International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
International Conference on Communication Computing & Virtualization 2016

A Comparative Analysis of Feature Extraction Techniques and Classifiers Inaccuracies for Bilingual Printed Documents (Gujarati-English)

Shailesh A. Chaudhari
M.Sc.(I.T.) Programme,
Veer Narmad South Gujarat University,
U.M. Road, Surat, India.

Ravi M. Gulati, PhD
Dept. of Computer Science
Veer Narmad South Gujarat University,
U.M. Road, Surat, India.

ABSTRACT

In a bilingual or multi-lingual optical character recognition system script identification is a challenging task. A remarkable research work on script identification have been noted in Indian or non-Indian context. As many commercial and official regional documents of different states of India are in bilingual containing one regional language of respective state and the other international intersperse language English. Therefore script identification is one of the primary tasks in multi-script document recognition. English words are mostly interspersed in regional documents of different states of India. In this paper script identification of Gujarati and English at word level is presented. For feature extraction two approach are used. In the first approach statistical features and in second approach the Gabor features of a word using Gabor filters with suitable frequencies and orientations are extracted. The proposed system uses two classifiers k-NN and SVM with different kernel functions used to classify the extracted features in one of the script. From the experiment it has been perceived that SVM outperform then k-NN.

General Terms

Script Recognition, Calssifier, Algorithms et. al.

Keywords

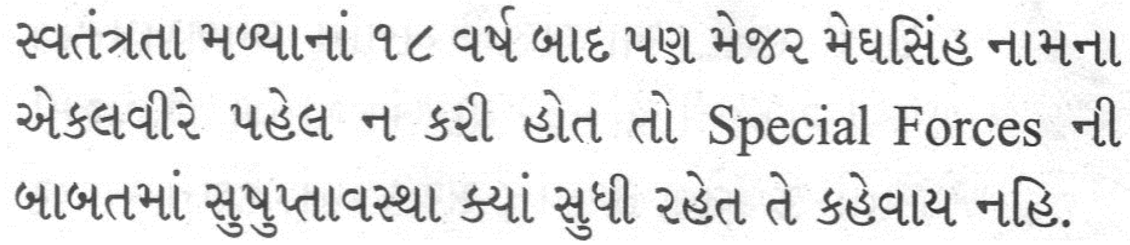
Gabor Filter, Support Vector Machine, Feature Extraction.

1. INTRODUCTION

Optical character recognition (OCR) is one of the important tasks of document image analysis. The OCR is simple when the document to be recognized is monolingual, but it becomes difficult when the document is bilingual or multi-lingual. In India many regional documents, magazines, books and reports are bilingual in nature containing one regional language and other international language English. For example an official document of Gujarat state contains Gujarati and English words. The processing of such type of document is a challenging task for OCR researchers. A script identifier simplifies the task of OCR by improving the accuracy of recognition.

There are 18 official Indian languages which are written using 12 different scripts and English has preferred to be the binding language in India. Indian script recognition has been in emerging stage. Almost all state official documents in Indian, normally contain English words mixed with other words in its own language. As shown in figure1, a Gujarati document of Gujarat state, contains intermixed English words. Therefore, these kinds of documents, word level script identification is necessary for a Character Recognition System.

In this paper, a script identification technique at word level for printed bilingual (Gujarat-English) document is proposed. Here, two types of features statistical and Gabor are used. A comparative study of classifier accuracy is discussed using k-NN and SVM with different kernel functions. Experiments are conducted to check the font type and font size dependency in the training and testing dataset.



સ્વતંત્રતા મળ્યાનાં ૧૮ વર્ષ બાદ પણ મેજર મેઘસિંહ નામના
એકલવીરે પહેલ ન કરી હોત તો Special Forces ની
બાબતમાં સુષુપ્તાવસ્થા ક્યાં સુધી રહેત તે કહેવાય નહિ.

Fig.1. Sample bi-script document image showing intermixed English and Gujarati words

The organization of the paper is as follows: Section 2 begins with description of related work. Section 3 describes Gabor filter design and Feature extraction algorithms. The SVM Classifier is discussed in detail in section 4. Experimental analysis and results are discussed in section 5 followed by conclusion and future enhancements in section 6.

.

2. RELATED WORK

The interspersion of different scripts in a single document may be at paragraph, line, word or character level is describe in [1,2]. The use of script identification system depends on the minimum size of the text from which features are extracted reliably. There are main two approaches used to identify script from multi-lingual document: Local approach and Global approach. The local approach analyses a document image based on connected components and such components require segmentation of the image as a pre-processing step, while global approach employs analysis of regions containing at least two lines and do not require type of fine segmentation.

In the scenario of Indian language document analysis, major works is done by Chaudhari and Pal [3, 4, 5, 12]. They used headline as a feature to distinguish Roman lines from the Bangla and Devnagari lines. They also used various structural features, horizontal projection profiles, Water reservoirs, Contour tracing (left and right profiles) to differentiate other Indian languages.

A texture-based approach was presented by Padma and Vijaya [6]. They extract the Haralick texture features from the co-occurrence matrix by the Wavelet Packet Decomposition and then used these features for the script identification in a machine printed document. In wavelet packets, each detail coefficients vector is also decomposed as in approximation vectors.

A two stage block level script identification scheme was reported by Patil and Subbareddy [7]. In the first stage of the document image is dilated using 3X3 matrix masks in four directions horizontal, vertical, right diagonal, and left diagonal.And in the second stage, these four modified forms of the image and the original image are used as a feature vector. An appearance based model for script identification at paragraph and word level was proposed by Vikram and Guru [9]. They used PCA (Principal Component Analysis) and FLD (Fisher's Linear Discriminator) appearance based models for feature extraction.

An initial attempt in word level script identification had been made to separate Tamil and English scripts by Dhanya et. al[10]. They used the spatial spread horizontal projection profiles of a word in upper and lower zones, the character density and the directional energy distribution using Gabor filter with suitable frequencies and orientations at a word level for feature extraction. A SVM (Support Vector Machine) based technique was proposed for word level script identification from Indian documents containing English, Devnagari and Bengali by Chnda et. al. [11]. They derived 64-dimensional chain-code features for classification. The rejected samples at first level of classification are further treated with 400-dimensional gradient feature for classification at next level.

A morphological reconstruction based approach was proposed by Dhandra et.al [8,15,16] for script identification at word level. They used morphological erosion and opening by reconstruction of words in four directions horizontal, vertical, right, and left diagonal using line structuring elements and also the hole filling is performed for those character which contains loop. A Gabor feature extraction technique was proposed by Kunte et.al [13] to recognize the script at word level using pre-trained neural classifier.

A font and size independent OCR system for recognizing Tamil and English scripts at character level was presented by Aparna et. al [14]. They employed features like geometric moments and discrete cosine transform coefficients.

3. FEATURE EXTRACTION

Feature extraction plays an important role in any character recognition system. The objective of feature extraction is to describe the pattern with least number of features that are effective in differentiating pattern classes. In this research Gabor filters are used, because they have been proven to be powerful tool for feature extraction.

Before proceeding to the recognition stage, some set of features are extracted from the image. These features are a condensed representation of the contents of the image that are most relevant to the task of recognition stage. The extracted features are further processed for the preparation of a vector of values that are then passed to the classifier.

Following types of approaches are used in feature extraction: Statistical feature and Gabor feature.

3.1 Statistical feature extraction

The words of both the scripts, English andGujarati, can be separated into 3 distinct zones. Fig. 2(a) and Fig. 2(b) shows the upper, middle and lower zone for Gujarati and English script.

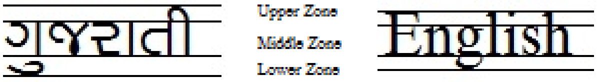


Fig.2(a) Gujarati Word **Fig. 2(b) English Word**

Here four types of statistical feature are extracted from each word image.

1) *Upper Zone Pixel Density*: it is division of the total number of on pixels in upper zone by size of the upper zone.
2) *Lower Zone Pixel Density*: it is division of the total number of on pixels in lower zone by size of the lower zone.
3) *Middle Zone Pixel Density*: it is division of the total number of on pixels in middle zone by size of the middle zone.
4) *Vertical Character Density*: it is division of the total number of characters by height of the word.

3.2 Gabor feature extraction

The Gabor filters give the optimal resolution in both the spatial and frequency domains to extract local features of an image. Gabor filters can be used as a directional feature extractor because they can effectively capture the energy coefficients in various directions, similar to Human Visual System(HVS). The Gabor filter is a band-pass linear filter whose impulse response is defined by a complex sinusoid harmonic function which is multiplied by a Guassian function as in eq..(1).

$$G_{(x,y)} = g_{(x,y)} s_{(x,y)} \quad \text{..........(1)}$$

Where s(x,y) is a complex sinusoid harmonic function, and g(x,y) is a Gaussian function known as envelope. The Complex sinusoid can be defined as in eq..(2).

$$e^{\left(j\left(2\pi\left(\mathcal{U}_0\mathcal{X}+\mathcal{V}_0\mathcal{Y}\right)+\mathcal{P}\right)\right)} \quad \text{...........(2)}$$

Where (u0, v0) is the spatial frequency and P is the phase of the sinusoid. The complex sinusoid can be thinking of as two real functions that represent real and imaginary part as in eq. (3) and eq.(4).

$$Real(S(x,y)) = \cos(2\pi(\mathcal{U}_0\mathcal{X} + \mathcal{V}_0\mathcal{Y}) + \mathcal{P} \text{.............(3)}$$

$$Img(S(x,y)) = \sin(2\pi(\mathcal{U}_0\mathcal{X} + \mathcal{V}_0\mathcal{Y}) + \mathcal{P} \text{.............(4)}$$

Thus a 2D Gabor filter with orientation θ and frequency f can be written as in eq...(5).

$$G_{x,y,\theta,f,6x,6y} = \exp^{-1/2\left(\frac{x'^2}{6x^2}+\frac{y'^2}{6y^2}\right)} X\,(cos\,2\prod \text{fx}' + \text{j}\,sin\,2\prod \text{fx}')$$

.................(5)

Where Ϭx and Ϭy are standard deviation of Gaussian envelope along x and y directions and x’ and y’ are represented as:.

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

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

In order to utilize the frequency spectrum effectively a range of both frequencies and orientations of the Gabor filters must be considered. The aim is to provide an interested even coverage of the frequency components, while maintaining a minimum of overlap between filters.To accomplish the measure of independence between the extracted features co-efficient, a proper variance values are considered.

Aset of variance radial frequencies and a sequence of orientations is considered. A multi-bank Gabor filter is created by five different values for spatial frequency (f=0.0625, 0.125, 0.25, 0.5, 1) and six different values for filter orientation (θ=0, 30, 60, 90, 120, 150, 180). Therefore a 2D Gabor filter bank of 30 filters composed of 11 channels is created.

3.2.1 Gabor Filter Design Algorithm:

```
Create the Gabor array Garray[] of size 30.
count=1;
For f=0.0625 to 1
$Garray[count] = e^{-1/2\left(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2}\right)} \times (\cos\theta f x' + j sin\theta f x')$
    For θ=0^0 to 180^0
            count=count+1
        θ=θ+30
    end
    f=f*2
end
```

3.2.2 Feature Extraction Algorithm:

1. Read the given input text word image $I_w(x,y)$.
2. Apply Gabor filter-bank on word image $I_w(x,y)$.
3. Convolve the output image with Even Symmetric Filter.
4. Compute standard deviation from the output of convolution of input word image $I_w(x,y)$.

Thus a feature vector of 30 (5-frequency X 6-orientations) is obtained which is used for classification experiment.

4. CLASSIFICATION

The core task of classification is to use the feature vectors provided by feature extraction algorithm to classify and assign the object/pattern to a category. To perceive the behavior of proposed algorithm, a comprehensive study has been made through experimental tests that are conducted on bi-script database using SVM classifier.

4.1 k-NN

k-NN is a supervised learning algorithm used for script classification. The four features are extracted from the test image *X* and these feature values are compared with feature values stored in the knowledge base. The k-NN algorithm uses euclidean distance to measure the distance between the test sample and the *k* neighbors. After careful determinationof the k nearest neighbours, a simple majority of these k-nearest neighbours is to be predicted for the query image word.

4.2 Support Vector Machine (SVM)

SVM is a binary classifierwhich classify dataset using by finding optimal hyper plane. It is an also supervised learning method that can be used for classification. It was invented by Cortes and Vapnik [17] in early 1946.

The power of SVM lies in its ability to transform data to a high dimensional space where the data can be separated using a hyper plane. SVM is a well-developed technique to create optimal hyper plane which distinct two classes by maximizing the distance or margin between two classes. Therefore the optimization process for SVM learning with different parameters of hyper plane needs to be understood.

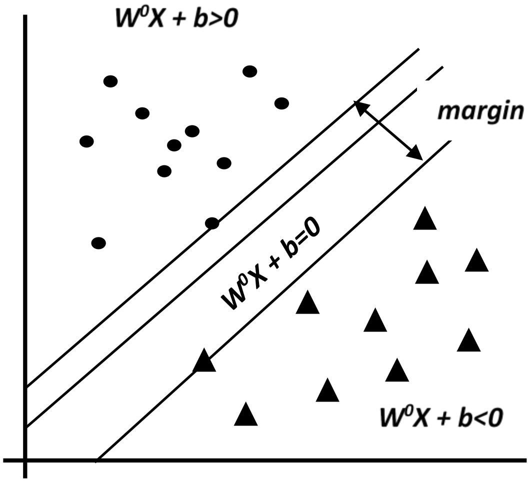


Fig. 3. Two class support vector

As shown in figure 3 the hyper plane can be represented as in eq..(6)

$$w^0.x + b = 0 \qquad \text{...........(6)}$$

Where $w^0.x + b = 0$ $w^0$is the normal to the hyper plane and b is is the bias of the hyper plane from the origin.

The points that lies on the hyper plane should satisfy the eq..(6). Let's assume that training set contains pairs of (Xi, Yi), where i=1,...n feature vectors and y=+ 1 are class labels. They must satisfy the following conditions.

$$w^0.x + b > 0 \quad if \quad Yi = 1$$

$$w^0.x + b < 0 \quad if \quad Yi = -1$$

The aim of the optimization process is to maximize the margin. That means to maximize the distance. The decision function for input pattern x with binary classifier can be represented as in eq..(7).

$$f(x) = \sin(w^0.x + b) \qquad ....(7)$$

In reality there are many such hyper plane which requires to maximize the margin. The distance between the margin $w^0.x + b = 1$ and $w^0.x + b = -1$ can be represented by $2/\| w \|$.

Therefore the optimization process to create support vector machine is defined as in eq..(8).

$$1/_2 \ \| w \|^2 - (w^0.x + b) \qquad ....(8)$$

To avoid to deal directly with the high dimensional space and excessive computations, several kernel functions are used as shown in table 1.

Table 1. SVM Kernel Functions

Linear Kernel	$K(x.y) = X^T Yi$
Polynomial Kernel	$K(x,y) = (X^T Yi + 1)^d$
RBF Kernel	$K(x.xi) = exp^{\{-\gamma \lvert x-y \rvert^2\}}$

5. EXPERIMENTS AND DISCUSSIONS

To validate effectiveness of proposed algorithm and evaluate the classifier performance, different experiments have been performed such as: a) Global script recognition accuracy based on 5 fold cross on the entire dataset. b) Script recognition accuracy of words with fonts style and size not present in training dataset.

5.1 Data Set Preparation

Due to lack of availability of standard databases authors developed their own dataset to show the efficiency of proposed features. The documents of both Gujarati and English words are printed with laser printer and scanned at a resolution of 300 dpi. Thus the dataset of 10000 words has been prepared with 5000 Gujarati words and 5000 English words.

5.2 Global Script Recognition Accuracy

Here 5-fold cross validation scheme is used for recognition result computation. First randomly generated 5-fold cross validation indices of the length of each of the dataset is created. These indices contain equal proportions of the integers 1 through 5. These integers are used to define a partition of whole datasets into 5 disjoint subsets. One subset is used for testing and remaining four subsets for training for each dataset. This is done 5 times for classifier SVM with different parameters, each time changing the testing dataset to different subset and considering remaining subsets for training. Thus 5 sets of feature vectors containing training and testing dataset in the ratio of 4:1 is obtained.

The experiments are carried out using different kernel functions of SVM classifier and kNN calssifier. The main cause of performance difference among different types of SVM classifier is linked to feature data distribution. Here testing is performed using Linear, Polynomial and RBF kernel on Gabor features. The average accuracy for Gujarati and English word recognition using kNN classifierand SVM classifier with different kernel functions is shown in Fig.4.

It can be noted that Gabor features with linear kernel function and polynomial kernel function showed the average accuracy 100.00% and 99.97%, leading as compared to RBF kernel function with average accuracy 98.38% and with kNN average accuracy 92.49%. The confusion matrix for Gujarati and English words using linear, polynomial, RBF kernel functions of SVM and kNN is shown in Table 2.

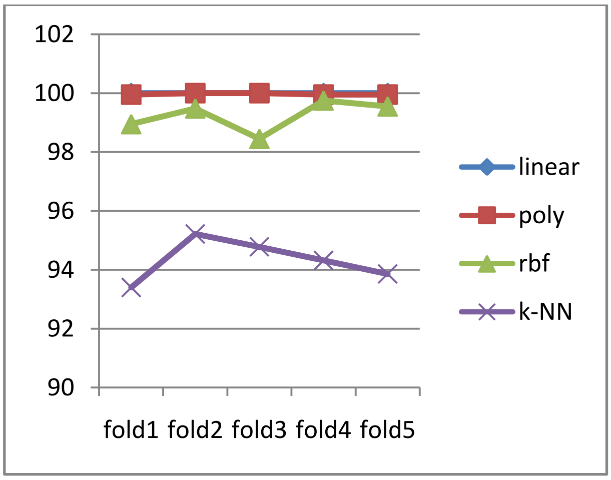


Fig. 4. Average Recognition Accuracy using different classifier

Table 2. Confusion matrix of bi-script using different classifier

Classifier	Words	Gujarati	English
SVM-RBF	Gujarati	4919	0
	English	81	5000
SVM-Linear	Gujarati	5000	0
	English	0	5000
SVM-Polynomial	Gujarati	4999	2
	English	1	4998
k-NN	Gujarati	4734	12
	English	266	4988

5.3 Accuracy with Fonts Style and Size not present in Training Dataset

In real world applications, the robustness of an algorithm with respect to distinct font and size words is a key factor. To show the efficiency of the proposed features, we performed the experiments using different fonts in training and test dataset. This is done by creating a new dataset for testing. These datasets have mutually exclusive font words. The result of new test dataset is given in Table 3. It is clear that Gabor feature with polynomial kernel function of SVM classifier gives the maximum accuracy that is 98.42 and 98.97% for Gujarati and English words respectively.

6. CONCLUSION AND FUTURE PERSPECTIVES

In this research an attempt is made on word level script identification and a comparative study of feature extraction techniques and classifier accuracies is carried out. In this work tow features: statistical feature and Gabor features are used. For classification k-NN and SVM classifier are used and compared for identification of script of printed Gujarati and English words. From results it observed that Gabor feature gave more accurate result then statistical features with both the classifiers. By using, Gabor features of words, the heights accuracy obtained is 98.70% with SVM-Polynomial kernel and for statistical features the heights accuracy obtained is 96.45% with SVM-Polynomial kernel. This is the first work on English and Gujararti script identification at word level. Here the work is reported only for Gujarati and English words, so in future it can be extended for other Indian and non-Indian words also.

Table 3. Average Accuracy of classifier with different feature set

Classifiers	Statistical Feature	Gabor Feature
SVM-RBF	94.98%	96.02%
SVM-Linear	92.2%	98.52%
SVM-Polynomial	96.45%	98.70%
k-NN	90.69%	94.34%

7. REFERENCES

[1]. Ghosh D., Dube T., Shivaprasad A. P., Script Recognition A Review. *IEEE, Transactions on Patter Analysis and Machine Intelligence* 2010. vol. 32, no. 12, pp. 2142-2161.

[2]. Chaudhari S., Gulati R., A Survey on Script Identification in Multi-script Indian Documents. *VNSGU journal of Science and Technology* 2012. Vol 3, Issue 2, pp. 138-152.

[3]. Chaudhuri.B.B, Pal.U, An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi). *In Proc. 4th ICDAR*, Uhn. 1997.

[4]. Pal U., Chaudhuri B.B., Script Line Separation from Indian Multi-Script Documents. *Proc. Int'l Conf. Document Analysis and Recognition.* 1999. pp. 406-409.

[5]. Pal U., Chaudhuri.B.B, Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line. *Proc. 6th Intl. Conf: Document Analysis and Recognition (ICDAR'OI).* 2001. pages 790-794.

[6]. Padma M.C., Vijaya P.A. Global Approach for Script Identification using Wavelet Packet Based Features. *International Journal of Signal Processing, Image Processing and Pattern Recognition*. 2010.Vol. 3, No. 3.

[7]. Patil B., Subbareddy N.V. Neural network based system for script identification in Indian documents. *Sadhana* 2002. Vol. 27, part-i1, pp 83-97.

[8]. Dhandra B.V., Nagabhushan P., Hangarge M., Hegadi R., Malemath V.S., Script Identification Based on Morphological Reconstruction in Document Images. *Proc. IEEE Int'l Conf. Pattern Recognition*. 2006. vol. 2, pp. 950-953.

[9]. Vikram T.N., Guru D.S. Appearance based models in document script identification. *ICDAR '07 Proceedings of the Ninth International Conference on Document Analysis and Recognition* - 2007. Volume 02.

[10]. Dhanya.D, Ramakrishnan.A.G, Peeta B. P. Script Identification In Printed Bilingual Documents. *Sadhana*, 2002. Vol. 27, Part-1, Pp. 73-82.

[11]. Sukalpa C., Pal S., Katrin F., Pal U. Two-stage Approach for Word-wise Script Identification. *10th International Conference on Document Analysis and Recognition*. 2009.

[12]. Pal U., Sinha S., Chaudhuri B.B. Multi-Script Line Identification from Indian Documents. *Proc. Int'l Conf. Document Analysis and Recognition.* 2003. pp. 880-884.

[13]. Kunte R.S., Sudhaker S. A Bilingual Machine-Interface OCR for Printed Kannada and English Text Employing Wavelet Features. *10th International Conference on Information Technology.* 2007.

[14]. Aparna KG, Dhanya D., Ramakrishnan AG, Bilingual (Tamil – Roman) Text Recognition on Windows, *Tamil Internet. California, USA* 2002.

[15]. Dhandra BV, Mallikarjun H., Hegadi R., Malemath VS 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 2006. pp. 389-393.

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

[17]. Cortes C, Vapnik VSupport vector network. *Machine Learning.* , 1995. 20:273–297.