ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

Handwritten Character Recognition Using ICA and Eigen Vector Approach

Beena K

Lecturer (Electronics), Maharajas Technological Institute, Thrissur, Kerala, India

ABSTRACT: This paper proposes a method for Handwritten English Language Character Recognition using Independent Component Analysis (ICA) and Eigen Vector Approach. English language text is searched in image based English text. In this method two databases of images are created; first one for training purpose and other for testing purpose. Characters to be recognized are placed in test data base. Training database consists of eight characters of English language (a to h) indifferent hand written shapes. Eigen values and Eigen vectors of all the images to beplaced in the Training Data base are calculated. A feature vector for each image of the Train Data base is obtained by using ICA algorithm. In ICA every image in the training set is represented as a linear combination of weighted Eigen vectors. These Eigen vectors (independent components) are obtained from covariance matrix of a training image set. The weights are found out after selecting a set of most relevant Eigen vectors. Feature vector is also created for each image to be identified and placed in 'Test Data base'. Recognition is performed by projecting a test image onto the subspace spanned by the Eigen vectors and then classification is done bymeasuring the minimum Euclidean distance. MATLAB has been used as a simulation tool.

KEYWORDS: Eigen Vector, Handwritten character recognition, ICA, MATLAB, preprocessing, segmentation, text identification

I. INTRODUCTION

Optical character recognition is an active area of research. It is the conversion of hand written and machine written text to computer readable/ editable form. Researchers have proposed several methods for recognition of different languages such as Urdu, English, Gurumukhi, Chinese, Malayalam, etc. Commercial Optical Character Recognition (OCR) for most of these languages is also developed and available in market. The goal of OCR is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process OCR involves several steps including segmentation, feature extraction, and classification.

Handwritten character recognition has been one of the most fascinating and challenging research areas in the field of image processing and pattern recognition. It contributes immensely to the advancement of an automation process and can improve the interface between man and machine in numerous applications. In general, handwriting recognition is classified into two types as off-line and on-line handwriting recognition methods. In the off-line recognition, the writing is usually captured optically by a digital camera and the completed writing is available as an image. But, in the on-line system the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available. The on-line methods have been shown to be superior to their off-line counterparts in recognizing handwritten characters due to the temporal information available with the former. However, several applications including mail sorting, bank processing, document reading and postal address recognition require off-line handwriting recognition systems. As a result, the off-line handwriting recognition also continues to be an active area for research towards exploring the newer techniques that would improve recognition accuracy. The experimental result

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

shows that Independent Component Analysis (ICA) decomposition enables us to separate reflection, shadows and specularities from natural scenes.
The present paper proposes hand written English character recognition by using two approaches, ICA algorithm for feature extraction and Eigen Vector Approach for classification of the handwritten character.

II. METHODOLOGY

INDEPENDENT COMPONENT ANALYSIS (ICA)

Classification using values of all the pixels give good results. The extra dimensionality can lead to random noise which will make measurement less accurate. By lowering the number of dimensions of the data it is possible to run the classification algorithm much faster and on larger datasets. A technique that can reduce dimensionality is ICA, where the data is projected on its k directions of largest variance.
Main aim of ICA is dimensionality reduction and feature extraction. By reducing the dimensionality of data, the speed of computation increases without loosing important information. ICA is applied on Eigen vector approach to reduce the dimensionality of a large dataset. It exploits inherently non Gaussian features of the data and employs higher moments. ICA is required when data cannot be ensembled, when data appears to be noisy. It minimises higher order stastistics such as kurtosis, thus minimising mutual information of the output.
ICA identifies independent components for non gaussian signals but one cannot determine the variance of the independent components and cannot rank the order of dominant components.

EIGEN VECTOR APPROACH

It is an efficient method for character recognition due to its simplicity, speed and learning capability. Eigen faces are set of Eigen vectors used in the computer vision of character recognition. The eigenvector are the independent component of a distribution of characters. The Eigen vectors of covariance matrix are a set of images where an image with N by N pixels is considered a point in N, two dimensional space. Extract the relevant information from the image encode it and compare one image encoding with a database of image encoded similarly. The number of possible Eigen vectors is equal to the number of image in the training set. The primary reason for using fewer Eigen face is computational efficiency.

Eigen Values and Eigen Vectors

The Eigen vector of a linear operator are nonzero vectors, which when operated by the operator results in a scalar multiple of them. Scalar is then called Eigen value (λ) associated with the Eigen vector (X). Eigen vector is a vector that is scaled by linear transformation. When a matrix acts on it, only the vector magnitude is changed not the direction.

$AX=\lambda X$, where A is a vector function.

$\text{Det}(A-\lambda I)=0$, when evaluated becomes a polynomial of degree n, called the characteristic polynomial of A. If A is N by N then there are n roots of the characteristic polynomial. Thus there are n Eigen values of A satisfying the equation $A X_i = \lambda_i X_i$ where i=1,2,.....n. If the Eigen values are all distinct, then there are n associated linearly independent Eigen vectors, whose directions are unique, which span an n dimensional Euclidean space.
Image Representation

Training set of M images of size NxN are represented by vector of size N^2. Feature vector of a image is stored in a NxN matrix. Those two dimensional vector is changed to one dimensional vector. Each image is represented by the vector Γ_i.
Mean Image

It is calculated by $\Psi = (1/M) \sum_{i=1}^{M} \Gamma i$ each image differs from the average by $\phi_i = \Gamma_i - \Psi$ called mean centered image.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

Covariance Matrix

A covariance matrix is constructed as $C= AA^T$ where $A=[\phi_1,\phi_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\phi_M]$ of size N^2xN^2 . In statistics C captures the correlation between all possible pairs of measurements for which the diagonal and off diagonal terms are called the variance and covariance respectively. The correlation value reflects the noise and redundancy in the measured data. Eigen vectors corresponding to AA^T can easily be calculated with reduced dimensionality where AX_i is the Eigen vector and Λ_i is the Eigen value.

Eigen Space

The Eigen vectors of the covariance matrix resemble the input image but look ghostly called Eigen face. Eigen vectors correspond to each Eigen face and discard the faces for which Eigen values are zero thus reducing the Eigen space to an extent.

A face image can be projected into the face space by $\omega_i = U_i (\Gamma i-\Psi)$; i=1, 2, ...M, where U_i is the i^{th} Eigen face. The weight is obtained as above form a vector,$\Omega_i =[\omega_1,\omega_2,\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \omega_M]$.

Testing Sample Classification

a) Read the test image and separate the character from it.

b) Calculate the feature vector of the test face. The test image is transformed into its Eigen vector components. First we compare the line of our image with mean image and multiply their difference with each Eigen vectors. Each value would represent a weight and would save on a vector Ω^T.

$\Omega_{test} = U_i^T(\Gamma_{test} - \Psi)$

$\Omega^T = [\omega_{test}, \omega_1, \omega_2, \ldots\ldots\ldots\ldots\ldots\omega_M]$

c) Compute the average distance (Euclidean distance) between test feature vector and all the training feature vectors. Mathematically recognition is finding the minimum Euclidean distance $\varepsilon_M = \sqrt{(\Omega_{test}-\Omega_i)^2}$ where i=1,2, ...M. The Euclidean distance between two weight vectors thus provides a measurement of similarity between the corresponding images.

d) The class with minimum Euclidean distance shows similarity to test image.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

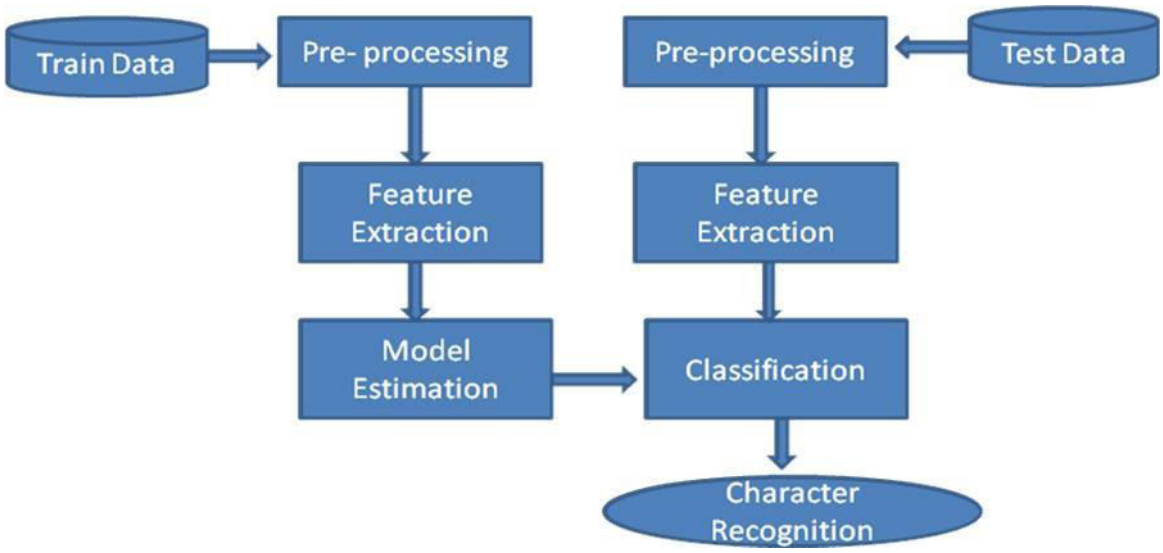
International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

Training and Testing

TRAINING **TESTING (RECOGNITION)**



Training and Testing Block Diagram

PRE- PROCESSING: Processing the data to a suitable form. Eg. Binarization, Segmentation

FEATURE EXTRACTION: Reduce the amount of data by extracting relevant information results in a vector of scalar values. The extracted information from the training set provides important cues of the structures such as intensity, position and shape.

MODEL ESTIMATION: Estimate a model for each class of the training data from the finite set of feature vectors

CLASSIFICATION: Comparing feature vectors to the various models and finding out the closest match

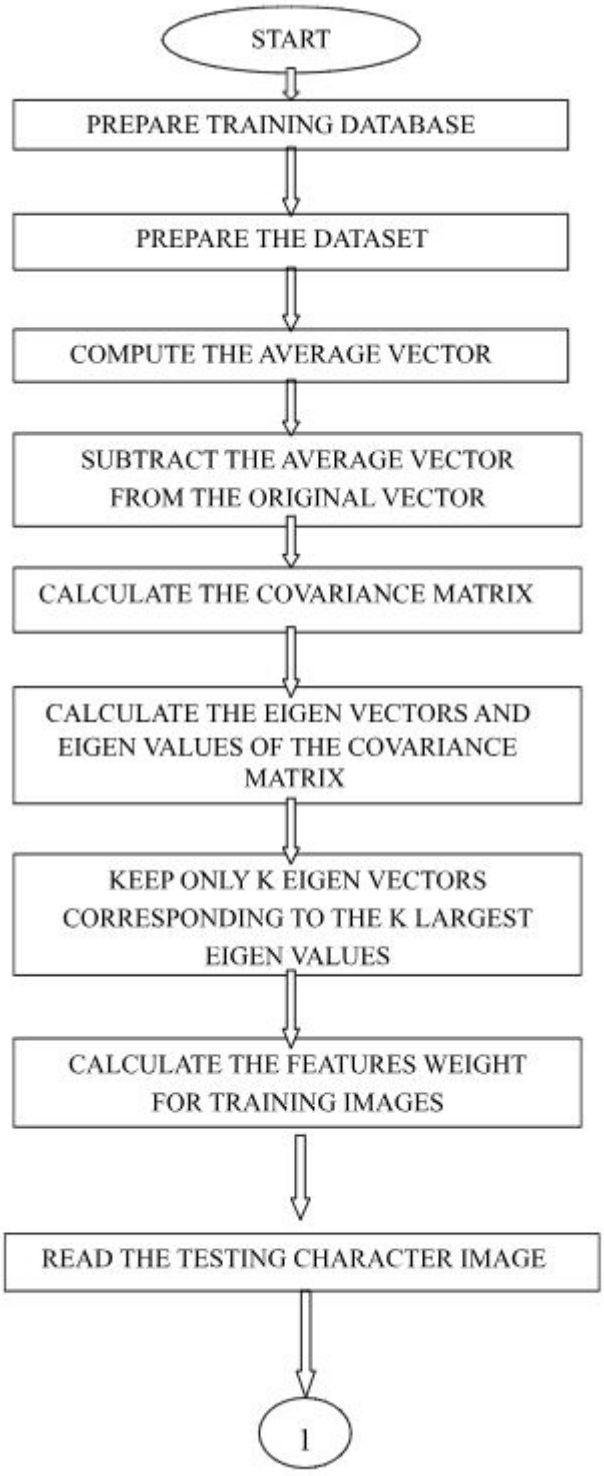
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

Methodology Flow chart

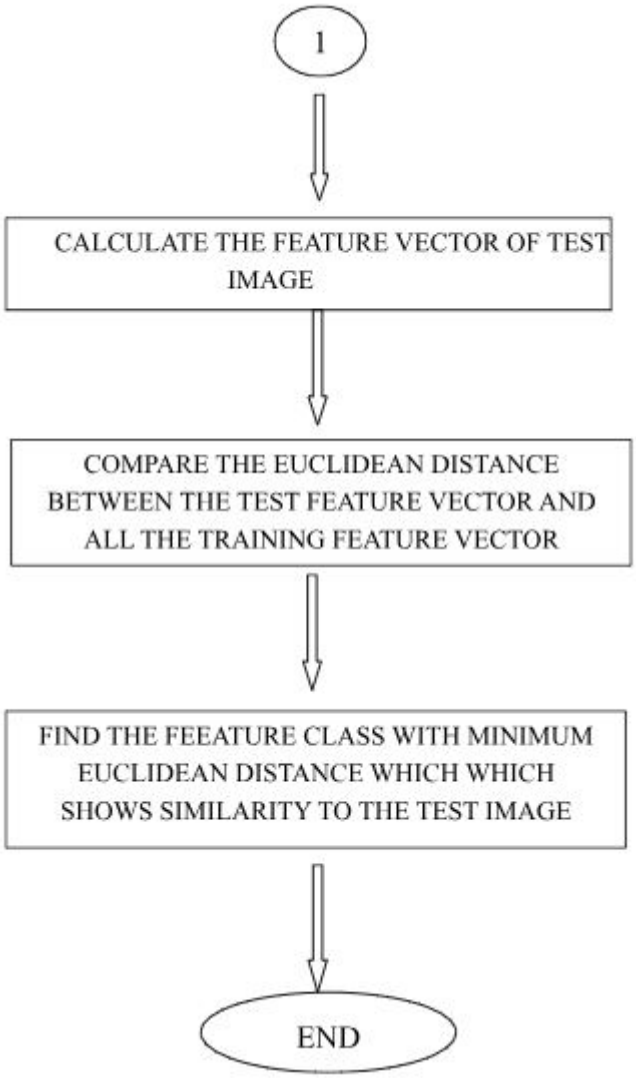


ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016



Methodology Flow chart

SOFTWARE TOOL

MATLAB

MATLAB is the high-level language and interactive environment used by millions of engineers and scientists worldwide. It lets you explore and visualize ideas and collaborate across disciplines including signal and image processing, communications, control systems, and computational finance. MATLAB

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python.
Features

1. High-level language for numerical computation, visualization, and application development
2. Interactive environment for iterative exploration, design, and problem solving
3. Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration, and solving ordinary differential equations
4. Built-in graphics for visualizing data and tools for creating custom plots
5. Development tools for improving code quality and maintainability and maximizing performance
6. Tools for building applications with custom graphical interfaces
7. Functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET, and Microsoft Excel

Development Tools
MATLAB provides high level programming language and development tools with which it is possible to develop and utilize algorithm and applications quickly. It enables to execute commands or group of commands one at a time with no compile or link to achieve the optimal solution.
The various development tools are
MATLAB Editor - Provides standard editing and debugging features, such as setting breakpoints and single stepping
M-Lint Code Checker - Analyzes your code and recommends changes to improve its performance and maintainability
MATLAB Profiler - Records the time spent executing each line of code
Directory Reports - Scan all the files in a directory and report on code efficiency, file differences, file dependencies, and code coverage
Usage of the interactive tool GUIDE allows to design and to edit user interface.

Graphical User Interface Development Environment (Guide)
MATLAB's Graphical User Interface Development Environment provides a rich set of tools for incorporating GUIs in M functions. Using GUI the process of 1) laying out a GUI (ie its button, pop up menus etc) and 2) programming the operation of the GUI are divided into two easily managed and independent tasks.
The resulting graphical M function is composed of two identically named files,

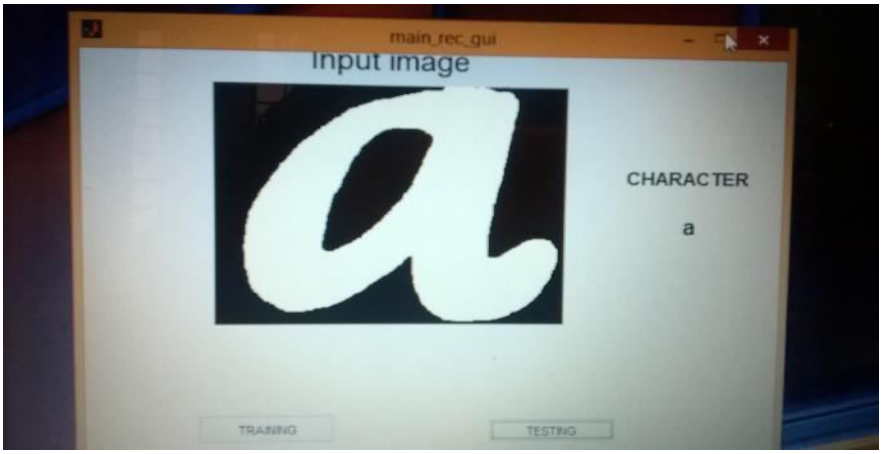
1. A file with extension .fig called FIG file contains a complete graphical description of all the functions of GUI objects and their spatial arrangements. A FIG file contains binary data that need not be parsed when the GUI based M function is executed.
2. A file with extension .m called a GUI M file contains the code that controls the GUI operation. This file include functions that are called when the GUI is launched and excited and call back functions that are executed when a user interacts with GUI objects eg. Pushing a button.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016



Input Image and Output Character Representation using GUIDE

Hardware Requirements

Hardware Requirements for 32 Bit MATLAB

Operating Systems	**Processors**	**Disk Space**	**RAM**
Windows 8.1 Windows 8 Windows 7 Service Pack 1 Windows Vista Service Pack 2 Windows XP Service Pack 3 Windows XP x64 Edition Service Pack 2 Windows Server 2012 Windows Server 2008 R2 Service Pack 1 Windows Server 2008 Service Pack 2 Windows Server 2003 R2 Service Pack 2	Any Intel or AMD x86 processor supporting SSE2 instruction set*	1 GB for MATLAB only, 3–4 GB for a typical installation	1024 MB (At least 2048 MB recommended)

Hardware Requirements for 32 Bit MATLAB

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

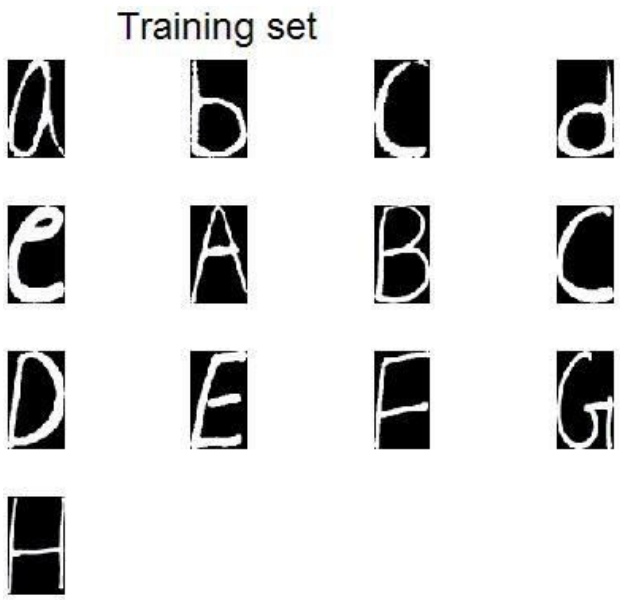
Vol. 5, Issue 3, March 2016

III. IMPLEMENTATION, RESULT AND ANALYSIS

TESTING PARAMETERS

Matlab R2013a is used for coding. The camera image is converted into gray scale image, as gray scale images are easier for applying computational techniques in image processing. The gray scale image is scaled for a particular pixel size as 48 x
84 because input images can be of different matrix size when we take it for recognition.
Training Data Set

Database for different set of conditions is maintained. In this work I worked with 8 English alphabets in different ways to create different set of images. Variations in input image orientation also changes the output, therefore input images for different orientation are also taken for recognition.



Training Images of Eight Alphabets

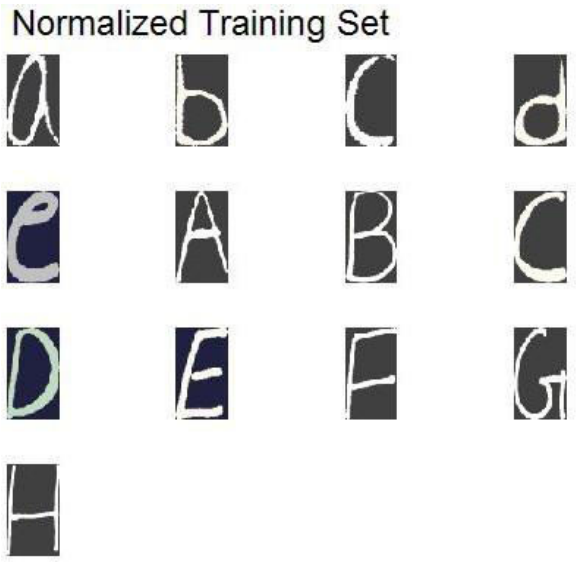
The images of character received from the acquisition stage have different distortions such as translation, rotation and scaling. The character normalization is applied for standardizing the size of the character.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016



Normalized Training Set Images

Mean Training Set Image Eigen Matrices

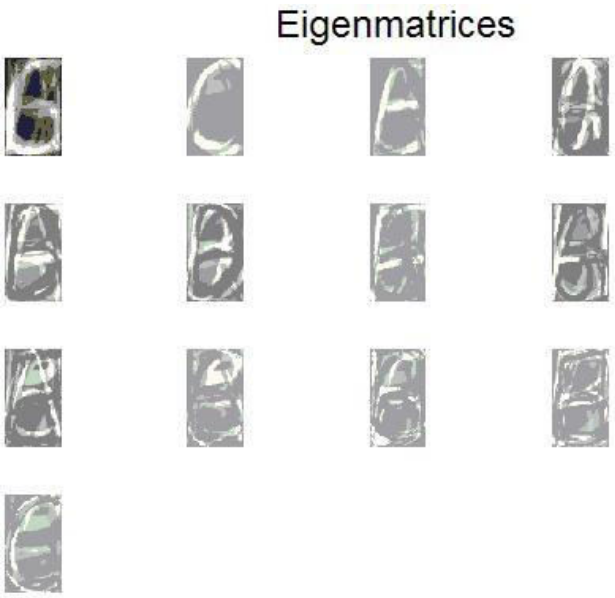
The Eigen vectors corresponding to the covariance matrix define the Eigen face which has ghostly appearance and a match is found if new image is close to these images.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016



Eigen Matrices of Training Set Images

Testing Conditions

Different intensity of light on image may change the recognition just as bright light causes image saturation. If the size of the image is varied the recognition may alter accordingly. A test image for recognition is tested by comparing to the stored dataset. The camera image pixel size is 320 x 240. The input image is resized into 48 x 84 pixel size

Test Image

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

Image Processing

Binarisation: The first step of image processing is binarization. The acquired RGB image is converted into gray image with 256 levels of gray scale. Then by selecting an appropriate threshold level image binarisation is carried out. The binarised image consists of only 1 and 0. By complementing this image we will get a white character on a black background, called whitening. This process is carried out because Matlab functions only on white image.

Binarised Image

Inverted Image

Segmentation: Segmentation is defined as the process of partitioning an image into a set of non overlapping regions and is carried out to find the better position of the shape points according to the appearance information. We can use segmentation for image compression, object recognition etc.

Segmented Image

Mean Image

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

Reconstructed Image

The character is recognized by finding the minimum Euclidean distance between the training and testing weight vectors.

Image after Classification

RESULT AND ANALYSIS

Thirteen different images were taken to test for eight different handwritten characters. The light intensity was kept low. Size variation of the test image is not altered much.

Input Image

Output Image

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Vol. 5, Issue 3, March 2016

IV. CONCLUSION

Two databases 'Train Data base' and 'Test Data base' have been created in MATLAB for the proposed work. Train Data base has been used for training and classification purposes. Characters to be recognized are placed in Test Data base. ICA transforms images in Train Data base. Training images are then projected onto Eigen space. Test images which are to be recognized are also projected onto Eigen space. Classification is done by the Eigen vector approach and the character to be recognized is shown by the algorithm.

REFERENCES

1. Khan, Rehan Ullah, Nasir Ahmad Khan, Khwaja Naveed. (2012), 'Urdu Character Recognition using Principal Component Analysis', International Journal of Computer Applications Vol. 60, No.11.
2. Jyostna Grace M. and Subhashini K. (2014), 'Image Binarization Based On Ica Approach for Optical Character Recognition', International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 3, No. 8.
3. Pradeep J, Srinivasan E. and Himavathi S. (2011), 'Diagonal Based Feature Extraction for Handwritten Alphabets Recognition System using Neural Network', International Journal of Computer Science & Information Technology (IJCSIT) Vol. 3, No. 1.
4. Munish Kumar M. K, Jindal, Sharma R.K, 'SVM Based Offline Handwritten Gurmukhi Character Recognition'.
5. Liton Chandra Paul and Abdulla Al Sumam (2012), 'Face Recognition using Principal Component Analysis Method', International Journal of Advanced Research in Computer Engineering & Technology Vol. 1, No. 9.
6. Seiichi Ozawa, Yoshinori Sakaguchi, Manabu Kotani, 'Feature Extraction of Handwritten Characters using Supervised and Unsupervised Independent Component Analysis' Faculty of Engineering, Kobe University, Kobe 657-8501, Japan
7. Mathew Turk and Alex Pentland, 'Eigen Faces for Recognition' The Media Laboratory Massachusetts Institute of Technology.
8. He Zhiguo, Zhong Yuquan, Cao Yudong, 'Handwritten Chinese Character Recognition based on Support Vector Machine and Kernel Independent Component Analysis' School of Computer Science, Panzhihua University, Panzhihua 617000, China.
9. Anil K. Jain, Fellow, Ieee, Robert P.W. Duin, Jianchang Mao (2000), 'Statistical Pattern Recognition', Senior Member, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1.
10. Lajish V, Sita G, Sunil Kumar Kopparapu, (2009) 'Handwritten Character Recognition Using PCA', State Space Point Distribution NCC 2009, January 16-18, IIT Guwahati.