Identifying the Character by Applying PCA Method using Matlab

P.Subbuthai
Department of Electronics and Instrumentation
Bharathiar University
Coimbatore
India

Azha Periasamy
Department of Electronics and Instrumentation
Bharathiar University
Coimbatore
India

S.Muruganand
Department of Electronics and Instrumentation
Bharathiar University
Coimbatore
India

ABSTRACT
Optical character recognition is getting more and more useful in daily life for various purposes. The aim of the paper is to find the number and English alphabets in the symbol of times new roman, arial, arial block size of 72, 48.Many researches have been done on many types of characters by using different approaches. In this recognition system was implemented by using of principal component analysis (PCA) algorithm. This algorithm is based on an Eigen value and Euclidean distance. PCA is practical and standard statistical tool in modern data analysis that has found application in different areas such as face recognition, image compression, and neuroscience.

General Terms
Binary. Edge, filling image and PCA

Keywords
PCA**,** Eigen value, Euclidean distance

1. INTRODUCTION
Character recognition system has received considerable attention in recent years due to the tremendous need for digitization of printed documents. Manual assignment of text data from images is time consumption and costly. For this then the automation of text extracted from images is one of the challenging area in the image processing. In this paper, numbers and English alphabet has to be recognized. The English letter can be consists of two cases, that are uppercase and lower case. In this paper focused on uppercase character in the style of Times new roman, Arial, Arial block of the size of 72, 48. English language is used all over the world for the communication purpose, also in many Indian offices such as railways, passport, income tax, sales tax, defense and public sector undertakings such as bank, insurance, court, economic centers, and educational institutions etc these approaches are done by principal component analysis. PCA is a linear transformation, which rotates the axes of image space along lines of maximum variance. The rotation is based on the orthogonal eigenvectors of the covariance matrix generated from a sample of image data from the input channels. The output from this transformation is a new set of images channels, which are also referred to as eigenchannels. The main use of this to reduce the dimensionality of a data set while retaining as much information as possible. The process of character recognition process can be divided into following stages namely preprocessing, feature extraction and recognition.

2. PREPROCESSING STEPS

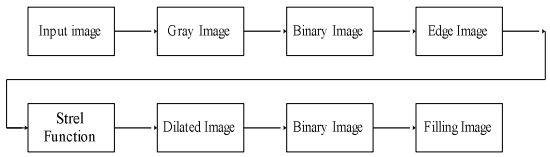

Fig 1: Block Diagram for Preprocessing

Preprocessing operations generally fall into three categories: image acquisition, image conversion, morphological operation. Their respective blocks are shown in Figure 1. Digital image acquisition is the creation of digital images. Typically from a physical scene. The term is often assumed to imply or include the preprocessing, compression, storage printing and display of such images. An image conversion consists of three steps: RGB image to gray image and gray image to binary image and then finally binary image into an edge image. The first step of the image processing is binarization. The colorful image represented by 3 coefficients red, green and blue from the acquisition unit must be converted to the images with 256 levels of gray scale[1]. Then select an appropriate threshold to achieve the image binarization[2]. Following by converting the grayscale image into binary image which consists of only 0 and 1[3]. Then the gray image is converted into edge image. Dilation image and filling images are took place in morphological operation. Edges of images are detected using appropriate thresholding and then further dilated operation using appropriate structure element[4]. The dilated images are converted into filling image through binary image. The filling image is used to reduce the number of connected components and the command *bwlabel* is used to calculate the connected component[3]. The next step is to obtain the bounding box of character. Bounding box is referring to the minimum rectangular box that is able to encapsulate the whole character[5]. Single character has been detected from this bounding box of the character. For template matching the image is resized into 74*50 by using bilinear method.

3. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest

variance on the second coordinate and so on[6]. The main idea of using PCA for character recognition is to express the large 1-D vector of pixels constructed from 2-D character image into the compact principal components of the feature space[7]. PCA involves a mathematical procedure to transform a number of correlated variables into a number of uncorrelated variables[8]. The principal component analysis is one of the most successful techniques and compression. PCA can also do prediction, redundancy removal, feature extraction, data compression.

PCA is a fundamental multivariate data analysis method which is encountered into a variety of areas in neural networks, signal processing and machine learning. It is an unsupervised method for reducing the dimensionality of the existing data set and extracting important information. PCA does not use any output information, the criterion to be maximized is the variance. PCA can be applied to economically represent the input digital images by projecting them onto a low-dimensional space constituted by a small number of basis images.

PCA is applied to extract R number of principal components corresponding to top R Eigen values are chosen to the digit images. Here each digit image of size M× M is converted into a column vector y_i, by concatenating the pixels in row orders. The vector y_i is of size P × 1 where $P = M^2$.

The mean of the set of N training samples $[y_1, y_2, y_3, \ldots y_N]$ is defined by

$$m=\frac{1}{N}\sum_{i=1}^{N} y_i \quad (1)$$

Each vector y_i differs from the mean (m) by the difference vector (Q_i), where

$$Q_i = y_i - m \quad (2)$$

Let, $A= [Q_1\ Q_2, \ldots, Q_N]$ (3)

Where A has a dimension P×N

Then the covariance matrix is defined as

$$C = AA^T \quad (4)$$

Thus, matrix C is of dimension P× P and we can find out P different Eigen values (λ_js) and corresponding P different Eigen vectors (λ_js) of C by solving the following equations

$$C\lambda_j = \lambda_j\sqrt{}_j \quad j=1, 2,\ldots, N \quad (5)$$

Here each √j is a P dimensional vector and is also called the principal component. In the present work, the magnitudes of these P numbers of Eigen values of C are sorted and top R Eigen values and the corresponding Eigen vectors are retained. Thus the number of principal components considered here are R which is much lower than P. These principal components are used for extracting features from each of the digit images in the training as well as test samples. For the kth digits image, features are extracted using the following equation

$$fkj = v^T j(Y_k - m) \quad j=1, 2, \ldots, R \quad (6)$$

Therefore, the feature set for any digit consists of R number of elements.

4. EIGEN VECTORS AND EIGEN VALUES OF THE COVARIANCE MATRIX

Principal component analysis transforms a set of data obtained from possibly correlated variables into a set of values of uncorrelated variables called principal components. The number of components can be less than or equal to the number of original variables. The first principal component has the highest possible variance, and each of the succeeding components has the highest possible variance under the restriction that it has to be orthogonal to the previous component. We want to find the principal components, in this case eigenvectors of the covariance matrix of facial images. The first thing we need to do is to form a training data set. 2D image I_i can be represented as a 1D vector by concatenating rows. Image is transformed into a vector of length N= mn.

$$I=\begin{bmatrix} y_{11} & y_{12} & y_{1n} \\ y_{21} & y_{22} & y_{2n} \\ y_{m1} & y_{m2} & y_{mn} \end{bmatrix} \xrightarrow{\text{Concatenation}} \begin{bmatrix} y_{11} \\ \vdots \\ y_{mn} \end{bmatrix}_{1*N} = Y \quad (7)$$

Let M such vectors y_i (i = 1, 2,..., M) of length N form a matrix of learning images, Y. To ensure that the first principal component describes the direction of maximum variance, it is necessary to center the matrix. First we determine the vector of mean values **m,** and then subtract that vector from each image Vector.

$$m=\frac{1}{N}\sum_{i=1}^{N} y_i \quad (8)$$

$$Q_i = y_i - m \quad (9)$$

Averaged vectors are arranged to form a new training matrix (size N×M);

$$A= (Q_1, Q_2, Q_3) \quad (10)$$

The next step is to calculate the covariance matrix C, and find its Eigenvectors e_i and eigenvalues λ_i:

$$C= \frac{1}{M} \sum_{n=1}^{M} m_n m_n^T \quad (11)$$

$$Ce_i= \lambda_i e_i. \quad (12)$$

Covariance matrix C has dimensions N×N. From that we get N eigenvalues and eigenvectors. One of the theorems in linear algebra states that the eigenvectors e_i and eigenvalues λ_i can be obtained by finding eigenvectors and eigenvalues of matrix C1 = A^TA (dimensions M×M). If v_i and μ_i are eigenvectors and eigenvalues of matrix A^TA, then

$$A^TAv_i=\mu_i v_i \quad (13)$$

Multiplying both sides of equation (13) with A from the left, we get:

$$A\ A^TAv_i = A\mu_i v_i ,$$
$$A\ A^T(Avi) =(Av_i)\mu_i ,$$
$$C\ (Avi) = \mu_i(Av_i) \quad (14)$$

Comparing equations (12) and (14) we can conclude that the first M–1 eigenvectors e_i and eigenvalues λ_i of matrix C are given by Av_i and μ_i, respectively. Eigenvector associated with the highest eigenvalue reflects the highest variance, and the one associated with the lowest eigenvalue, the smallest variance. Eigenvalues decrease exponentially so that about 90% of the total variance is contained in the first 5% to 10% eigenvectors. Therefore, the vectors should be sorted by eigenvalues so that the first vector corresponds to the highest eigenvalue. These vectors are then normalized. They form a new matrix E so that each vector e_i is a column vector. The dimensions of this matrix are N×D, where D represents the desired number of eigenvectors. It is used for projection of data matrix A and calculatation of z_i vectors of matrix $Z= (z_1, \ldots, z_M)$

$$Z=E^TA. \quad (15)$$

Each original image can be reconstructed by adding mean image (m) to the weighted summation of all vectors e_i.

The last step is the recognition of characters. Image of the character we want to find in training set is transformed into a vector P, reduced by the mean value m and projected with a matrix of eigenvectors

$$\omega = ET\ (P - m) \quad (16)$$

5. EUCLIDEAN DISTANCE

The definition of geometric distance formula comes from the Pythagorean theorem. Euclidean distance is come from the extension of the Pythagorean theorem into higher dimensions and it is commonly used in distance in the place of pattern recognition. The distance between an input vector x and a training sample q can be calculated by the following equation[9].

$$F(x,q)^2 = \sum_{i=1}^{N}(x - p_i) \quad (17)$$

6. RESULTS AND DISCUSSION

The system is proposed to recognize the Numbers and English alphabet characters. English letters can be of 2 cases viz. Uppercase and lowercase, here we consider text only in upper case in the symbol of Times new roman, Arial, Arial Block of the size of 72 and 48.There are many application for this recognition.

1.Semiconductor device [integrated chip number] identification.
2. Security.
3. Identification of stolen Vehicles.
4. Object Identification.

Principal component analysis is a statistical feature extraction technique which has been applied successfully. For template matching resize the image into 74*50. Most of them use the gray scale image [or] binary image for feature extraction. Here the output of the preprocessing is fill image. Before the fill processing noise reduction are took place by using Dilation. Fill image extracts the character and recognize the character by using Euclidean distance. Identify the character based on the Euclidean distance between the character and of the character models. Here the character PIC 16 F877A is identified using PCA is shown in the following Figures. Times new roman and its size 72 is applied in PIC 16 F877A.

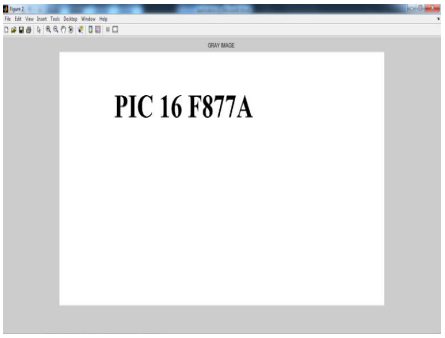

Fig 2: Input image

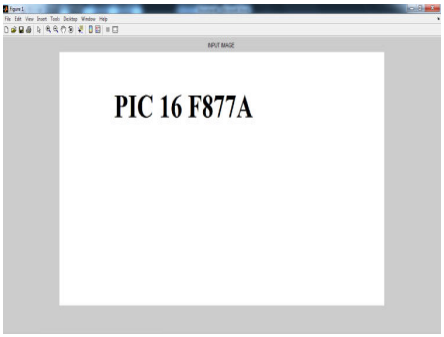

Fig 3:Gray image

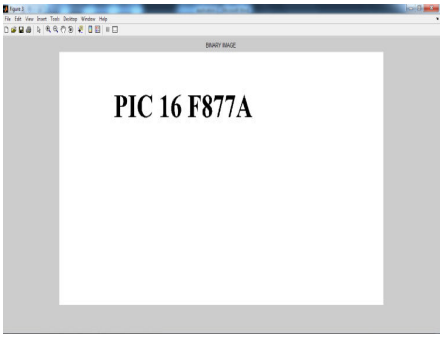

Fig 4: Binary image

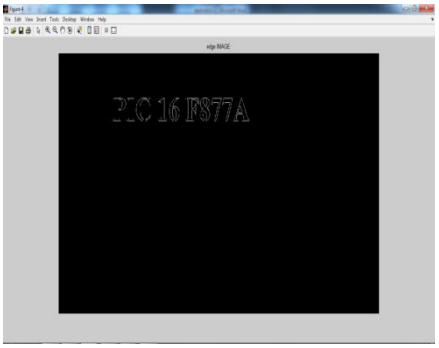

fig 5:Edge image

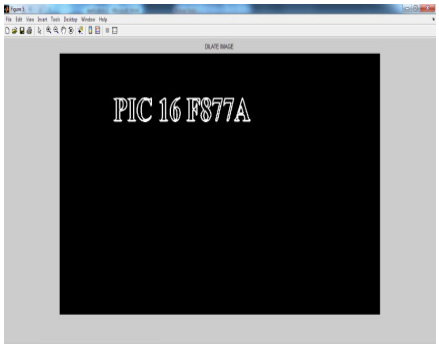

Fig 6:Dilated image

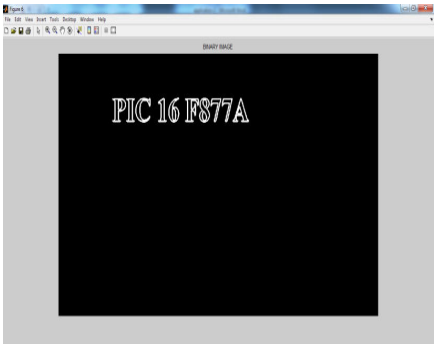

Fig 7:Binary image

Fig 8: Filled image

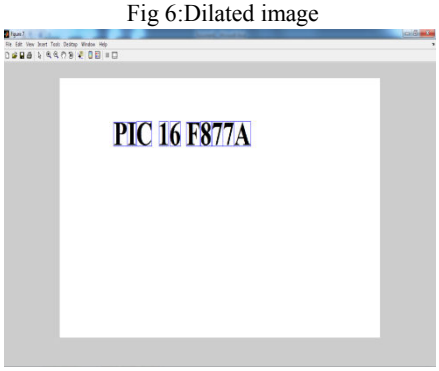

Fig 9: Cropping image

The input image is shown in Figure.2. It is an RGB image. RGB image is an array of color pixels, where each color pixel is represented as red, green, and blue. The range of the value is [0,255] or [0, 655535] for RGB images of class uint8 or uint16 respectively. Then the RGB image converted into Gray image, where the Gray image is an M*N dimensional, but RGB image is M*N*3 dimensional. Gray image is shown in Figure 3.

Figure 4, explains the conversion of gray image into binary image using gray threshold method. Binary image is a method to reduce color images into two colors, black and white.

Logically this greatly reduces the complexity of the image. One algorithm to perform binarization is the threshold algorithm. In binary image threshold decides the black and white of the image. Figure 5 shows edge detection of a binary image. The result of applying an edge detector to an Image may lead to a set of connected curves that indicate the boundaries of objects, the boundaries of surface orientation. It reduces the amount of data to be processed and may therefore filter out information that may be regarded as less relevant, while preserving the important structural properties of an image. Next step of the conversion is shown in Figure 6, it is the conversion of edge to dilate image using structural element. The dilation process is performed by laying the structural element on the image and sliding it across the image in a manner similar to convolution. Figure 7 shows the conversion of dilate image to binary image and again binary image is converted into fill image is shown in Figure 8. Principal Component Analysis is used to determine the most discriminating features between the images. Following operations are sused in PCA analysis.

1. Calculate the mean image
2. Calculate the deviation of each image from mean image.
3. Computing the difference image for each image in the training set
4. Merging all centered images.
5. Calculate the covariance matrix
6. Find Eigen vectors and Eigen values.

Suppose if the covariance matrix has a dimension of M*N, then we gets N Eigen value and Eigen vectors. Using fill image to find groupings then cropping the image using region props of bounding box is shown in Figure 9. Euclidean distances between the projected test image and the projection of all centered training images are calculated.

7. CONCLUSION

An efficient Recognition System for numbers and English characters has been implemented in this paper. We shave applied conventional PCA scheme for Subsequent recognition purpose. Various techniques are used in the preprocessing phase before implementing the classification of numbers and English characters.

To improve the performance of this prototype, the improved feature extraction method and the preprocessing techniques are possibly required. From these, decided to implement the Euclidean Distance in the final recognition product. Standard PCA is used to reduce dimensionality of each class and the orthogonal distance to the class subspace used for classification. Thus the numbers and English characters are identified by using principal component analysis in the symbol of Times new roman, Arial, Arial block of the size of 72, 48.

REFERENCE

[1] Mirosław Miciak," Character Recognition Using Radon Transformation and Principal Component Analysis in Postal Applications", Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 495 – 500.

[2] Yang Yang, Xuhui Gao, Guowei Yang a," Study the Method of Vehicle License Locating Based on Color Segmentation". Advanced in Control Engineering and Information Science, Procedia Engineering 15 (2011) 1324 – 1329.

[3] anish lazrus, siddhartha choubey, sinha g.r, "an efficient method of vehicle number plate detection and Recognition", International Journal of Machine Intelligence ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 3, 2011, pp-134-137.

[4] Dileep Kumar Patel, Tanmoy Som, Sushil Kumar Yadav, Manoj Kumar Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric", Journal of Signal and Information Processing, 2012, 3, 208-214.

[5] Velappa Ganapathy, and Kok Leong Liew," Handwritten Character Recognition Using Multiscale Neural Network Training Technique", World Academy of Science, Engineering and Technology 15 2008.

[6] J.T.Jolliffe, "principal componenet analysis",springer series in statistics,2nd ed,springer ,2002.

[7] k.kim," face recognition using principal componenet analysis",Dcs, university of Maryland,college park,usa 2003.

[8] Pramod Kumar Pandey, Yaduvir Singh, Sweta Tripathi, "Image Processing using Principle Component Analysis", International Journal of Computer Applications (0975 – 8887) Volume 15– No.4, February 2011.

[9] Dileep Kumar Patel, Tanmoy Som, Sushil Kumar Yadav, Manoj Kumar Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric", Journal of Signal and Information Processing, 2012, 3, 208-214.

AUTHOR'S PROFILE

P.Subbuthai received M.Sc and submitted dissertation M.Phil in the department of Electronics and Instrumentation from Bharathiar University, Coimbatore, Tamil Nadu, and India in 2011 and 2012 respectively. Her research interests include Digital Image Processing, Optical Character Recognition.

Azha Periasamy received his M.Sc degree in Applied Physics and Computer Electronics in 1988 from Urumu Dhanalakshmi College, Trichy, Tamil Nadu, and India. He was awarded M,Phil degree in 1995 from Bharathiar University, Coimbatore, Tamil Nadu, India. He is working as an Assistant Professor in the Department of Electronics and Instrumentation, Bharathiar University, Coimbatore, India. His field of interest is molecular Physics, VLSI System Design and Digital Image Processing.

S.Muruganand received his M.Sc degree in Physics from Madras University, Chennai, Tamil Nadu, India, and the Ph.D degree from Bharathiar University in 2002. He is working as an Assistant Professor in the Department of Electronics and Instrumentation, Bharathiar University, Coimbatore, India. His area of interest is Embedded Systems, Sensors, Physics and Digital Signal Processing.