Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Script Identification in Printed Indian Documents

Hamsaveni L*
Department of Studies in Computer Science, Manasagangotri University of Mysore, Mysore, India

Pradeep C
Department of Computer Science and Engineering, Rajarajeswari College of Engineering, Bangalore, India

Chethan H K
Department of Computer Science and Engineering, Maharaja Institute of Technology, Mysore, India

Abstract- Automatic identification of a script in a given document image facilitates many important applications such as automatic archiving of multilingual documents, searching online archives of document images and for the selection of script specific OCR in a multilingual environment. In this paper, we present a scheme to identify different Indian scripts from a document image. For given script we extracted different features like Gray Level Co-occurrence Method (GLCM) and Scale invariant feature transform (SIFT) features. The features are extracted globally from a given text block which does not require any complex and reliable segmentation of the document image into lines and characters. The features are fed into Nearest Neighbor classifier. Thus the proposed scheme is efficient and can be used for many practical applications which require processing large volumes of data. The scheme has been tested on 10 Indian scripts and found to be robust in the process of scanning and relatively insensitive to change in font size. The performance of script classification is measured using precision and recall. This proposed system achieves good classification accuracy on a large testing data set.

Keywords—SIFT, GLCM, Nearest Neighbour

I. INTRODUCTION

Document image analysis has been an active research area from a few decades, and that facilitates the establishment of paperless offices across the world. The process of converting textual symbols present on printed and/ or handwritten paper to a machine understandable format is known as optical character recognition (OCR) which is the core of the field of document image analysis. The OCR technology for Indian documents is in emerging stage and most of these Indian OCR systems can read the documents written in only a single script. As per the trilingual formula of Indian constitution [1], every state Government has to produce an official document containing a national language (Hindi), official language (English) and state language (or regional language).

According to the three-language policy adopted by most of the Indian states, the documents produced in an Indian state Karnataka, are composed of texts in the regional language-Kannada, the National language-Hindi and the world wide commonly used language-English. In addition, majority of the documents found in most of the private and Government sectors of Indian states, are tri-lingual type (a document having text in three languages). So, there is a growing demand to automatically process these tri-lingual documents in every state in India, including Karnataka.

The monolingual OCR systems will not process such multi-script documents without human involvement for delineating different script zones of multi-lingual pages before activating the script specific OCR engine. The need for such manual involvement can result in greater expense and crucially delays the overall image-to-text conversion. Thus, an automatic forwarding is required for the incoming document images to handover this to the particular OCR engine depending on the knowledge of the intrinsic scripts. In view of this, identification of script and/ or language is one of the elementary tasks for multi-script document processing. A script recognizer, therefore, simplifies the task of OCR by enhancing the accuracy of recognition and reducing the computational complexity.

II. PREVIOUS WORK

Existing works on automatic script identification are classified into either local approach or global approach. Local approaches extract the features from a list of connected components like line, word and character in the document images and hence they are well suited to the documents where the script type differs at line or word level. In contrast, global approaches employ analysis of regions comprising of at least two lines and hence do not require fine segmentation. Global approaches are applicable to those documents where the whole document or paragraph or a set of text lines is in one script only. The script identification task is simplified and performed faster with the global rather than the local approach. A sample work has been reported in literature on both Indian and non-Indian scripts using local and global approaches.

A. Local approaches on Indian scripts

Pal and Choudhuri [2] have proposed an automatic technique of separating the text lines from 12 Indian scripts (English, Hindi, Bangla, Gujarati, Tamil, Kashmiri, Malayalam, Oriya, Punjabi, Telugu and Urdu) using ten triplets formed by

grouping English and Devanagari with any one of the other scripts. This method works only when the triplet type of the document is known. Script identification technique explored by Pal [3] uses a binary tree classifier for 12 Indian scripts using a large set of features. B Patil and Subbareddy [4] have proposed a neural network based system for script identification of Kannada, Hindi and English languages. Dhandra et al., [5] have exploited the use of discriminating features (aspect ratio, strokes, eccentricity, etc,) as a tool for determining the script at word level in a bi-lingual document containing Kannada, Tamil and Devnagari containing English numerals. A method to automatically separate text lines of Roman, Devanagari and Telugu scripts has been proposed by Pal et al., [6]. In Lijun et al, [7] have developed a method for Bangla and English script identification based on the analysis of connected component profiles. Vipin [8] have presented an approach to automatically identify Kannada, Hindi and English languages using a set of features viz., cavity analysis, end point analysis, corner point analysis, line based analysis and Kannada base character analysis. Word-wise script identification systems for Indian scripts has been discussed in [24].

B. Global approaches on Indian scripts

Adequate amount of work has been reported in literature using global approaches. S Chaudhury et al., [9] has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu. G D Joshi et al., [10] have presented a script identification technique for 10 Indian scripts using a set of features extracted from logGabor filters. Dhanya et al., [11] have used Linear Support Vector Machine (LSVM), K-Nearest Neighbour (K-NN) and Neural Network (NN) classifiers on Gabor-based and zoning features to classify Tamil and English scripts. Hiremath [12] have proposed a novel approach for script identification of South Indian scripts using wavelet based co-occurrence histogram features. Ramachandra and Biswas [13] have proposed a method based on rotation invariant texture features using multi channel Gabor filter for identifying seven Indian languages namely Bengali, Kannada, Malayalam, Oriya, Telugu and Marathi. S R Kunte and S Samuel [14] have suggested a neural approach in on-line script recognition for Telugu language employing wavelet features. Nagabhushan et al., [15] have presented an intelligent pin code script identification methodology based on texture analysis using modified invariant moments. Peeta et al., [16] have presented a technique using Gabor filters for script identification of Indian bilingual documents.

C. Local and global approaches on non-Indian scripts

Sufficient amount of work has also been carried out on non-Indian languages. Spitz [17] has proposed a system, which relies on specific, well defined pixel structures for script identification. Such features include locations and numbers of upward concavities in the script image, optical density of connected components, the frequency and combination of relative character heights. This approach has been shown to be successful in distinguishing between Asian languages (Japanese, Chinese, and Korean) against European languages (English, French, German, and Russian). Wood et al., [18] have proposed projection profile method to determine Roman, Russian, Arabic, Korean and Chinese characters. Hochberg et al., [19] have presented a method for automatically identifying script from a binary document image using cluster-based text symbol templates. In Ding et al., [20], a method that uses a combined analysis of several discriminating statistical features to classify Oriental and European scripts is presented. Tan et al., [21] has proposed a rotation invariant texture feature extraction method for automatic script and language identification from document images using multiple channel (Gabor) filters and Gray level co-occurrence matrices for seven languages: Chinese, English, Greek, Koreans, Malayalam, Persian and Russian. A Busch et al., [22] has presented the use of texture features (gray level co-occurrence matrix and Gabor energy features) for determining the script of a document image. B.Kumar et al. [23] have used topological, structural features with rule based classifier for line based multi-script identification.

It can be seen from the references cited above that sample amount of work has been done in the area of document script/language identification. Even though some considerable amount of work has been carried out on Indian script identification, hardly few attempts focus on the all the languages. So, an intensive work needs to be done in this field as the demand is increasing. Also the existing methods have to be improved to reach a stage of satisfactory practical application. It is in this direction the research work proposes a model that automatically identifies the all the languages in given document. We propose a based classification scheme which uses a global approach and demonstrate its ability to classify 10 Indian language scripts. In section (3), we describe the preprocessing scheme in detail. Feature extraction is presented in section 4. Results of the scheme tested over a large data set are presented in section (5).

III. PREPROCESSING

Our scheme first segments the text area from the document image by removing the upper, lower, left and right blank regions. After this stage, we have an image which has textual and non-textual regions. This is then binarised after removing the graphics and pictures (at present the removal of non-textual information is performed manually, though page segmentation algorithms such as [12] could be readily been employed to perform this automatically). Text blocks of predefined size (100×200 pixels) are next extracted. It should be noted that the text block may contain lines with different font sizes and variable spaces between lines words and characters. Numerals may appear in the text.

IV. FEATURE EXTRACTION

Feature extraction is a necessary step for any classification task. For image object classification purpose, the use of texture and shape features has proved to be quite effective for many applications. There are many ways for calculating texture feature descriptors. GLCM is one of them. Many descriptors can be obtained from the co-occurrence matrix

calculated. The SIFT based descriptors describes a given object with respect to a set of interesting points which are invariant to scale, translation, partial occlusion and clutter. These feature descriptors have been used successfully for object recognition, robotic mapping etc.

In our work, for each script, we computed 4 texture features, contrast, homogeneity, correlation and energy. For each object, the SIFT algorithm generates a feature vector of 128 elements. So each image object is now represented by a feature vector of 132 elements.

A. GLCM Based Texture Feature Descriptors

Texture features based on spatial co-occurrence of pixel values are probably the most widely used texture feature descriptors having been used in several application domains like analysis of remotely sensed images, image segmentation etc. Cooccurrence texture features are extracted from an image into two steps. First, pair wise spatial co-occurrence of pixels separated by a given angular value are computed and stored in a grey level co-occurrence matrix. Second, the GLCM is used to compute a set of scalar quantities that characterizes the different aspects of the underlying texture. We have worked with four GLCM based descriptors, namely, Contrast, Correlation, Homogeneity and Energy [26].

B. SIFT Feature Descriptors

In computer vision, SIFT is used to detect and describe local features in an image. SIFT features are used for reliable matching between different views of the same object. The extracted features are invariant to scale, orientation and are partially invariant to illumination changes. The SIFT feature extraction is a four step process. In the first step, locations of the potential interest points are computed in the image by finding the extremas in a set of Difference of Gaussian (DOG) filters applied to the actual image at different scale-space. Then those interest points which are located at the areas of low brightness and along the edges are discarded. After that an orientation is assigned to the remaining points based on local image gradients. Finally local image features based on image gradient is calculated at the neighboring regions of each of the key points. Every feature is defined in the 4 x 4 neighborhoods of the key points and is a vector of 128 elements [27].

V. EXPERIMENTS AND RESULTS

A. Data Collection

At present, in India, standard databases of Indian scripts are unavailable. Hence, data for training and testing the classification scheme was collected from different sources. These sources include the regional newspapers available online [24] and scanned document images in a digital library [25].

1) Indian Language Scripts

India has 18 official languages which includes Assamese, Bangla, English, Gujarati, Hindi, Konkanai, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Rajasthani, Sanakrit, Tamil, Telugu and Urdu. All the Indian languages do not have the unique scripts. Some of them use the same script. For example, languages such as Hindi, Marathi, Rajasthani, Sanskrit and Nepali are written using the Devanagari script; 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. In all, ten different scripts are used to write these 18 languages. These scripts are named as Bangla, Devanagari, Roman(English), Gurumukhi, Gujarati, Malayalam, Oriya, Tamil, Kannada and Urdu. The image blocks of these images are shown in Fig. 1.

B. Nearest Neighbour (NN)

One of the simplest classifiers which we used is the Nearest Neighbour classifier [28][29]. The term of nearest can be taken to mean the smallest Euclidean distances in n-dimensional feature space. This takes a test sample feature in vector form, and finds the Euclidean distance between this and the vector representation of each training example. The training sample closest to the test sample is termed its Nearest Neighbour. Since the trained sample in some sense is the one most similar to our test sample, it makes sense to allocate its class label to the test sample. This exploits the 'smoothness' assumption that samples near each other are likely to have the same class.

C. Results

We have performed experiments with different types of images such as normal, bold, thin, small, big, etc. The training and testing set comprises of more than 100 samples. We have kept the same data file for testing and training for the classifier to analyze the result. In most of the documents the occurrence of Roman characters is very few as compared to that of other characters. Table 1 and Table 2 tabulate the confusion matrix obtained for both GLCM and SIFT feature using Nearest Neighbor classifier. We calculate precision, recall and F–measure for various scripts based on the confusion matrix. Figure 2 and Figure 3 gives a graphical representation of precision, recall, and F-measure values.

Precision= Number of Relevant / Total number of relevant

Recall= Number of Relevant / Total number of classified

Fmeasure=2. Precision.Recall /Precision.Recall

VI. CONCULSION

Based on our observation human ability to classify unfamiliar scripts we have examined the possibility of using only global analysis of scripts for identifying them. We have presented a set of features like GLCM and SIFT for accomplishing classification. These features have been used to develop a script classification scheme for Indian language

scripts. . It requires a very simple preprocessing followed by a feature extraction process. Test results of the proposed classification scheme have revealed that good performance accuracy (98%) is obtainable using global analysis thereby illustrating its strength and utility. The scheme can be extended to multiple scales to handle scripts printed at a different resolution. The proposed scheme can be used for other language scripts as well with minimal modification

REFERENCES

[1] U. Pal, S. Sinha and B. B.Chaudhri, (2003), Multi-Script Line Identification from Indian Documents, Proceedings of International Conference on Document Analysis and Recognition, pp. 880-884.

[2] Pal U., Chaudhuri B.B., (1999), Script line separation from Indian multi-script document, Proc. 5th Int. Conf. on Document Analysis and Recognition (IEEE Comput. Soc. Press), 406–409.

[3] Pal U. and Chaudhuri B.B., (2003), Script line identification from Multi script documents, IETE journal Vol. 49, No 1, 3-11.

[4] Basavaraj Patil S. and Subbareddy N.V., (2002), Neural network based system for script identification in Indian documents, Sadhana Vol. 27, Part 1, 83–97.

[5] Dhandra B.V., Nagabhushan P., Mallikarjun Hangarge, Ravindra Hegadi, Malemath V.S., (2006), Script Identification Based on Morphological Reconstruction in Document Images, The 18th International Conference on Pattern Recognition (ICPR'06), Vol.No. 11-3, 950-953.

[6] Pal U., Chaudhuri B.B., (1999), Script line separation from Indian multi-script document, Proc. 5th Int. Conf. on Document Analysis and Recognition (IEEE Comput. Soc. Press), 406–409.

[7] Lijun Zhou, Yue Lu and Chew Lim Tan, (2006), Bangla/English Script Identification based on Analysis of Connected component Profiles, Proc. 7th IAPR workshop on Document Analysis System, New land, 234-254.

[8] Vipin Gupta, G.N. Rathna, K.R. Ramakrishnan, (2006), A Novel Approach to Automatic Identification of Kannada, English and Hindi Words from a Trilingual Document, Int. conf. on Signal and Image Processing, Hubli, pp. 561-566.

[9] Santanu Chaudhury, Gaurav Harit, Shekar Madnani, Shet R.B., (2000), Identification of scripts of Indian languages by Combining trainable classifiers", Proc. of ICVGIP, India.

[10] Gopal Datt Joshi, Saurabh Garg, and Jayanthi Sivaswamy, (2006), Script Identification from Indian Documents, H. Bunke and A.L. Spitz (Eds.): DAS, LNCS 3872, 255–267.

[11] Dhanya D., Ramakrishnan A.G. and Pati P.B., (2002), Script identification in printed bilingual documents, Sadhana, vol. 27, 73-82.

[12] Hiremath P S and S Shivashankar, (2008), Wavelet Based Co-occurrence Histogram Features for Texture Classification with an Application to Script Identification in a Document Image, Pattern Recognition Letters 29, pp 1182-1189.

[13] Srinivas Rao Kunte R. and Sudhakar Samuel R.D., (2002), A Neural Approach in On-line Script Recognition for Telugu Language Employing Wavelet Features, National Workshop on Computer Vision, Graphics and Image Processing (WVGIP), 188-191.

[14] Peeta Basa Pati, S. Sabari Raju, Nishikanta Pati and A. G. Ramakrishnan, (2004), Gabor filters for Document analysis in Indian Bilingual Documents, 0-7803-8243-9/04/ IEEE, IClSlP, pp. 123- 126.

[15] Spitz A. L., (1994), Script and language determination from document images, Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 229-235.

[16] Wood S. L.; Yao X.; Krishnamurthy K. and Dang L., (1995), Language identification for printed text independent of segmentation, Proc. Int. Conf. on Image Processing, 428–431, IEEE 0- 8186-7310-9/95.

[17] Hochberg J., Kerns L., Kelly P. and Thomas T., (1997), Automatic script identification from images using cluster based templates, IEEE Trans. Pattern Anal. Machine Intell. Vol. 19, No. 2, 176–181.

[18] Ding J., Lam L. and Suen C. Y., (1997), Classification of oriental and European Scripts by using Characteristic features, Proc. 4th ICDAR , 1023-1027.

[19] Tan T. N., (1998), Rotation invariant texture features and their use in automatic script identification, IEEE Trans. Pattern Anal. Machine Intell. PAMI, Vol.20, No. 7, 751–756.

[20] Andrew Busch; Wageeh W. Boles and Sridha Sridharan, (2005), Texture for Script Identification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 11, pp. 1720-1732.

[21] B. Kumar, A. Bera and T. Patnaik, (2012), Line Based Robust Script Identification for Indian Languages, International Journal of Information and Electronics Engineering, vol. 2, pp. 189-192.

[22] R. Rani, R. Dhir and G. S. Lehal, (2013),Modified Gabor Feature Extraction Method for Word Level Script Identification- Experimentation with Gurumukhi and English Scripts, International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6, no. 5, pp. 25-38.

[23] A. K. Jain and Y. Zhong., (1996), Page segmentation using texture analysis. Pattern Recognition 29, 743–770.

[24] http://www.samachar.com/.

[25] Digital Library of India. http://dli.iiit.ac.in/

[26] R. M. Haralick, K. Shanmugam, and I. Dinstein, (1973), Textural Features of Image Classification, IEEE Transactions on Systems, Man and Cybernetics, % vol. SMC-3, no. 6.

[27] Lowe, D. G., (2004), Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 60, 2, pp. 91-110.

[28] Hall P, Park BU, Samworth R J . (2008), Choice of neighbor order in nearest-neighbor classification. Annals of Statistics36 (5): 2135–2152.

[29] Bremner D, Demaine E, Erickson J, Iacono J, Langerman S, Morin P, Toussaint G., (2005), Output-sensitive algorithms for computing nearest-neighbor decision boundaries. Discrete and Computational Geometry33 (4): 593–604.

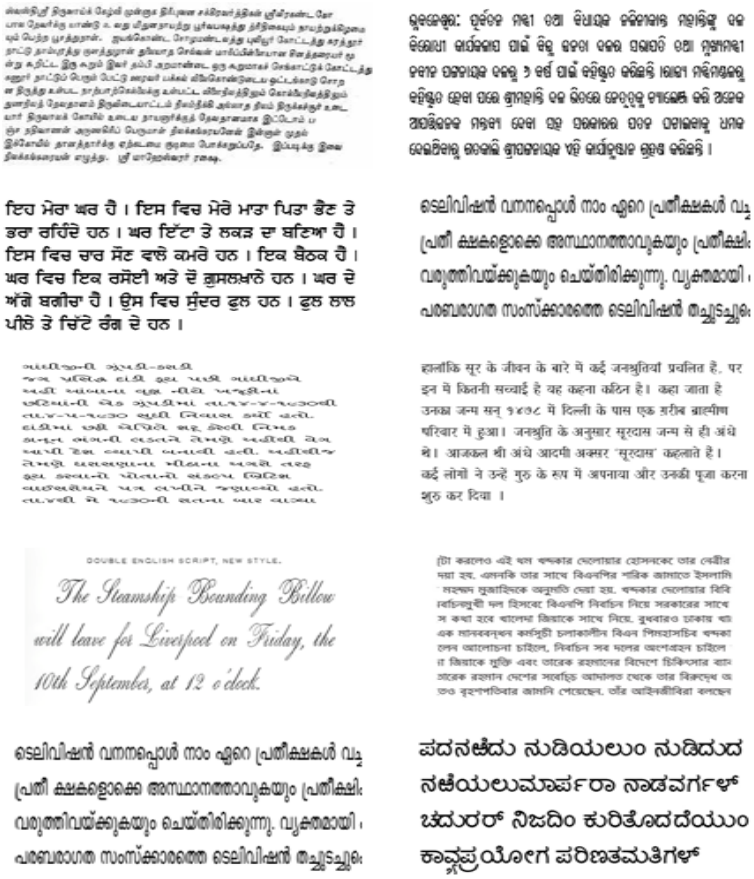

Figure 1. Indian Language Scripts

Table 1. Confusion Matrix of GLCM features for 10 different Indian scripts. (Here De=Devanagari, Ba=Bangla, Ta= Tamil, Gu= Gurumukhi, Ka= Kannada, Ma=Malayalam, Ro= Roman (English), Or= Oriya, Guj= Gujarati, Ur= Urdu.)

Actual	De	Ba	Ta	Gu	Ka	Ma	Ro	Or	Guj	Ur
De	92			04	04					
Ba		89		06				05		
Ta			84		05	06		05		
Gu	04	08		82						06
Ka	05		05		85	05				
Ma			05		05	80	05		05	
Ro		05	06				79	10		
Or					05	02	03	85	05	
Guj				04					92	04
Ur								04	04	92

Table 2. Confusion Matrix of GLCM features for 10 different Indian scripts.

Actual	De	Ba	Ta	Gu	Ka	Ma	Ro	Or	Guj	Ur
De	95			5						
Ba		92		8						
Ta		02	85	13						
Gu	02		10	88						
Ka					98	01		01		
Ma						90		10		
Ro		05	05				81	09		
Or					02	02	02	90	04	
Guj				01					98	01
Ur									01	99

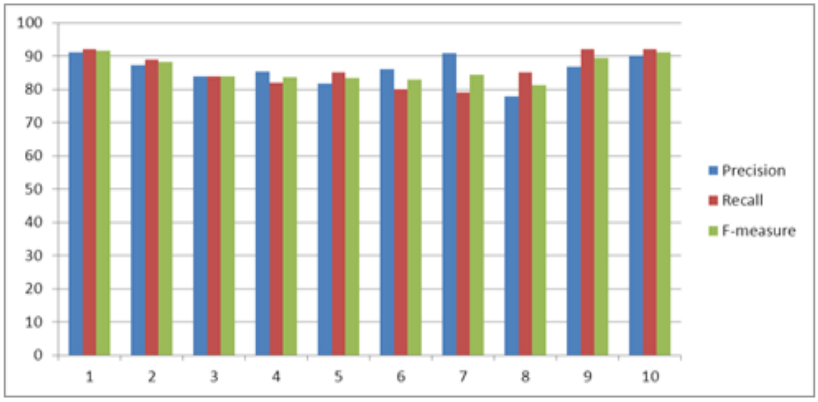

Figure 2.Shows accuracy of precision recall and F-measure for GLCM features

Figure 3. Shows accuracy of precision recall and F-measure for SIFT features

