Comparative Analysis of Gabor and Discriminating Feature Extraction Techniques for Script Identification

Rajneesh Rani[1], Renu Dhir[2], and G.S.lehal[3]
[1,2] Department of CSE, NIT Jalandhar,
Punjab, India. [1]ranir@nitj.ac.in, [2]dhirr@nitj.ac.in
[3]Department of CSE, Punjabi University,Patiala,
Punjab, India. [3]gslehal@gmail.com

Abstract. A considerable amount of success has been achieved in developing monolingual OCR systems for Indian Scripts. But in a country like India, where many languages and scripts exist, it is more common that a single document contain words from more than one script. Therefore a script identification system is required to select the appropriate OCR. This paper presents a comparative analysis of two different feature extraction techniques for script identification of each word. In this work, for script identification discriminating and Gabor filter based features are computed of Punjabi words and English numerals. Extracted feature are simulated with Knn and SVM classifiers to identify the script and then recognition rates are compared. It has been observed that by selecting the appropriate value of k and appropriate kernel function with appropriate combination of feature extraction and classification scheme, there is significant drop in error rate.

Keywords: Script Identification, Gabor Features, Discriminating Features, Support Vector Machines, Knn

1 Introduction

For a multilingual country like India where the documents contain more than one language, to develop an OCR is a great challenge. Mostly, two different kinds of techniques can be used to develop this type of system. One technique 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 technique 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 for determining the script of printed/handwritten documents can be typically classified into four categories [2, 3]: a) connected component based Script Identification b)Script Identification at text block level c) Script Identification at text line level d) Word level Script Identification.

Feature Extraction is an important phase for script identification system of a word. Feature Extraction has been defined as "Extracting from the raw data the information which most relevant for classification purposes, in the sense of minimizing the within-class pattern variability while enhancing the between-class pattern variability" [4].There are a number of techniques available for feature extraction for script identification [5-15]. Selection of a feature extraction technique is the single most important factor in achieving high performance of script identification systems. Gabor filters [10-12] can be used as a directional feature extractor. Other types of features are discriminating features [13-15] which means that every language can be identified based on its distinct visual appearance. These features can be extracted by using morphological reconstruction of an image. This paper presents a comparison of these two methods for identification of Punjabi words and English numerals.

The paper is organized as follows. The theory of Gabor filters and feature extraction using these is discussed in Section 2. Discriminating features of Punjabi words and English numerals have been described in Section 3. Section 4 deals with different classification techniques and finally Section 5 contains the experimental results and conclusion.

2 Gabor Filters

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

$$h(x, y) = g(x, y)s(x, y) \tag{1}$$

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

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

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

$$x' = x\cos\theta + y\sin\theta \tag{3}$$

$$y' = y\cos\theta - x\sin\theta \tag{4}$$

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.

In the present work, 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 $(\theta = 0^o, 30^o, 60^o, 90^o, 120^o, 150^o, 180^o)$ 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

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

3 Discriminating Features of Punjabi words and English Numerals

Punjabi words and English numerals have a distinct visual appearance as shown in Fig. 1.

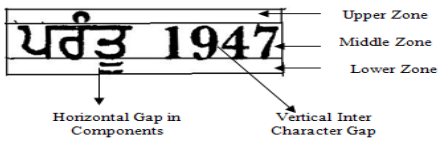

Fig 1. Sample image of Punjabi Word and English Numeral showing different Zones

After a careful study of shapes of Punjabi words and English numerals, nine features for automatic classification of English numerals and Punjabi words' script are:

F1: Average Aspect ratio (AAR): The average aspect ratio (AAR) is defined as:

$$AAR = \frac{1}{N}\sum_{i=1}^{N}\frac{height(component_i)}{width(component_i)} \quad (5)$$

Here N is the number of connected components of input word image.

F2: Average Eccentricity (AE): The average eccentricity (AE) is defined as

$$AE = \frac{1}{N}\sum_{i=1}^{N}\frac{len\_maj\_axis(component_i)}{len\_\min\_axis(component_i)} \quad (6)$$

Here N is the number of connected components of input word image.

F3-F6: Based on Stroke Density in a Direction (SD): Features F3, F4, F5 and F6 are based on stroke densities in vertical, horizontal, left diagonal and right diagonal directions. The stroke density in a direction is computed as:

$$SD = \frac{\sum_{i=1}^{N} no\_onpixels\_instroke_i}{size\_of\_word} \quad (7)$$

Here N is the number of strokes in that direction.

To extract the stroke density in a direction, we have performed the morphological opening operation on the input binary word/numeral image with line structuring element having length=k × Mean (Connected_components_Height) and angle depending on the direction.

F7: Pixel Ratio after Filling Holes (PRFH): For fill holes, we choose the marker image, f_m to be 0 everywhere except on the image border, where it is set to 1-f. Here f is the original image.

$$f_m(x\,y) = \begin{cases} 1\text{-}f(x,y) \text{ if } f(x,y) \text{is on the border of } f \\ \text{otherwise } 0. \end{cases} \tag{8}$$

PRFH is computed as:

$$PRFH = \frac{Sumof\_onpixels\_afterfillhole}{size\_of\_word} \tag{9}$$

F8: Vertical Inter Character Gap (VICG): To extract the value of this feature, vertical projection histogram is taken of the image. If any vertical projection profile value is equal to zero then that means there is a gap between two characters and the value of this feature is set to1 otherwise is set to 0.

F9: Horizontal Break in Components (HBIC): To extract the value of this feature, horizontal projection histogram is taken of the image. If any horizontal projection profile value is equal to zero then that means there is a gap between components of a word and the value of this feature is set to1 otherwise is set to 0.

4 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.

KNN (k nearest neighbor) Classification:

The k- nearest neighbor (k-nn) approach attempts to compute a classification function by examining the labeled training point sin n dimensional space. Then the Euclidean distance is calculated between the test point and all reference points q in order to find k nearest neighbors. A test sample is labeled with the same class label as the label of the majority of its K nearest neighbors. Nearest Neighbor is a special case of k-nn, where k=1.

SVM (Support Vector Machines) Classification:

SVM is a kind of learning machine whose fundamental is statistics learning theory. 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). If the data is nonlinear, there arises the need of mapping the data to higher dimensional feature space by function ϕ . So the linear classifier is extended to nonlinear classifier by computing the dot product in the input space rather than in the feature space via

constructing a kernel function. Variant learning machines are constructed according to different kernel functions and thus construct different hyper planes in the feature space. Different types of kernel functions used in the reported work are: Linear, RBF, Polynomial and Sigmoid

5 Experimental Results and Discussion

The experiments are done in Matlab 7.4(R2007a). In order to investigate the effectiveness of each method, data set of 4505 words has been created form various documents. 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.

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. Then the average accuracy is calculated.

Table 1 provides the details of recognition results for different subsets with different kernel functions using SVM.

Table 1: Script Identification Results Using SVM with Discriminating Features and Gabor Features

Input	Classification Accuracy with Different Kernel Functions in %			
	Linear Kernel	Polynomial Kernel	RBF Kernel	Sigmoid Kernel
Discriminating Features	97.23	94.96	95.53	93.85
Gabor Features	99.75	99.82	96.67	57.82

Table 2 provides the details of recognition results for different subsets with Knn with different values of K.

Table 2: Script Identification Results Using KNN with Discriminating Features and Gabor Features

Input	Classification Accuracy with Different Values of K			
	K=1	K=-3	K=5	K=7
Discriminating Features	99.02	99.13	98.98	98.93
Gabor Features	97.62	97.11	96.91	96.40

It has been observed that for discriminating features, KNN Classifier gives the better results and for Gabor features, SVM Classifier gives the better results. Again it has been observed that different kernel functions and different values of K, for each of features, give better results. However error rate is more for increasing the value of K

beyond 7. None gives 100% accuracy. So a combination of these classifiers and these feature extraction techniques can be used to get more accurate results.

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 and 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 and 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. P.A. Devijver and J. Kittler, "Pattern Recognition: A statistical Approach, London: prentice –Hall,1982".
5. 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.
6. D Dhanya, A.G Ramakrishnan and Peeta Basa pati, "Script identification in printed bilingual documents," Sadhana, vol. 27, part-1, pp. 73-82, 2002.
7. 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
8. M.C. Padma and 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.
9. 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
10. 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.
11. 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
12. Peeta Basa Pati and A.G. Ramakrishnan " Word level multi-script identification" in the Pattern Recognition Letters 29 pp. 1218-1219, 2008.
13. B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi, and 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.
14. B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi and 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
15. B.V.Dhandra and Mallikarjun Hangarge, " On Separation of English Numerals from Multilingual Document Images", In the journal of multimedia , Vol 2, No 6, pp. 26-33, 2007.