Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 5, May 2014, pg.1211 – 1215

RESEARCH ARTICLE

An Analysis of Subspace Methods for Large South Indian Datasets

Krishna Murthy C.R [1], C.Naveena[2], T.C Manjunath[3]

[1]HKBK College of Engineering, Bangalore, INDIA

[2]HKBK College of Engineering, Bangalore, INDIA

[3]HKBK College of Engineering, Bangalore, INDIA

[1] krishna.murthy.crk@gmail.com; [2] naveena.cse@gmail.com; [3] principal@hkbkeducation.org

Abstract: Optical Character Recognition (OCR) is one of the important fields in image processing and pattern recognition domain. Handwritten Character Recognition has always been a challenging task. The complexity of accurate recognition of Multi Lingual South Indian Scripts makes its recognition a challenging task for the researchers. Multi Lingual characters are a challenging task because of the high degree of similarity between the characters. This paper presents an analysis of subspace methods for recognition of handwritten isolated Multi Lingual South Indian Scripts for the Kannada, Tamil, Malayalam languages. The study was carried out with a huge dataset containing 33,640 handwritten samples. The proposed method preprocesses the 841 different classes of characters obtained from scanned documents of the Multi Lingual South Indian Scripts for the Kannada, Tamil, Malayalam languages. Both Principal Component Analysis (PCA) & Fisher Linear Discriminant Analysis (FLDA) approaches are used to extract the features of characters. For classification Probabilistic Neural Network (PNN) approach is used with the combination of both PCA & FLDA feature extraction method. Based on classification of character the computed results performance of both PCA & FLDA based PNN classification was analyzed & discussed here.

Keywords: Principal Component Analysis, Fisher Linear Discriminant Analysis, Probabilistic Neural Network, Handwritten Character Recognition, Normalization

I. Introduction

One of the most interesting OCR researches is recognizing handwritten characters. This particular problem is more challenging than the counterparts of typewritten and printed texts. OCR is an area of pattern recognition and processing of handwritten character is motivated largely by desire to improve man and machine communication. An OCR system may be designed to work for either of on-line and off-line purposes. On-line OCR systems collect

input data by recording the order of strokes made by the write on an electronic bit-pad, and off-line OCR systems do the same by recording the pixel by pixel digital image of the entire writing with a digital scanner[1]

India is a multi-lingual country containing nearly 22 official languages. namely, Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri (Meithei), Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu. Offline HCR systems are matured only in few languages like English, Japanese, Arabic and Chinese[2]. Even though, sufficient studies have performed in foreign scripts like Chinese, Japanese and Arabic characters, only a very few work can be traced for handwritten character recognition of Indian scripts especially for the South Indian scripts[3]. However, there are not many reported efforts at developing OCR systems for South Indian Languages.

As per the literature survey many works have been observed on recognition of handwritten characters. C.Naveena et al [4] proposed a system for offline recognition of Bi-lingual Kannada and English isolated characters. The experiments were performed using the database containing 22,600 samples of Kannada and English and compared the affect of four different similarity measure techniques namely Euclidean distance, Modified squared Euclidean distance, Correlation distance and Angle distance for an unconstrained handwritten character recognition. The strength of these similarity measures are estimated between feature vectors with respect to the recognition performance of the Gabor-PCA method. Rajashekararadhya et al [5] presented a zone-based feature extraction algorithm scheme for the recognition of Multi-Lingual off-line handwritten numerals of four popular South Indian scripts for Kannada, Tamil, Telugu and Malayalam. The Nearest Neighbor, Feed Forward Backpropagation Neural Network and Support Vector Machine classifiers are used for subsequent classification and recognition. V.N Aradhya et al [6] describes recognition system for totally unconstrained handwritten characters for south Indian language of Kannada is proposed. The proposed feature extraction technique is based on Fourier Transform and well known Principal Component Analysis (PCA). The system trains the appropriate frequency band images followed by PCA feature extraction scheme. For subsequent classification technique, Probabilistic Neural Network (PNN) is used. B.V.Dhandra et al [7] presented a novel approach for Kannada, Telugu and Devanagari handwritten numerals recognition based on global and local structural features is proposed. Probabilistic Neural Network (PNN) Classifier is used to classify the Kannada, Telugu and Devanagari numerals separately. Algorithm is validated with Kannada, Telugu and Devanagari numerals dataset by setting various radial values of PNN classifier under different experimental setup. Jagyanseni Panda et al [8] develops an efficient way for recognition of odia numerals by using a non-linear classifier. While recognition of odia numerals, considered the gradient features and curvature feature are used these for the recognition of isolated handwritten odia digits. These features are then reduced using the Principal Component Analysis. The reduced features are then separately used for training the single layer perceptron network adopted for the classification task. The network is trained as in the single layer perceptron network and the weights are updated using lms algorithm. Aji George et al [9] proposed an efficient and robust algorithm for recognition of handwritten isolated Malayalam character using a new type of feature extraction, namely, a combination of statistical features and features obtained by finding the contourlet transform of the pixel value of image is proposed. Experimental results show that these 32 features with feed forward propagation neural network.

The motivation behind our work was to develop a character recognition system, which can further be developed to recognize handwritten South Indian scripts such as Kannada, Tamil, Malayalam, because there is no complete dataset is available for South Indian Scripts. In each South Indian languages consists of more than 250 classes of characters and it is a challenging task to develop a HCR system with these large dataset. However, there are not many reported efforts at developing OCR systems for South Indian Scripts. Therefore, our attempt here is a humble effort towards this noble goal of developing a OCR system, in public domain, for South Indian scripts such as Kannada, Tamil, Malayalam.

The organization of the paper is as follows: In Section 2, we describe the concept of proposed method and its stages. In Section 3, experimental results and analysis were shown in detail. Finally, describes the conclusion of our study.

II. Proposed Method

The proposed method consists of three stages. The first stage consists of Preprocessing. Second stage is the Feature Extraction is performed using PCA and FLDA. These three stages described below.

A. Dataset and Preprocessing

The standard database for South Indian scripts is neither available freely nor commercially, hence, we have collected the sample of Kannada, Tamil and Malayalam handwritten vowels, consonants, modifier glyphs, numerals characters from different professionals belonging to schools, colleges, and employees are collected and created the data set for 33,640 unconstrained Kannada, Tamil, and Malayalam characters from 120 different writers. The collected data set containing multiple lines of isolated handwritten numerals are scanned through a flat bed HP scanner at 300 DPI and preprocessed the each scanned image through Binarization using Otsu's global thresholding technique[10] and is stored in bmp file format. Normalization methods aim to remove the variations of the writing and obtain standardized data. After we perform the binarization operation, we normalize the image size into 50 x 50 pixels.

B. Feature Extraction:

This paper proposes two subspace methods i. Principal Component Analysis (PCA) & ii. Fisher Linear Discriminant Analysis (FLDA) for extracting features in context of Handwritten Kannada, Tamil, Malayalam character recognition.

1) *Principal Component Analysis:* PCA is a classical feature extraction and data representation technique also known as Karhunen-Loeve Expansion[11]. It is a linear method that projects the high-dimensional data onto a lower dimensional space[4].

Let the training set of character images be $X=(X_1, X_2, \ldots\ldots, Xm)$. The average character of the set is defined by $\psi = \frac{1}{M}\sum_{n=1}^{M} x_m$ each face differs from the average by the vector $\phi_i = x_i - \psi$.

This set of very large vectors is then subject to principal component analysis. A set of M orthonormal vectors u_n, describes the distribution of the data. The K^{th} vector u_k is chosen such that

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(u_k^T \phi_n)^2 \qquad (1)$$

is maximum, subject to

$$U_l^T U_k = \delta_{lk} = \begin{cases} 1, if\ l = k \\ 0, otherwise \end{cases} \qquad (2)$$

The vectors u_k and scalars λ_k are the eigen vectors and eigen values, respectively of the covariance matrix.

$$C = \frac{1}{M}\sum_{n=1}^{M} \emptyset_n \emptyset_n^T = A\,A^T \qquad (3)$$

Where the matrix $A=[\phi_1, \phi_2, \ldots\ldots \phi_m]$. The matrix C, however is N^2 by $N^2$and determine the N^2 eigen vectors and eigen values.

The collection of eigen vectors is called feature matrix. This feature matrix represents a collection of different individuals. Further in classification feature matrix is distinguished into individual handwritten character class.

2) *Fisher Linear Discriminant Analysis:* Fisher Linear discriminant analysis is usually performed to investigate differences among multivariate classes, to determine which attributes discriminate between the classes, and to determine the most parsimonious way to distinguish among classes[12].

Steps involved in the feature extraction using FLD for a set of images are as follows.

Suppose that there are M training samples A^k (k=1, 2… M), denoted by m by n matrices, which contain C classes, and the i^{th} class C_i has n_i samples. For each training character image, define the corresponding character image as follows:

1. Calculate the within class scatter matrix (S_w) for the i^{th} class, a scatter matrix (S_i) is calculated as the sum of the covariance matrices of the centered images in that class

$$S_i = \sum_{x \in Xi} (x - m_i)\,(x - m_i\ )^T \qquad (4)$$

where m_i is the mean of the images in the class. The within class scatter matrix (S_w) is the sum of all the scatter matrices:

$$S_w = \sum_{i=1}^{c} S_i \qquad (5)$$

2. Calculate the between class scatter matrix (S_b): It is calculated as the sum of the covariance matrices of the differences between the total mean and mean of each class.

$$S_b=\sum_{i=1}^{c} n_i\,(m_i - m)(m_i - m)^T \qquad (6)$$

where n_i is the number of images in the class, m_i is the mean of the images in the class and m is the mean of all the images.

3. Solve the generalized eigenvalue problem: Solve for the generalized eigenvectors (v) and eigenvalues (λ) of the within class and between class scatter matrices:

$$S_b V=\lambda S_w V \qquad (7)$$

4. Keep first C-1 eigenvectors: Sort the eigenvectors by their associate eigenvalues from high to low and keep the first C-1 eigenvectors W.

5. For each sample Y in training set, extract the feature

$$Z=Y^T * W \qquad (8)$$

Then, use the probabilistic neural network classifier for classification.

D. Classification

This paper proposes Probabilistic Neural Network(PNN) as a classifier method for recognition of offline handwritten characters. Probabilistic Neural Networks use a statistical approach in their prediction algorithm. In particular, the PNN models the popular Bayesian classifier, a technique which minimizes the expected risk of classifying patterns in the wrong category. One of the main criticisms of Bayes' classification techniques is the lack of information about the class probability distributions. The inherent advantage of the PNN architecture is the speed with which it can be trained and can handle data that has points outside the norm thus performing better than other neural architectures [6].

III. Experimental Results

In proposed method the experiment was carried out on a large dataset that we collected 33,640 handwritten characters from different writers belonging to different age groups, qualification and professions. The proposed method groups the 33,640 of Kannada, Tamil, Malayalam characters into 841 classes of characters & process the 841 different classes of characters each with 40 individual writers obtained from scanned documents of the Multi Lingual South Indian Scripts for Kannada, Tamil, Malayalam languages. Each samples of handwritten Kannada, Tamil, & Malayalam consists of Vowels, Consonants, Modifier glyphs (Diacritics), and Numerals. Each experiment is repeated 17 times by varying number of projection vectors *t* (where *t* = 1...50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, and 841). Since *t*, has a considerable impact on recognition accuracy, we chose the value that corresponds to best classification result on the character dataset. All of our experiments are carried out on a PC machine with P4 2.4GHz CPU and 512MB RAM memory under Matlab 7.0 platform.

For a total of 40 samples, the system is trained by varying number of samples s (where s=1...7,14,21,28,35) and remaining samples are used during testing. In experimentation we can list out many observations.

- For character classes between 1-50 with train set=35 & test set=5, obtained the Recognition Accuracy(%) 73.2 for PCA based approach & 72.4 for FLDA based approach. Also we conclude that if increases number of character in the trained set better recognition accuracy is obtained among test set.
- For character classes between 1-841 with train set=35 & test set=5, obtained the Recognition Accuracy(%) 58.64 for PCA based approach & 58.81 for FLDA based approach. Also we conclude that FLDA approach works well with better recognition accuracy than PCA approach. Because in the scattered feature matrix FLDA finds most discriminant projection by maximizing between-class distance and minimizing within-class distance.
- Because of huge data set with 841 of classes of character classes in proposed method we obtained the Recognition Accuracy(%) at maximum of 73.2 for PCA based approach & 72.4 for FLDA based approach.

- In the evaluation of both Feature Extraction approaches (PCA & FLDA) with the classifier PNN from all experiments, the recognition accuracy of FLDA approach is superior than PCA approach. Hence we conclude that recognition accuracy of Multi Lingual Handwritten Characters using FLDA based PNN approach better recognition accuracy can be obtained.

IV. Conclusions

In this paper we have used large dataset(33640 characters) of South Indian Scripts for Kannada, Tamil and Malayalam characters to recognize isolated Handwritten Character Recognition. The proposed method extracts features from well known FLDA and PCA based subspace methods. For classification purpose, we explored PNN classification combined with feature extraction PCA and FLDA for Multi-Lingual isolated handwritten character recognition is presented. Based on the classification the handwritten character recognition accuracy of both the feature extractor approach PCA & FLDA with PNN classifier are analyzed. The performance of the system is tested on totally unconstrained handwritten Kannada, Tamil and Malayalam characters. The dataset containing handwritten characters have very complex structure and shape due to different handwriting styles of each individual writers. The recognition performance for 841 classes of characters are found to be encouraging. The proposed system is tested for Vowels, Consonants, Modifier glyphs (Diacritics), and Numerals. There are about 841 different characters to be tested and need for improvement in recognition accuracy.

References

[1]C.Naveena, V.N. Aradhya "Handwritten Character Segmentation for Kannada Scripts", *World Congress on Information and Communication Technologies*,2012.

[2]Sneha, Satish "Handwritten Character Recognition for major Indian Scripts: A Survey**",** *International Journal of Computer Science &Engineering Technology*, volume. 4, no.04,2013.

[3]J.John, Pramod, K.Balakrishnan, "Handwritten Character Recognition of south Indian Scripts: a review", *National Conference on Indian Language Computing, Kochi*, 2011.

[4]C.Naveena, V.N.Aradhya, S.K.Niranjan, "The Study of Different Similarity Measure Techniques in Recognition of Handwritten Characters", *ICACCI*, 2012.

[5]V.Rajashekararadhya, P.Ranjan," Handwritten Numeral/Mixed Numerals Recognition Of South-Indian Scripts: The Zonebased Feature Extraction Method", *Journal of Theoretical and Applied Information Technology*, volume.7,no.1,pp. 063 - 079**,**2009.

[6]V.N.Aradhya, S.K.Niranjan, G.Kumar, "Probabilistic Neural Network based Approach for Handwritten Character Recognition" *International Conference ACCTA*,volume.1,2010.

[7]B.V.Dhandra, R.G.Benne, M.Hangarge, "Kannada, Telugu and Devanagari Handwritten Numeral Recognition with Probabilistic Neural Network: A Novel Approach", *Recent Trends in Image Processing and Pattern Recognition*,2010.

[8]J.Panda, M.Panda1, A.Samal, N. Das," Odia Handwritten Digit Recognition Using Single Layer Perceptron", *International Journal Of Electronics And Communication Engineering & Technology*,volume.5,issue .4, pp.80-88,2014.

[9]A.George, F.Gafoor, "Contourlet Transform Based Feature Extraction For Handwritten Malayalam Character Recognition Using Neural Network", *International Journal of Industrial Electronics and Electrical Engineering*, volume.2, issue.4,2014.

[10] N. Otsu, "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 1, pp.377–393, 1979.

[11]Karhunen,"Representation and separation of nonlinear PCA type learning", Technical Report A-17,1993.

[12] Fisher R A, Ann. Eugen , Vol 7, 1936, pp. 179-188.

