World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:5, No:1, 2011

Identification of Printed Punjabi Words and English Numerals Using Gabor Features

Rajneesh Rani, Renu Dhir, and G.S. Lehal

***Abstract*—**Script identification is one of the challenging steps in the development of optical character recognition system for bilingual or multilingual documents. In this paper an attempt is made for identification of English numerals at word level from Punjabi documents by using Gabor features. The support vector machine (SVM) classifier with five fold cross validation is used to classify the word images. The results obtained are quite encouraging. Average accuracy with RBF kernel, Polynomial and Linear Kernel functions comes out to be greater than 99%.

***Keywords*—**Script Identification, Gabor Features, Supprt Vector Machines

I. INTRODUCTION

OPTICAL Character Recognition (OCR) is one of the important tasks of machine learning. OCR means recognition of machine printed text by computer and its conversion to an editable form, which can be further processed as per the requirements. All official documents, magazines and reports can be converted to electronic form using a high performance OCR system.

Development of a OCR is a great challenge for a multilingual country like India where the documents contain more than one language. Generally, there are two kinds of approaches for developing this type of system. One approach is combined database approach [1]. That is the database of reference characters has alphabets from all of its languages in which the document is printed. So database is larger at the recognition level of individual character. The second approach is based on the identification of the script of each character before taking the characters for recognition. This helps in reduced search in the database at the cost of script recognition task. A number of techniques have been developed for determining the script of printed/handwritten documents and these can be typically classified into four categories [2, 3]: a) connected component analysis b) text block level analysis c) text line level analysis d) word and character level analysis.

It has been revealed from literature survey that a considerable amount of research has been done for script identification. Wood *et al.* [4] proposed a method based on projection profile to determine Arabic, Russian, Korean, Roman and Chinese Languages. Hochberg *at al.* [5] discussed a method for script identification based on cluster based templates. Spitz [6] determined the script of Asian and European languages by examining the upward concavities of connected components. Tan *et al.* [7] described a method based on texture analysis for automatic script and language identification using multi channel Gabor filters and co-occurrence matrices for seven languages: Chinese, English, Koreans, Greek, Malayalam, Persian and Russian.

Due to diversity of languages/scripts English has proven to be the binding language in India. Indian documents normally contain text words in its state language and numerals in English. Normally for these types of documents word level script identification is required.

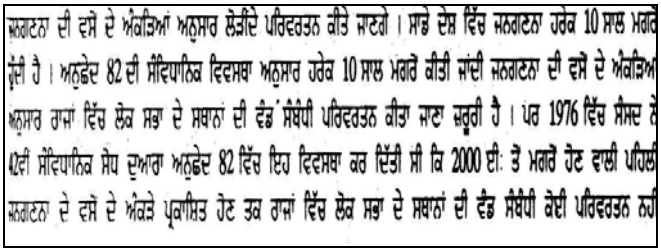

Fig. 1. Sample bi-script document image showing interspersed English Numerals and Punjabi Words

From the literature survey, it is clear that a considerable amount of work has been done for script identification at word level for Indian languages. The algorithms proposed are by Dhanya e*t al.* [8] identified Roman and Tamil Scripts using Gabor filters and spatial spread features, Pal e*t al.* [9] described a method based on water reservoir, conventional, topological and structural features,to identify the Devanagri, English and Telgu words. Padma e*t al.* [10] proposed a method based on discriminating features to identify Kannada ,Hindi and English text words. Pati e*t al.* [11,12,13] have proposed word level script identification based on 32 Gabor features using Gabor filters. These algorithms deal with only text word separation not with English Numeral Separation from multilingual documents. Dhandra e*t al.* [14,15,16] e proposed the methods for English Numeral separation based on morphological reconstruction. But the work done is on South Indian languages, Punjabi language is ignored here. For Punjabi documents, Roman words have been separated as in [17], but again English numerals have not been identified here. Sharma e*t al.* [18] proposed a rejection based method for digit extraction from Gurumukhi (Punjabi) documents. Individual digits are extracted from the documents not the whole numeral as a word using this method. Here, we also made an attempt to identify script of English Numerals and Punjabi words through Gabor Features and SVM classifier.

Rajneesh Ranir is Assistant Professor in the National Institute of Technology Jalandhar, Punjab, India , e-mail: ranir@ nitj.ac.in.

Renu Dhir is Assosciate Professor in the National Institute of Technology Jalandhar, Punjab, India , e-mail: dhirr@ nitj.ac.in.

G,.S Lehal is Professor in the department of Computer Science and Engineering at Punjabi University Patiala., Punjab, India e-mail-gslehal@gmail.com

The paper is organized as follows. Introduction to Gabor Filters and features extraction technique using these has been described in section II. The section III contains a proposed script identification system. The experiment details and results obtained are presented in section IV. Conclusion and Discussions are given in section V.

II. Gabor Filters

Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features that are effective in discriminating pattern classes [8].

Gabor filters can be used as a directional feature extractor. These filters can effectively capture the concentration of energies in various directions. Human Visual system is sensitive to spatial orientation with an approximate angular bandwidth of 30 degree and spatial frequency. Gabor filters can be modeled to closely resemble the HVS. These modeled filters are capable of providing multi-resolution analysis and can be used to extract directional features.

A Gabor Filter is a linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function [19].

$$h(x,y) = g(x,y)s(x,y) \quad (1)$$

Where $s(x,y)$ is a complex sinusoid, known as carrier and $g(x,y)$ is a Gaussian shaped function, known as envelope [19]. Thus the 2-D Gabor filter can be written as

$$h_{x,y,\theta,f} = e^{-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2}\right)} . e^{j2\pi fx} \quad (2)$$

Where σ_x and σ_y explain the spatial spread and are the standard deviations of the Gaussian envelope along x and y directions. x' and y' are the x and y co-ordinates in the rotated rectangular co-ordinate system given as

$$x' = x\cos\theta + y\sin\theta \quad (3)$$

$$y' = y\cos\theta - x\sin\theta \quad (4)$$

Any combination of θ and f, involves two filters, one corresponding to sine function and other corresponding to cosine function in exponential term in Equation 2. The cosine filter, also known as the real part of the filter function, is an even symmetric filter and acts like a low pass filter, while the sine part being odd-symmetric acts like a high pass filter.

Gabor filters having Spatial frequency (f = 0.0625, 0.125, 0.25, 0.5, 1.0) and orientation $(\theta = \frac{n\pi}{6})$ where n varies in the range 0 to 6, have been used in our reported work.

III. Script Identification System Architecture

The architecture of script identification system for Punjabi words and English numerals on document images is shown in Figure 2.

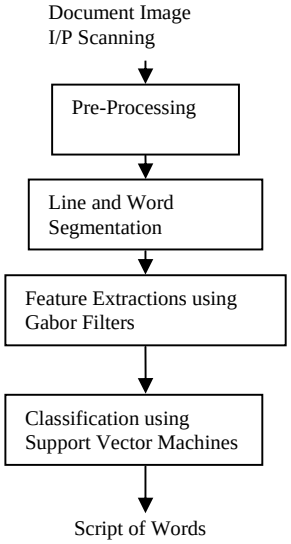

Fig. 2. Script Identification System Architecture

The system in the reported work is composed of following phases:

- Segmentation
- Feature Extraction
- Classification

The documents considered are after preprocessing that is after digitization, removing noise and skewness.

A. Segmentation

In present work, the segmentation process is performed in two successive stages: line segmentation and word segmentation. For line and word segmentation horizontal and vertical projection profiles are respectively used. A text line is located between scan lines whose horizontal projection profile histogram values are greater than some threshold value. After a text line is detected, its vertical projection profile is determined and if a number of successive vertical projection profiles are greater than some predefined threshold value, a word is considered to exist between these vertical lines.

B. Feature Extraction

In the proposed system, multi-bank Gabor filters having five different values for Spatial frequency (f = 0.0625, 0.125, 0.25, 0.5, 1.0) and six different values for orientation $(\theta = 0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ)$ are chosen to give a total of 70 Gabor filters with a combination of 35 even and 35 odd filters. From the output of each Gabor filter mean and standard deviation are computed, which serves as Gabor features. Thus for each word we get a feature vector of 140 values given by

$$F = [\mu_1, \sigma_1, \mu_1, \sigma_1, \mu_1, \sigma_1 \ldots\ldots\ldots \mu_{70}, \sigma_{70}]$$

C. Script Classification

The objective of classification is to identify the script of words taken form the test set. Features extracted from the words are sent to the Classifier. In the report work SVM (Support Vector Machine) classifier has been used to identify the script.

SVM is a kind of learning machine whose fundamental is statistics learning theory. SVM is basically used to solve two class problems. For these, it finds the optimal hyper-plane which maximizes the distance, the margin, between the nearest examples of both classes, named support vectors (SVs). Given a training database of N data: $[x_n \mid n = 1, 2 \ldots\ldots N]$, the linear SVM classifier is give as

$$F(x) = \sum \alpha_j x_j . x + b$$

where x_j are the set of support vectors and the parameters α_j and b have been determined by solving a quadratic problem[20].If the data is non linear, there arises the need of mapping the data to higher dimensional feature space by function ϕ . So the linear classifier is extended to non linear classifier by computing the dot product in the input space rather than in the feature space via constructing a kernel function. SVM operate on kernel evaluations of the feature vectors x_i or x_j given as

$$K(x_i, x_j) = \phi(x_i^T).\phi(x_j)$$

Variant learning machines are constructed according to different kernel functions ad thus constructs different hyper planes in the feature space.

Different types of kernel functions used in the reported work are:

Linear Kernel: $K(x_i, x_j) = x_i^T x_j$

Polynomial Kernel: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$

RBF kernel: $K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0$

IV. Experimental Results

The proposed method is implemented in Matlab 7.4(R2007a). In order to investigate the effectiveness of the proposed method, data set of 4505 words has been created form various documents as described in the segmentation phase in section III. The documents considered contain only text lines. Documents are created in different fonts and printed from a laser printer. Then these documents are scanned. Fonts used are AnmolLipi and Anmol Kalmi for Punjabi words and Times New Roman and Calibri for English Numerals. So from all these documents 4505 words are segmented, out of which 1900 and 2605 are English Numerals and Punjabi words.

Feature vector of size 140 is computed for each of these words using Gabor Filters as explained in section II.. Then this feature vector has been used for classification using SVM. Experiments are carried with different kernel functions to determine the best kernel function. Fivefold cross-validation has been used for result computation with five iterations.

Fivefold defines the data set of 4505 words into five disjoint subsets each having 901 words. Here, four subsets are used for training and one is used for testing. So this process is repeated five times leaving one different subset for evaluation each time.

Table1 provides the details of recognition results for different subsets with different kernel functions.

TABLE I SCRIPT IDENTIFICATION RESULTS FOR ENGLISH NUMERALS AND PUNJABI WORDS

Input	Classification Accuracy with Different Kernel Functions in %		
	Linear Kernel	**Polynomial Kernel**	**RBF Kernel**
Fold1	99.89	99.89	95.22
Fold2	99.67	99.89	97.00
Fold3	99.89	99.89	96.67
Fold4	99.67	99.89	97.33
Fold5	99.67	99.78	97.22
Average Accuracy	99.75	99.86	96.68

The results obtained are quite encouraging and the best average accuracy is 99.86% with Polynomial Kernel function.

V. Conclusion

In this paper we have described a simple, novel and effective method for Punjabi text words and English numerals script identification. The aim of this paper is to facilitate the multilingual OCR. After segmenting the document in horizontal and vertical directions, a feature vector of size 140 is extracted from each word image using Gabor Filters. These features are passed to SVM classifiers for classification with different kernel functions. The polynomial kernel function gives the best results. The experimental results show that this scheme is effective to identify English Numerals and Punjabi words, which further helps to feed individual word to specific OCR system.

References

[1] D Dhanya and A G Ramakrishnan, “Simultaneous Recognition of Tamil and Roman Scripts”, in the Proc. Tamil Internet, Kuala Lumpur, pp. 64-68, 2001.

[2] Rajneesh Rani, Renu Dhir , “A Survey: Recognition of Scripts in Bi-Lingual/Multi-Lingual Indian Documents” in national journal of PIMT Journal of Research Vol. 2 No. 1 pp. 55-60 , March- August, 2009.

[3] S.Abirami, Dr. D. Manjula, “A Survey of Script Identification Techniques for Multi-Script Document Images” in international journal of Recent trends in Engineering Vol. 1 No. 2 pp. 246-249 May,2009.

[4] S.Wood, X.Yao, K.Krishnamurthi and L.Dang “language identification from for printrd trxt independent od fsegmentation,” Proc of International conference on Image Processing, pp. 428-431,1995.

[5] J.Hochberg, P.Kelly, T Thomas and L Kerns, “Automatic script identification from document images using cluster based templates,”

IEEE Trans. on Pattern Anaylsis and Machine Intelligence, vol 19, pp. 176-181, 1997.

[6] A.L.Spitz, "Determination of the script and language content of document images," IEEE Transactions on pattern Anaylsis and Machine Intelligence, Vol 19, pp.234-24,1997.

[7] T.N. Tan, "Rotation invariant textutre features and their use in automatic script identification," IEEE Trans on Pattern Anaylsis and Machine Intelligence, vol. 20, pp 751-756, 1998.

[8] D Dhanya, A.G Ramakrishnan and Peeta Basa pati, "Script identification in printed bilingual documents," Sadhana, vol. 27, part-1, pp. 73-82, 2002.

[9] U.Pal. S.Sinha and B.B Chaudhuri, "Word-wise Script identification from a document containing English ,Devnagari and Telgu Text," in the proc. of NCDAR, pp. 213-220,2003

[10] M.C. Padma , Dr. P.A. Vijya, " Language Identification of Kannada, Hindi and English Text Words through Visual Discriminating features", in the international journal of Computational Intelligence Systems, Vol.1 No.2 pp. 116-126, May -2008.

[11] Peeta Basa pati, S. Sabari Raju, Nishikanta Pati and A.G. Ramakrishnan, "Gabor filters for document analysis in Indian Bilingual Documents," In the Proc. Of ICISIP, pp. 123-126, 2004.

[12] Peeta Basa Pati and A.G.Ramakrishnan, "HVS inspired system for Script Identification in Indian Multi-Script Documents", In Proc. of 7th International Workshop on Document Analysis System, Nelson Newland, pp. 380-389, 2006

[13] Peeta Basa Pati, A.G. Ramakrishnan " Word level multi-script identification" in the Pattern Recognition Letters 29 pp. 1218-1219, 2008.

[14] B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi, V.S.Malemath, "Word-wise Script Identification from Bilingual Documents based on Morphological Reconstruction," in the proc. of First IEEE International Conference on Digital Information Management, pp. 389-394, 2006.

[15] B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi, V.S.Malemath, "Word–wise Script Identification based on Morphological Reconstruction in Printed Bilingual Documents," in the proc. of IET International Conference on Vision Information Engineering VIE, Bangalore pp. 389-393, 2006

[16] B.V.Dhandra, Mallikarjun Hangarge, " On Separation of English Numerals from Multilingual Document Images", In the journal of multimedia , Vol 2, No 6, pp. 26-33, 2007.

[17] Renu Dhir, Chandan Singh and G.S.Lehal, "A Structural Feature Based Approach for Script Identification of Gurmukhi and Roman Character and Words" in the proc. of 39th Annual National Convention of Computer Society of India (CSI) held at Mumbai, India, 2004

[18] Dharamveer Sharma, Gurpreet Singh Lehal, Preeti Kathuria ," Digit Extraction and Recognition from Machine printed Gurmukhi documents" in the Proc. Of International workshop on Multilingual Ocr Article no 12, 2009.

[19] R Anjeev Kunte and R D Sudhaker Samuel, " A Bilingual machine-Interface OCR fr Printed Kannada and English Text Employing Wavelet Features" in the prproc of 10th International Conference on Information Technology, pp.202-207, 2007.

[20] G.G.Rajput,S.M Mati, "Fourier Descriptor based Isolated Marathi Handwritten Numeral Rcognition" in International Journal od Computer Applications Vol. 3 No.4 pp.9-13,June=2010