard deviation to obtain features of dimension 18. It was observed that, these features extracted by using Gabor filters (Algorithm-1) are not sufficient in terms of recognition accuracy. Hence, to improve the accuracy two separate algorithms were proposed using DCT (Algorithm-2) and Wavelets (Algorithm-3), respectively. The first set of features is obtained by performing algorithm-1 and algoritm-2 on the input word image. The second set of features is obtained by performing algorithm-1 and algoritm-3 on the input word image. These features are then fed to the K-NN and SVM classifier to identify the script. The feature extraction algorithm s is given below.

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(c)).
4) Apply thinning operation (Figure 2(d)).
5) Crop the image by placing bounding box over the portion of line.
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.

Algorithm-2 (Gabor combined with DCT)

Input: Image in gray scale at word level.
Output: Feature vector
Method:

1) Perform steps 1 through 7 of algorithm-1 to obtain the preprocessed and convolved images (total 18).
2) Perform following steps.
   a) Apply DCT to the preprocessed image and compute the standard deviation of the DCT.
   b) Apply DCT for each convolved input images and 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 word level.
Output: Feature vector
Method:

1) Perform steps 1 through 7 of algorithm-1 to obtain the preprocessed convolved images (total 18).
2) Perform following steps.
   a) Apply wavelet to the preprocessed image and compute the standard deviation of the for each frequency bands. This forms 4 features.
   b) Apply DCT to the preprocessed image. Then apply wavelet to the DCT image and compute the standard deviation of each frequency bands, namely, approximation coefficients (cA), vertical coefficients (cV), horizontal coefficients (cH), and diagonal coefficients (cD). This forms 4 features.
   c) Perform Wavelet (Daubechies 9) decomposition for each convolved input images 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 18 images. This forms 4 x 18=72 features.
3) Concatenate features obtained in step2 (a), (b) and (c) to get the feature vector of length 80.

III. SCRIPT RECOGNITION

K-NN and SVM classifiers are adopted for recognition purpose. K-NN classifier 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. Support Vector Machine (SVM) is a universal constructive learning procedure based on the statistical learning theory. SVMs are currently among the best performers for a number of classification tasks ranging from text to genomic data. We have used SVM classification system devised to assess the

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 4, October 2013

potential of SVMs in scripts classification. SVMs can be applied to complex data types beyond feature vectors by designing kernel functions for such data. Tuning SVMs remains a black art: selecting a specific kernel and parameters is usually done in a try-and-see manner. SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. The SVM is a machine learning method basically used for two-class recognition problems. SVM is used in conjunction with the Radial Basis Function (RBF) kernel, a popular, powerful kernel. RBF kernel nonlinearly maps each sample into a higher dimensional space, so it can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, a grid search was performed in order to find the optimal values for both the variance parameter γ of the RBF kernel and the cost parameter C of SVM using cross-validation. Basically pairs of (C, γ) are tried and the one with the best cross – validation accuracy is picked.

IV. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed multi-script identification system on a dataset of 900 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. 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 and SVM 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 with both classifiers are tabulated in Table 1 and 2. The results clearly shows that features extracted by using Gabor function yield good results. Gabor filters provide good features for the text images at line level as compared to other methods found in the literature. Two sets of features are extracted from the proposed algorithms. Features obtained by performing algorithm 1 and algorithm 2 are combined in First feature set. Features obtained by performing algorithm 1 and algorithm 3 are combined in Second feature set. We got good result with SVM classifier when we used Gabor feature. The results are promising when we applied DCT/Wavelet to the Gabor convolved images as compared to the Gabor convolved images.

V. CONCLUSION

In this paper, feature extraction algorithms for script identification from multi script handwritten documents are presented. Here combined Gabor with DCT/wavelets is used for feature extraction in bi-script identification scheme. Experiments are performed at line level for bi-script. KNN and SVM classifiers are used in recognition phase that yielded better results. Remarkable recognition rate of is achieved for bi-script. 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, we can increase the recognition accuracy by removing skew using the method defined in [10] and by combining the multiple classifiers.

VI. ACKNOWLEDGMENT

We 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.

REFERENCES

[1] N. Otsu , A Threshold Selection Method from Gray-Level Histogram, IEEE Transaction Systems, Man and Cybernetics, vol 9, no.1, pp.62-66, 1979.

[2] Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, “Script and language identification for handwritten document images”, IJDAR, vol.2, pp. 45-52, 1999.

[3] U. Pal and Chaudhuri B.B., “Script Line Separation from Indian Multi-Script Documents”, 5th ICDAR, pp.406-409,1999.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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] Gonzalez and Woods, Digital Image processing, 3/e, Pearson Education, 2008.

[10] 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.

[11] 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.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 4, October 2013

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

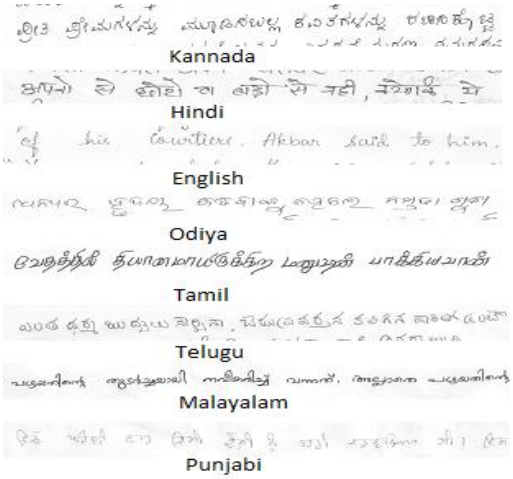

Fig 1: Sample handwritten line images in different scripts (in binary).

Fig 2: Pipeline process for feature extraction.

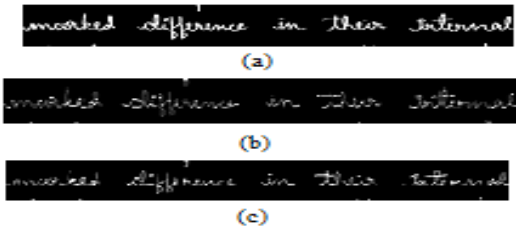

Fig 3: Gabor filtered images for zero degree orientation and frequencies a, b, and c

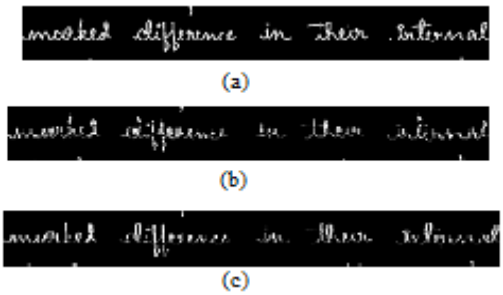

Fig 4: Gabor filtered images for 30 degree orientation and frequencies a, b, and c

Table 1: Recognition results of bi-script line images using KNN classifier

Bi-scripts	DCT applied to Gabor convolved images (Algorithm 1-- 54 Features Algorithm 2-- 19 Features)	Wavelets applied to Gabor convolved images (Algorithm 1-- 54 Features Algorithm 3-- 80 Features)
Kannada, English	90%	93%
Hindi, English	99%	99%
Malayalam, English	99%	99%
Punjabi, English	97%	97%
Tamil, English	97%	94%
Odiya, English	97%	97%
Telugu, English	94%	97%
Bengali, English	96%	98%

Table 2: Recognition results of bi-script line images using SVM classifier

Bi-scripts	DCT applied to Gabor convolved images (Algorithm 1-- 54 Features Algorithm 2-- 19 Features)	Wavelets applied to Gabor convolved images (Algorithm 1-- 54 Features Algorithm 3-- 80 Features)
Kannada, English	90%	92%

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 4, October 2013

Hindi, English	99%	99%
Malayalam, English	97%	99%
Punjabi, English	95%	90%
Tamil, English	97%	94%
Telugu, English	97%	98%
Bengali, English	90%	96%
Odiya, English	95%	99%