International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.6, No.5 (2013), pp.25-38
http://dx.doi.org/10.14257/ijsip.2013.6.5.03

# Modified Gabor Feature Extraction Method for Word Level Script Identification- Experimentation with Gurumukhi and English Scripts

Rajneesh Rani[1], Renu Dhir[1] and Gurpreet Singh Lehal[2]

[1]*Dr B R Ambedkar National Institute of Technology Jalandhar, Jalandhar-144011, India*
*Department of Computer Science and Engineering, Punjabi University Patiala-147002, India*
*ranir@nitj.ac.in, dhirr@nitj.ac.in, gslehal@gmail.com*

***Abstract***

*Script Identification is one of the challenging step in the Optical Character Recognition system for multi-script documents. In Indian and Non-Indian context some results have been reported, but research in this field is still emerging. This paper presents a research work in the identification of Gurmukhi and English scripts at word level. It also identifies English Numerals from Gurmukhi text. Gabor feature extraction is one of most popular method for script recognition. This paper presents a zone based gabor feature extraction technique. The given word image after normalization is divided into different zones of different sizes and then features from each of these zones are extracted in various directions using gabor filters. Script is then determined by using SVM classifier. The experimental tests carried out in the field of Gurmukhi and English Script recognition show that the proposed technique leads to improvement over the traditional Gabor feature extraction without zoning. In future, this can also be extended for other scripts.*

***Keywords:*** *Filters; Zoning; Script Identification; SVM Classifier*

## 1. Introduction

Creating a paperless environment is in its great demand because of emergence and widespread applications of computers and multimedia technologies [1]. Document analysis and recognition is required after converting all the printed documents into images. A lot of research has been done on Optical Character Recognition in last 100 years [2]. However, most systems are script specific in the sense that they can read characters in one particular script only. *Script* is defined as the graphic form of the writing system used to write statements expressible in language [3]. The documents like language translation books, passport application form, exmination question papers and most official documents contain words from more than one script. Therefore, in this multilingual and multiscript world, OCR systems need to be capable of recognizing characters irrespective of the script in which these are written.If all the characters of different scripts are handled simultaneously at the classification stage, then accuracy of the overall OCR system will decrease. So for successful implementation of OCR system for multi-script documents, it is necessary to separate regions of different scripts in the document before feeding these to individual OCR systems. The regions of different scripts in a document can be paragraphs, lines, words and characters. This addresses the need of developing script identification techniques at paragraph, line and word level [4].

ISSN: 2005-4254 IJSIP

In India, a document contain mostly words in its state language mixed with English words and numerals as shown in Figure 1, where state language is Punjabi language written in Gurmukhi script. This has motivated us to identify the script of text words and English numerals in multilingual document images.


ਲੋਕ ਸਭਾ ਦੀ ਰਚਨਾ **Composition of the Lok Sabha** ਲੋਕ ਸਭਾ ਨੂੰ ਹੇਠਲਾ ਸਦਨ ਵੀ ਕਿਹਾ ਜਾਦਾ ਹੈ ਇਸ ਹੇਠਲੇ ਸਦਨ ਭਾਵ ਲੋਕ ਸਭਾ ਦੇ ਚੁਣੇ ਹੋਏ ਮੈਬਰਾ ਦੀ ਵੱਧ ਤੋ ਵੱਧ ਸੰਖਿਆ 550 ਨਿਸਚਿਤ ਕੀਤੀ ਗਈ ਹੈ ਇਹਨਾ 550 ਮੈਬਰਾ ਵਿਚੋ 530 ਮੈਬਰ ਰਾਜਾ ਦੀ ਅਤੇ 20 ਮੈਬਰ ਸੰਘੀ ਖੇਤਰਾ **Union Territories** ਦੀ ਪ੍ਰਤੀਨਿਧਤਾ ਕਰਦੇ ਹਨ ਦੂਜੇ ਸਬਦਾ ਵਿਚ ਭਾਰਤੀ ਸੰਘ ਦੇ 28 ਰਾਜਾ ਵਿਚੋ ਵੱਧ ਤੋ ਵੱਧ 530 ਅਤੇ 7 ਸੰਘ ਖਤਰਾ ਵਿਚੋ ਵੱਧ ਤੋ ਵੱਧ 20 ਮੈਬਰ ਲੋਕ ਸਭਾ ਲਈ ਚੁਣੇ ਜਾ ਸਕਦੇ ਹਨ ਇਸ ਤੋ ਇਲਾਵਾ ਜੇਕਰ ਰਾਸਟਰਪਤੀ ਨੂੰ ਵਿਸਵਾਸ ਹੋਵੇ ਕਿ ਐਗਲੋ-ਇੰਡੀਅਨ ਜਾਤੀ ਨੂੰ ਆਮ ਚੋਣਾ ਵਿਚ ਉੱਚਿਤ ਪ੍ਰਤੀਨਿਧਤਾ ਪ੍ਰਾਪਤ ਨਹੀ ਹੈ ਸਕੀ ਤਾ ਉਹ ਇਸ ਜਾਤੀ ਦੇ ਦੋ ਮੈਬਰ ਲੋਕ ਸਭਾ ਵਿਚ ਨਾਮਜਦ ਕਰ ਸਕਦਾ ਹੈ ਐਗਲੋ-ਇੰਡੀਅਨ ਜਾਤੀ ਦੇ ਇਹ ਦੋ ਮੈਬਰ ਲੋਕ ਸਭਾ ਦੇ ਕੁਲ 550 ਮੈਬਰਾ ਤੋ ਇਲਾਵਾ ਹੋਣਗੇ ਇਹਨਾ ਦੋ ਮੈਬਰਾ ਨੂੰ ਸਾਮਲ ਕਰਕੇ ਲੋਕ ਸਭਾ ਦੇ ਮੈਬਰਾ ਦੀ ਕੁਲ ਗਿਣਤੀ 552 ਹੋਵੇਗੀ ਲੋਕ ਸਭਾ ਦੀ ਕੁੱਲ ਸੰਖਿਆ ਵਿਚੋ 79 ਸਥਾਨ ਅਨੁਸੂਚਿਤ ਜਾਤੀਆ **Scheduled Castes** ਲਈ ਅਤੇ 40 ਸਥਾਨ ਅਨੁਸੂਚਿਤ ਕਬੀਲਿਆ **Scheduled Tribes** ਲਈ ਰਾਖਵੇ ਰੱਖੇ ਗਏ ਹਨ ਇਹਨਾ ਸਥਾਨਾ ਤੇ ਇਹਨਾ ਜਾਤੀਆ ਜਾ ਕਬੀਲਿਆ ਦੇ ਲੋਕ ਹੀ ਚੋਣ ਲੜ ਸਕਦੇ ਹਨ

**Figure 1. Sample of Punjabi Text Written in Gurmukhi Script Interpersed with English Words and Numerals**

In this paper, a modified Gabor Feature Extraction technique has been proposed to determine scripts from document images. The proposed feature extraction technique converts the original word image into its different zones at first, seacond and third level. Then gabor features are extracted from each of the subimages obtained from different zones. The underlying script of word is then determined by SVM classifier using these extracted features. Since the experiments are for discrimination of Gurmukhi, English words and English numerals, an overview about Gurmukhi, English Script and English numerals and its character set as shown in Figure 1.1 has been explained here.

### 1.1. Gurumukhi Script

Gurmukhi script is used primarily for the Punjabi language, which is the world's 14th most widely spoken language. The populace speaking Punjabi is not only confined to North Indian states such as Punjab, Haryana and Delhi but is spread over all parts of the world.There is rich literature in this language in the form of scripture, books, poetry. Gurmukhi script alphabet consists of 41 consonants and 12 vowels and 2 half characters which lie at the feet of consonants in Figure 2.a.

**Figure 2.a. Gurmukhi Script**

| a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |

**Figure 2.b. English Script**

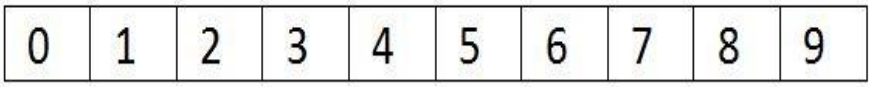



**Figure 2.c. English Numerals**

**Figure 2. Gurmukhi Script, English Script and English Numerals Character Set**

The characters of words are connected mostly by this line called head line and so there is no vertical intercharacter gap in the letters of a word and formation of merged characters is a norm rather than an aberration in Gurmukhi script The words are, however, separated with blank spaces.A word in Gurmukhi script can be partitioned into three horizontal zones. The upper zone denotes the region above the head line, where vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. The middle zone is the busiest zone. The lower zone represents the area below middle zone where some vowels and certain half characters lie in the foot of consonants [5].

### 1.2. English Script

There are twenty six for each of upper and lower case letters as in Figure 2.b. While the capital letters of English script cover the middle and upper zones, most of lower case letters that have a spatial spread that covers only the middle zone or the middle and the lower ones [1].

### 1.3. English Numerals

English numerals as shown in Figure 2.c have same features as that of English upper case letters, there is no concept of lower case letters. Also number of horizontal and vertical strokes is less as compared to English characters.

The remainder of this paper is as follows: Related script and language identification work has been reviewed in Section 2. The feature extraction technique using zone based gabor filters has been discussed in Section 3. The SVM classifier used for script identification is explained in Section 4. Experimental results have been presented in Section 5. Conclusion is given in Section 6.

## 2. Related Works

Spitz described a technique by examining the upward concavities of connected components to determine script of Asian and European languages [6].Tan proposed a method to identify the script of seven languages: Chinese,English, Koreans, Greek, Malayalam, Persian and Russian, based on taxture anaylsis using multi channel Gabor filters and co-occurrence matrices [7].Wood *et al.*, have used horizontal and vertical projection profiles of document images to identify the script of the document. They argued that horizontal and vertical projection profiles of Roman, Cyrillic, Arabic and Korean scripts have dominant peaks at different positions [8]. Hochberg *et al.*, presented a system that automatically identifies the script form using cluster based templates [9].

Script identification at paragraph/ block level for Indian languages has been reported in [10, 11]. Line Level script identification techniques for Indian languages have been discussed in [12-14].Most of the Indian documents have script changing with words. From the literature survey, it has been revealed that a major amount of work has been carried out for script identification at word level. Based on the presence and absence of shiorekha, a bilingual OCR for printed documents has been proposed by Jawaher *et al.*, for identification of Devanagari and Telugu scripts [15]. The script line identification techniques in [13, 14] have been modified in [16, 17] for script word separation in printed Indian multiscript documents by including some new features such as headline feature, distribution of vertical strokes, left and right profiles, deviation feature, loop feature, tick feature *etc*. Based on these structural features, systems for recognizing English, Devanagari and Urdu [18] and English and Tamil [19] have been reported. English, Hindi and Kannada words have also been discriminated by using structural features by Padma et al [20]. R.Dhir *et al.*, [21] have presented an automatic script identification system for Roman and Gurmukhi script words and characters.Peeta Basa Pati *et al.*, [22-24] have reported a system to identify the script of each word in a document image. They have started with a bi-script which is later extended to tri-script and then to eleven-script scenarios. Effectiveness of Gabor and Discrete Cosine Transformations (DCT) features has been evaluated using KNN, Linear Discriminate and SVM classifiers. In their work, the consideration of English numerals has not been mentioned. B.V. Dhandra *et al.*, [25] have proposed word level English numeral identification from Indian documents having scripts Kannada, Devanagari, Tamil, Oriya and Malayalam using morphological reconstruction.However the work reported is only on identification of English Numerals not for English words. And also the work reported is only for south Indian scripts.So, from this a motivation is there to extract English words and numerals from Gurmukhi script. To extract English numerals from Gurmukhi text, some results have been reported in our previous work [26].This work is an extension of that work, by modifying the technique of feature extraction and by identifying the English words and numerals from Gurmukhi words.

## 3. Feature Extraction

Features of an image/patten are the symbolic properties which are used to differentiate it from other image/pattern. Here, an attempt has been done to find features that maximize the differentiation between Gurmukhi words, English words and English numerals. The features are based on the energy distribution in the given word image and its different zones in various directions.After careful observation of Gurmukhi Script words, English Script words and numerals ,as discussed in Section 1, it reveals that these words have different behaviour in different zones.This motivates us to use the energy features of a word image in different zones and directions for identification of its script. Gabor filters, which are capable of providing multi-resolution analysis, have been used to find directional energy features [27].

### 3.1. 2 Dimensional Gabor Function

A Gabor function G(x,y) is a linear function as given in Equation 1, defined by multiplication of harmonic function with a Gaussian function [26]

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

where s(x,y)is a complex sinusoid harmonic function, known as carrier and g(x,y) is a Gaussian shaped function, known as envelope. Thus the 2D Gabor filter with orientation $\theta$ and centerd at frequency $f$ can be written as in Equation 2

$$G_{x,y,\theta,f,\sigma_x,\sigma_y} = \exp^{-\frac{1}{2}(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2})}(\cos 2\pi f x' + j\sin 2\pi f x') \quad (2)$$

where $\sigma_x$ and $\sigma_y actas$ the spatial spread and are the standard deviations of the Gaussian envelope along x and y direction and $x'$ and $y'$ are defined as:

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

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

$$G_{Even}(x,y,\theta,f,\sigma_x,\sigma_y) = \exp^{-\frac{1}{2}(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2})} \times \cos 2\pi f x' \quad (3)$$

$$G_{Odd}(x,y,\theta,f,\sigma_x,\sigma_y) = \exp^{-\frac{1}{2}(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2})} \times \sin 2\pi f x' \quad (4)$$

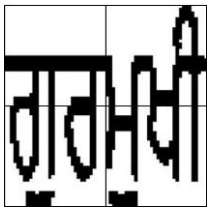
For a given image I(x,y), the convolution of Gabor function at frequency f and angle $\theta$ with the given image yields Gabor filtered output image O(x,y) as given in Equation 5

$$O(x,y) = I(x,y) * G(\theta,f) \quad (5)$$

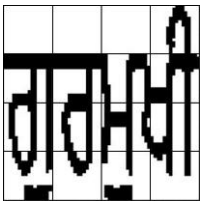
### 3.2. Zone-Based Gabor Feature Extraction

Local matching is a common approach in pattern recognition whose general logic is to first locate several local components having distinct features and then classify these images by comparing and combining the corresponding local stattistics. Hence for script identification of a word, as there is no other component for obtaining local information,it is divided into different zones of same size at two levels as shown in Figure 3. Also, as the words can have any number of characters, so the size of the word image varies.To divide the image into different zones of same size, normalization of the given word image is required, so normalization has been done for the given word image into size 32 × 32. The Zone based Gabor feature extraction as shown in Figure 4 is achieved by the following algorithm.

Normalized Image of size 32 × 32
Figure 3.a



First level Zoning
Figure 3.b



Second Level Zoning
Figure 3.c

**Figure 3. Example of Two Level Zoning of a Normalized Word Image**

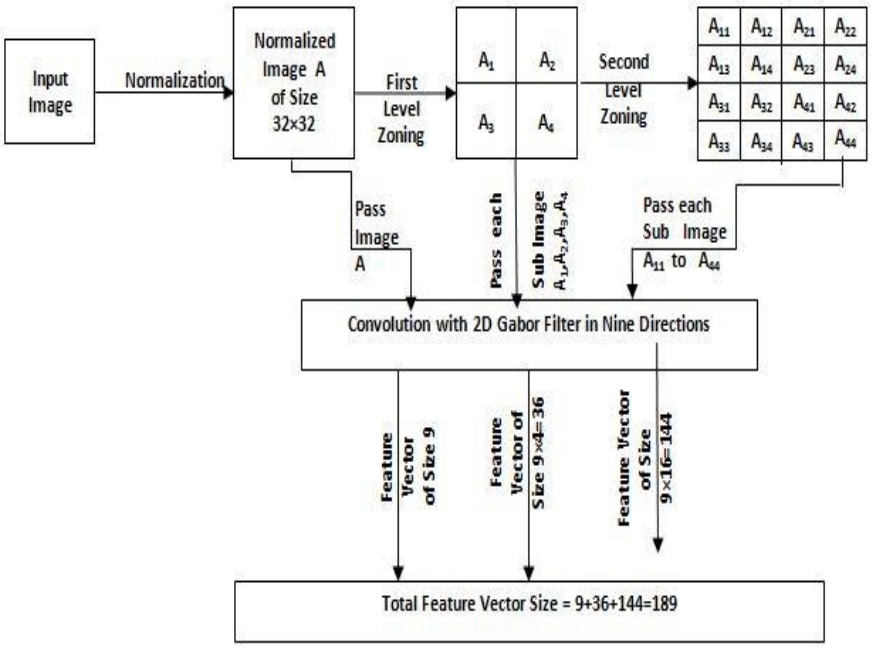



**Figure 4. Zone based Gabor Feature Extraction**

*A*lgorithm: Gabor Feature Extraction

1. Normailze the given word Image into size $32 \times 32$ and call it A.
2. Apply first level zoning by partitioning the normalized image into four equal non overlapping subregions of size $16 \times 16$ and call these as $A_1, A_2, A_3$ and $A_4$.
3. Apply second level zoning by further partitioning each subregion into four equal non overlapping sub-subregions of size $8 \times 8$ and call these as $A_{11} \ldots A_{14}$, $A_{21} \ldots A_{24}$, $A_{31} \ldots A_{34}$ and $A_{41} \ldots A_{44}$ and thus obtain 16 small regions in different parts of the image.
4. Repeat step 5 and 6 for each of the twenty one images (one obtained in step 1, four in step 2 and sixteen in step 3) $I_Z(x, y) = A, A_1, A_2, A_3, A_4, A_{11} \ldots\ldots A_{44}$.
5. Convolve $I_Z$ with odd symmetric and even symmetric gabor filters given in Equations 3 and 4 in nine different angles of orientation (0, $\pi/9$,$2\pi/9$, $3\pi/9$, $4\pi/9$, $5\pi/9$, $6\pi/9$, $7\pi/9$, $8\pi/9$) with frequency 2/n to genearte output gabor filtered images $O_{even,\theta}(I_Z)$ and $O_{oddz,\theta}(I_Z)$, where $n \times n$ is the size of $I_Z$ .
6. Evaluate the energy content of the image in each direction $\theta$ and normalize this energy by dividing it by the size of each corresponding image region as given in equation .

$$Energy_\theta(I_Z) = \frac{\sqrt{(O_{even,\theta}(I_Z))^2 + (O_{odd,\theta}(I_Z))^2}}{n \times n} \quad (6)$$

Thus each zone region $I_Z$ gives nine features.The feature vector $Fv$ of zone region $I_Z$ is represented as:

$$Fv(I_Z) = Energy_{0\pi/9}(I_Z), Energy_{1\pi/9}(I_Z), \ldots\ldots\ldots Energy_{8\pi/9}(I_Z) \quad (7)$$

So, a feature vector of size 189 ($1 \times 9 + 4 \times 9 + 16 \times 9$) is obtained.

## 4. Classification

The main task of classification is to use the feature vectors provided by feature extraction algorithm to assign the object/pattern to a category [28]. Support Vector Machine(SVM) is a classification technique successfully used in a wide range of applications.

### 4.1. SVM Classifier

Binary (two-class) classification using support vector machines (SVMs) is a very well developed technique to find the optimal hyperplane to maximize the distance or margin between two classes .

Given a training set of instance-label pairs $(x_i, y_i), i = 1,2,\ldots. l$ where $x_i \varepsilon R^n$, i.e. having n features for a particular training sample and $y_i \in \pm 1$, i.e. class label either 1 or -1 for corresponding training instance $x_i$. If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin".The distance between these two hyperplanes is $\frac{2}{\|w\|}$, so $\| w \|$ should be minimium [29].

If there exists no hyperplane that can split the 'yes' and 'no' examples, the Soft Margin method will choose a hyperplane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. The objective function is then increased by a function which penalizes non-zero $\xi_i$ and the optimization becomes a trade off between a large margin and a small error penalty.The support vector machines (SVM) require the solution of the following optimization problem, *i.e.*, minimization of error function[30] as given in Eq. 8 :

$$min_{w,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i \tag{8}$$

subject to the constraints:

$$y_i(W^T \phi(x_i) + b) \geq 1 - \xi_i$$

and

$$\xi_i \geq 0$$

where C is the penality parameter, W is the vector of coefficients, b a constant and $\xi_i$ are parameters for handling non-separable data (inputs). The index i labels the N training cases or instances. Here $y_i \in \pm 1$ are the class labels and $x_i$ are the independent variables. The kernel $\phi$ is used to transform data from the input (independent) to the feature space.

In testing phase, for a given input pattern x, the decision function of an SVM binary classifier is

$$f(x) = sign(\sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b) \tag{9}$$

where:

$$sign(u) = \begin{cases} 1 for u > 0 \\ -1 for u < 0 \end{cases}$$

$b$ is the bias, $\alpha_i$ is the langrage multiplier and $K(x, x_i)$ is the kernel function [32].

There are several number of kernels used in support vector machines. Some of the popularily used kernel functions are:

• Linear Kernel:

$$K(x, x_i) = x^T x_i \tag{10}$$

• Polynomial Kernel:

$$K(x, x_i) = (x^T x_i + 1)^d \tag{11}$$

where d is the degree of polynomial.

• Gaussian (RBF)Kernel:

$$K(x, x_i) = \exp(-\gamma * \| x - x_i \|^2) \tag{12}$$

where $\gamma = (1/2\sigma^2)$ and $\sigma$ is the standard deviation of the $x_i$ values.

The solution of multi (More than two classes) classification is by combining several binary classifiers.There are two approaches for combining binary SVM classifier: 'One Vesus All' (OVA) and 'One Versus One' (OVO).

For our case, as there is three class problem, so number of binary SVM classifiers for both approaches are three. As one vs one approach is having less number of training examples for each binary classifier, so OVO approach has been used for the present work.

## 5. Experimental Results and Discussion

In this section, a description has been given about the experimental setup and results obtained to examine the performance of extracted Zone-based Gabor features with SVM classifier using different kernel functions. The comarison of the results with Gabor Features by pati *et al.,* [24] and our previous work [26] has been done on dataset prepared for the present work.

### 5.1. Dataset Preparation

Since a standard database for Indian script recognition is not available, a dataset of 11400 words has been prepared. To prepare the database, Gurmukhi documents from different sources like books, magazines, newspapers have been taken and scanned. After binarization, noise cleaning line segmentation and word segmentation has been done using horizontal and vertical projection profiles respectively. The collected data set contains 5212 Gurmukhi Words, 4288 English Words and 1900 English Numerals.

### 5.2. Experiment Results with Different Kernel Functions

To obtain the recognition results, 10-fold cross validation has been used. First, creation of randomly generated 10-fold cross-validation index of the length of size of dataset has been done. This index contains equal proportions of the integers 1 through 10. These integers are used to define a partition of whole dataset into 10 disjoint subsets. We used one division for testing and remaining nine divisions for training. This has been done 10 times, each time changing the testing dataset to different division and considering remaining divisions for training. Thus 10 sets of feature vectors containing training and testing dataset in the size ratio of 9:1 has been got.

Our experiments are carried out using different kernal functions of SVM classifier with 'OVO' approach.The main cause of performance difference among different types of SVM classifiers is linked to feature data distribution[34]. Linear SVM should be largely enough to achive high accuracy when data is linearly seperable. However, such situations are very rare and therefore it is necessary to find hyper-surfaces separating two classes. Therefore we have tested our results using Linear, Polynomial and Gaussian (RBF) kernel. Detailed results of different SVMs are given in Tables 1, 2 and 3. From linear SVM we obtained 92.87% accuracy, from Polynomial kernel we obtained average accuracy 93.28% and from RBF kernel we got average accuracy 99.39% from our script identification scheme.

**Table 1. Results for Script Identification using Linear Kernel Function**

| | | Recognized Script | | |
|---|---|---|---|---|
| | | Gurumukhi Words | English Words | English Numerals |
| Actual Script | Gurumukhi Words | 93.28% | 6.04% | 0.68% |
| | English Words | 5.50% | 93.05 % | 1.45% |
| | English Numerals | 1.47% | 7.21% | 91.32% |
| Average Accuracy | | 92.87% $\pm$ 0.2551% | | |

**Table 2. Results for Script Identification using Polynomial Kernel Function**

| | | Recognized Script | | |
|---|---|---|---|---|
| | | Gurumukhi Words | English Words | English Numerals |
| Actual Script | Gurumukhi Words | 93.80% | 5.43% | 0.77% |
| | English Words | 5.34% | 93.38 % | 1.28% |
| | English Numerals | 1.84% | 6.53% | 91.63% |
| Average Accuracy | | 93.28% $\pm$ 0.2490% | | |

**Table 3. Results for Script Identification using Gaussian(RBF) Kernel Function**

| | | Recognized Script | | |
|---|---|---|---|---|
| | | Gurumukhi Words | English Words | English Numerals |
| Actual Script | Gurumukhi Words | 99.33% | 0.65% | 0.02% |
| | English Words | 0.37% | 99.46 % | 0.17% |
| | English Numerals | 0.05% | 0.58% | 99.37 % |
| Average Accuracy | | 99.39%$\pm$ 0.2026% | | |

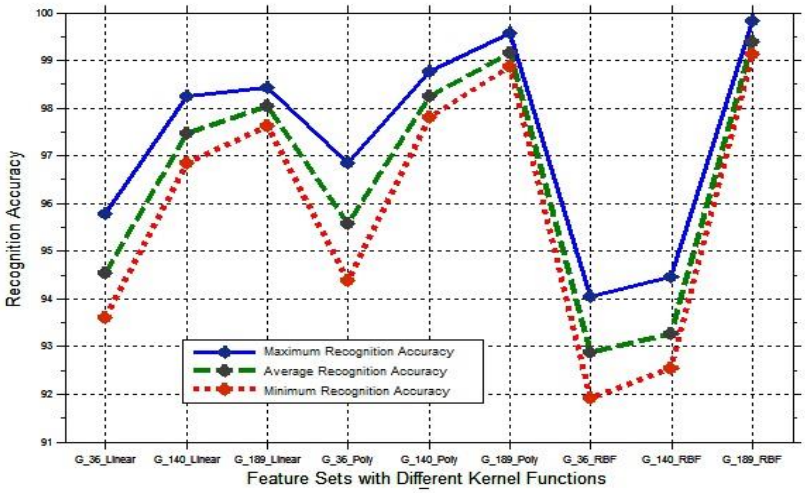



Figure 5.a. Comparison of Recognition Accuracy

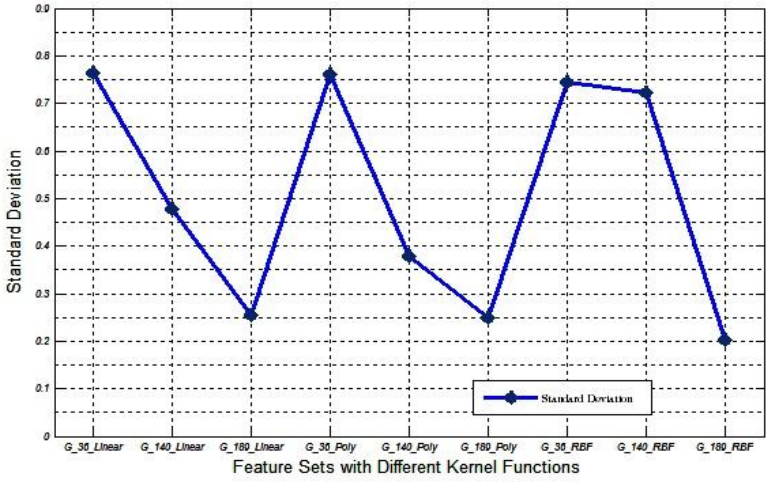



Figure 5.b. Comparison of Standard Deviation

**Figure 5. Comparison of Proposed Method with Existing Methods**

In order to further analysis of classification error depending on the kind of SVM chosen, we have computed standard deviation of the recognition accuracies obtained from ten subsets.The standard deviations of recognition accuracies obtained from Linear, Polynomial and Gaussian(RBF) SVM are 0.2551,0.2490 and 0.2026 respectively.From Tables 1, 2 and 3, it is clear that Gaussian (RBF) SVM gives highest accuracy with lowest standard deviation.This means that Gaussian SVM is best in separating the hypersurfaces defined by its support vectors to identify script of Gurmukhi words, English words and English Numerals than the other separating hyper-surfaces.

### 5.3. Comparison with Existing Methods

There are many papers that discuss script recognition problem but a few of them are for Gurmukhi and Roman scripts.Our previous method [26] identifies English Numerals from Gurmukhi text using 140 Gabor features. Pati *et al.*, [24] used a set of 36 features for

Gurmukhi and Roman script.To the best of our knowledge, there is no reportd work that simultaneously identifies English Words and numerals from Gurmukhi script. So, we have tested these two methods on our present dataset and compared these results with our proposed method.

Figure 5.a shows the minimum, average and maximum accuracy obtained with all types of feature sets with different kernel functions. Figure 5.b gives the standard deviation for various feature-classifier combinations.Here code G_36_Linear means 36 Gabor features extracted by method [24] with Linear kernel function of SVM Classifier.It may be observerd from Figure 5. athat the G_189_RBF is leading in all for minimum,average and maximum accuracy.Again from Figure 5.b it is clear that G_189_RBF has the lowest standard deviation of recognition accuracies obtained from ten subsets.From these two figures it is clear that our proposed Gabor Feature Extraction has more accuracy with low standard deviation for all types of kernel functions of SVM than the other similar methods with corresponding kernel function.But overall, the proposed technique with RBF kernel function is most efficient for script identification of Gurmukhi Words, English Words and English Numerals than other similar works by Gabor Features and SVM kernel functions.

## 6. Conclusion

In this paper, a Zone based approach based on Gabor filters for word level script identification has been proposed. First the given word image is divided into Zones at two leveles after normalization. Then Gabor features are extracted from the normalized image and from different zones of the image at each level. Then script of Gurmukhi Words, English words and Numerals has been identified by using these features and SVM classifier with different kernel functions. The testing of this scheme has been done on 11,400 words and obtained 99.39% accuracy with RBF kernel function of SVM classifier. To the best of our knowledge, this is the first work which identifies English words and Numerals from Gurmukhi script. In future more rigorous investigations are needed to study the effectiveness of the proposed scheme for identification of script of other Indian and non-Indian scripts.

## References

[1] S. Abirami and S. Murugappan, “Scripts and Numerals Identification from Printed Multilingual Document Images”, Proceedings of International Conference on CS & IT, **(2011)**, pp. 129.
[2] http://en.wikipedia.org/wiki/Optical _ character _ recognition.
[3] D. Ghosh, T. Dube and A. P. Shivaprasad, “Script Recognition A Review”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 12, **(2010)** December, pp. 2142-2161.
[4] R. Rani and R. Dhir, “A Survey: Recognition of Scripts in Bi-Lingual/Multi-Lingual Indian Documents”, National Journal of PIMT Journal of Research, vol. 2, no. 1, **(2009)** March-August, pp. 55-60.
[5] G. S. Lehal and C. Singh, “Feature Extraction and Classification for OCR of Gurmukhi Script”, Vivek vol. 12. no. 2, **(1999)**, pp. 2-12.
[6] A. L. Spitz, “Determination of Script and Language Content of Document Images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 3, **(1997)** March, pp. 233-245.
[7] T. N. Tan, “Rotation Invariant Texture Features and Their Use in Automatic Script Identification”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 7, **(1998)** July, pp. 751-756.
[8] S. L. Wood, X. Yao, K. Krishnamurthi and L. Dang, “Language Identification for Printed Text Independent of Segmentation”, Proceedings of International Conference on Image Processing, vol. 3, **(1995)** October, pp. 428-431.
[9] J. Hochberg, P. Kelly, T. Thomas and L. Kerns, “Automatic Script Identification from Document Images Using Cluster Based Templates”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, **(1997)** February, pp. 176-181.
[10] G. D. Joshi, S. Garg and J. Sivaswamy, “Script Identification from Indian Documents”, Proceedings of. IAPR International Workshop Document analysis Systems, **(2006)** February, pp. 255-267.

[11] S. Chaudhury and R. Sheth, "Trainable Script Identification Strategies for Indian Languages", Proceedings of International Conference on Document Analysis and Recognition, **(1999)** September, pp. 657-660.
[12] U. Pal and B. B. Chaudhari, "Script Line Separation from Indian Multi-Script Documents", Proceedings of International Conference on Document Analysis and Recognition, **(1999)** September, pp. 406-409.
[13] U. Pal and B. B. Chaudhuri, "Identification of Different Script Lines from Multi-Script Documents", Image and Vision Computing, vol. 20, no. 13, **(2002)** December, pp. 945-954.
[14] U. Pal, S. Sinha and B. B. Chaudhri, "Multi-Script Line Identification from Indian Documents", Proceedings of International Conference on Document Analysis and Recognition, **(2003)** August, pp. 880-884.
[15] C. V. Jawahar, M. N. S. S. K. Pavan Kumar and S. S. Ravi Kiran, "A Bilingual OCR for Hindi-Telugu Documents and its Applications", Proceedings of International Conference on Document Analysis and Recognition, **(2003)** August, pp. 408-412.
[16] S. Sinha, U. Pal and B. B. Chaudhri, "Word-Wise Script Identification from Indian Documents", Proceedings of IAPR International Workshop on Document Analysis Systems, **(2004)** September, pp. 310-321.
[17] S. Chanda, S. Sinha and U. Pal, "Word-Wise English Devanagari and Oriya Script Identification", Speech and Language Systems for Human Communication, **(2004)**, pp. 244-248.
[18] S. Chanda, R. K. Roy and U. Pal, "English, Devanagari and Urdu Text Identification", Proceedings of International Conference on Cognition and Recognition, **(2005)** December, pp. 538-545.
[19] S. Chanda, R. K. Roy and U. Pal, "English and Tamil Text Identification", Proceedings of National Conference on Recent Trends In Information Systems, **(2006)** July, pp. 184-187.
[20] M. C. Padma and P. Nagabhushan, "Identification and Separation of Text Words of Kannada, Hindi and English Languages through Discriminating Features", Proceedings of National Conference on Document Analysis and Recognition, **(2003)** July, pp. 252-260.
[21] R. Dhir, C. Singh and G. S. Lehal, "A Structural Feature Based Approach for Script identification of Gurumukhi and Roman Characters and Words", Proceedings of 39th Annual National Convention of Computer Society of India, **(2004)** December.
[22] P. Basa Pati, S. Sabari Raju, N. Pati and A. G. Ramakrishnan, "Gabor filters for document analysis in Indian Bilingual Documents", Proceedings of International Conf. on ISIP, **(2004)**, pp. 123-126.
[23] P. Basa Pati and A. G. Ramakrishnan, "HVS Inspired System for Script Identification in Indian Multi- Script Documents", Proceedings of International Workshop on Document Analysis System, **(2006)** February, pp. 380-389.
[24] P. Basa Pati and A. G. Ramakrishnan, "Word level multi-script Identification", Pattern Recognition Letters, **(2008)**, pp. 1218-1229.
[25] B. V. Dhandra and M. Hangarge, "On Separation of English Numerals from Multilingual Document Images", Journal of multimedia, vol. 2, no. 6, **(2007)** November, pp. 26-33.
[26] R. Rani, R. Dhir and G. S. Lehal, "Comparative Analysis of Gabor and Discriminating Feature Extraction Techniques for Script Identification", Proceedings of ICISIL, Communications in Computer and Information Science 139, Springer, **(2011)** March.
[27] D. Dhanya, A. G. Ramakrishnan and P. Basa Pati, "Script identification in printed bilingual documents", Sadhana, vol. 27, no. 1, **(2002)** February, pp. 73-82.
[28] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification", 2nd ed. Wiley, **(2000)**.
[29] Support Vector Machines (SVM), StatSoft website [Online]. Available at: http://www.statsoft.com/textbook/support-vector-machines/.
[30] Support Vector Machine, Wikipedia website [Online]. Available at: http://en.wikipedia.org/wiki/Support-vector-machine.
[31] C.-W. Hsu, C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification", [Online] Available: http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf, **(2010)** April.
[32] N. Das and J. Mohan Reddy, "A statistical-topological feature combination for recognition of handwritten numerals", Applied Soft Computing Elsevier, vol. 12, **(2012)**, pp. 2486-2495.
[33] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines", IEEE transactions on Neural Networks, vol. 13, no. 2, **(2002)** March, pp. 415-425.
[34] S. Chanda and U. Pal, "Word-Wise Thai and Roman Script Identification", ACM transactions on Asian Language Information Processing, vol. 8, no. 3, **(2009)** August, pp. 11.1-11.21.

## Authors

**Rajneesh Rani** she has received the B.Tech and M.Tech degrees, both in Computer Science and Engineering, from Punjab Technical University, Jalandhar, India in 2001 and Punjabi University Patiala, India in 2003 respectively. From 2003 to 2005, she was a lecturer in Guru Nanak Dev Engineering College, Ludhiana. She is currently working as an assistant professor in NIT Jalandhar since 2007. Her teaching and research include Image Processing, Pattern Recognition, Machine Learning, Computer Programming and Document Analysis and Recognition.

**Dr. Renu Dhir** received the B.Tech degree in Electrical India in 1983, M.Tech in Computer Science and Engineering from TIET Patiala, India in 1997 and PhD in Computer Science and Engineering from Punjabi University Patiala, India in 2007.She is currently associate professor in the Department of Computer Science and Engineering, NIT Jalandhar. Her teaching and research include Image Processing, Pattern Recognition, Machine Learning and Network Security.

**Dr. Gurpreet Singh Lehal** received B.Sc and M.Sc (Maths Hons.) from Punjab University, Chandigarh, India. He has done M.E (Computer Science) from TIET Patiala, India and PhD from Punjabi University Patiala. He is currently working as a Professor in the Department of Computer Science and Engineering, Punjabi University, Patiala. He is also acting as a Director of Advanced Center for Technical Development of Punjabi Language Literature and Culture. For many years, he has done research in image processing, pattern recognition and Natural Language Processing and has developed many technologies and language software. He is a member of IEEE.