Unsupervised Learning Dimensionality Reduction Algorithm PCA For Face Recognition

Vivek Kumar
Department of Computer Science
National University of Science &Technology-MISIS
Moscow, Russian Federation
vivekkumar0416@gmail.com

Dr. Denis Kalitin, Prayag Tiwari
Department of Computer Science
National University of Science &Technology-MISIS
Moscow, Russian Federation
kalitindv@gmail.com, prayagforms@gmail.com

***Abstract*—Somechallenges of developing face recognition system is to train examples of different poses, illuminations, contrast and background conditions. Even though if the system works okay with the tested data set; it fails big time to perform accurately with new data set with different attributes. This manuscript is focusing on the mentioned problem statement by deploying PCA and implementing the concept of splitting the dataset in training, cross-validation, and test set to compute misclassification error. It enables the algorithm to be a better fit to be used for the new training set. Principal component analysis (PCA) popularly known as dimensionality reduction algorithm often referred as black box, which is used widely but poorly understood. The hope with this manuscript is to provide road-map for a deep understanding the PCA and mathematics of its step by step process to develop the desired recognition system. The interesting part of this manuscript is the selection of database. The source of images were databases of AT&T labs, CVL Face Database, Face 94, Essex database and real times images taken by us with a standard professional camera. The images included in the database are of 10,000 people from various racial origins, it spans evenly over the age group that is from 18 to 60 years old. All the experiments were executed on MATLAB and the results were obtained with fair accuracies which are discussed at the end.**

***Keywords*—principal component analysis;face recognition; database; misclassification error**

I. INTRODUCTION (HEADING 1)

A human can perceive and distinguish a large number of faces in their lives, even after meeting after a long time. Even if some changes have come in existence like small or big changes in appearance, aging, color, hairs human still can recognize the faces, and this analyzing capability is remarkable. This led to the evolution of designing the systems which can work in closer proximity to human system functioning.

Principal component analysis (PCA) has been called a standout among the most significant outcomes from applied linear algebra. It is a widely used mathematical tool for high dimension data analysis and is deployed in several types of analysis just within the domains of visualization and computer graphics alone, PCA has been used for dimension reduction [1], face and gesture recognition [2], dimension reduction [1], motion analysis and synthesis [3], clustering [4], and many more. It is a way of identifying patterns indata and expressing the data so as to highlight their similarities and contrasts. Since patterns in data can be elusive to find in high dimensional data, where the luxury graphical representation is not accessible [5]. The other primary preferred standpoint of PCA is that, once these patterns are found in the data, it can be further compressed, i.e. by reducing the number of dimensions, without much loss of information and thus very efficiently used in image compression. With a small extra effort, PCA gives a guide to how to lessen an intricate dataset to a lower dimension to uncover the sometimes covered up, a simplified structure that often underlies it.

In spite of the fact that PCA is a powerful tool for reducing dimensions and uncovering relationships among complex data items, it has been customarily seen as a “black box" approach [6]. The transformation of coordinates from original data space into Eigenspace makes it tedious for the user to understand the underlying relation.

II. STRUCTURE OF PCA

To understand the implementation methodology PCA-based face recognition system, it is important to first know its processes and sub-processesassociated with it.Face recognition system is determined by how to extract feature vector exactly and to classify them into a group accurately [7]. Therefore, it is necessary to closely focus at the feature extractor and classifier.

Basically, the face recognition contains three processes which are image preprocessing, feature extraction and clustering. Within these three process, there are further different steps which are presented below in a lucid manner.

A. Data Preprocessing

To use the test Image for face recognition system, it is needed to be in usable form beforehand. Therefore this step is used to solve some common problems like illumination variations, normalization, face detection, facial features detection, headposes estimation and face image dimensions normalization [8].

*1) RGB to grayscale Image conversion:*For many applications color information does not help to identify important edges or other features [9].

*2) Cropping:*It is very obvious that test image which is undergoing the implementation, can be very large or hasunne-

ISBN:978-1-5090-6471-7/17/$31.00 ©2017 IEEE

Fig. 1. Cropped image

cessary objects besides subject in thebackground. So, cropping is done to re-frame an image and separate the face in order to reduce theimpact of background.

*3) Binarization:*This process is used to transform a grayscale image to a binary image. This type of image format stores an image as a matrix and contain only two classes of pixels, white as thebackground and black as foreground. Thresholding is a basic way for binarizing the gray scale image. A grayscale image has pixel intensity values from 0 to 255 very similar to thecolour image. Values below 128 forms one group and above 255 second. Assign a 0 for black and a 1 for white. It can be represented as by the following equation.

$$B(x,y) = \begin{cases} 0 & if\ T(x,y) \ \le T(x,y) \\ 1 & otherwise \end{cases} \quad (1)$$

In the above equation B(x,y) is the result of binary image processing and T stands for the threshold for distinguishing the two groups.

B. Feature Extraction

Prior to developing PCA, there are several steps we need to undergo like normalization and depending upon the data feature scaling. Along with that, in the development of face recognition, we need to find out quantities such feature empirical mean, Eigenvalue and Eigen matrix, covariance matrix. This phase can be summed up in following stages:

- To extract the information from the image and acquire data.
- Organizing the data set.
- Mean Normalization.
- Information extraction from the image and acquired data.
- Constructing the covariance matrix from the acquired data.
- Constructing the covariance matrix from the acquired data.
- Sort Eigenvectors and finding principal Eigenvectors.
- Transformation of sets in a new subspace.
- Calculate difference matrix and find minimum difference.
- Map and store the data.

III. FORMULATION OF PCA

A. Mathematics of PCA

The prime objective of PCA is to reduce the data from n-dimensions to k-dimensions i.e. it tries to project the data by finding a lower dimensional subspace onto, so as to minimize the squared projection errors and the sumof squared projection errors. This seems to have a lot of mathematics and complex derivation procedure. So, in this section, we will elaborate the mathematical concept of finding the mentioned values and matrices.

Suppose we are given a dataset $x = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ which have attributes of m different types of objects. So as preprocessing step i.e. mean normalization we first compute the mean of each feature and then we replace each feature,$x^{(i)}$ with $x^{(i)}$- μ, and so this makes each feature now have exactly zero mean. The mean vector is N ×1 column vector μ

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \quad (2)$$

The different features have very different scales. In order to ensure that different attributes are all treated on the same scale, we rescale each coordinate to have unit variance. Let say

$$\sigma_j^2 = \frac{1}{m} \sum (x_j^i)^2 \quad (3)$$

And replace each $x^{(i)}$with$x^{(i)}/\sigma_j$. Having done the preprocessing of the data we compute now the covariance matrix is symbolized byΣ (Greek letter sigma), defined by the following equation

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)})(x^{(i)})^{\mathrm{T}} \quad (4)$$

The transpose of μ is 1× Nvector, therefore, thecovariance matrix is a symmetric N× N square matrix. By PCA we can eliminate the redundancy by transforming the original feature space into a subspacein terms of Principal Components (PCs). To compute the covariance matrix we need to compute eigenvectors of the matrix sigma. It can be done by singular value decomposition [10]. But using MATLAB is rather an easy and one step process for which the code is

```
[U, S, V] = svd (Sigma);
```

By SVD we get the Eigenvalue and Eigen matrix namely 'U' a N×N square matrix having the columns of vectors $\{u^{(1)} \dots\dots u^{(m)}\}$. 'S' a diagonal N× N matrix having vector elements $\{s^{(1)} \dots\dots s^{(m)}\}$ in the diagonal only. As we need to reduce the dimension and project the data onto vectors $u^{(1)}$tou$^{(k)}$. Let's take the first k column of the matrix U and construct a matrix say it $U_{reduce} \in \mathbb{R}^{n\times k}$. To obtain a k dimensional vector say it 'Z' we need to establish a relation between Z and U_{reduce} which is given by; Z= $(U_{reduce})^{T} \times X$ where Z$\in \mathbb{R}^{k\times 1}$.

Once eigenvectors of the covariance matrix are found, we sort them by eigenvalue, in decreasing order. We can ignore the components of lesser significance for small Eigenvalueswith-

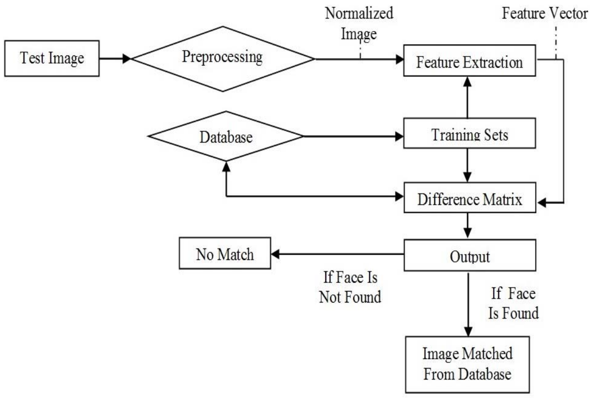


Fig. 2. Diagram of face recognition system

out loosing significant information. The higher the eigenvalue, the more characteristic features particular eigenvector portrays about a face. Now we form a feature vector, which is just a fancy name for a matrix of vectors which is constructed by using the eigenvectors which we kept from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns.

B. Choosing the number of principal components (k)

The average squared error in this context can be represented by this equation where$x^{(i)}$, is the original data and x^{i}_{approx} is the projected data on the lower dimensional surface.

$$\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x^{i}_{approx}\right\|^{2} \tag{5}$$

The total variation in the data can be written as:

$$\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}\right\|^{2} \tag{6}$$

A typical way to choose $k$is to choose it to be that smallest value where ninety-nine percent of variance can be retained.

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x^{i}_{approx}\right\|^{2}}{\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}\right\|^{2}} \tag{7}$$

An excellent style manual for science writers is [7].

IV. DEVELOPING FACE RECOGNITION SYSTEM

A common belief among researchers is that a large training set can be a great help in significant improvement of algorithm's performance. So they spend ample time of their research in finding data and developing a large training set. However, it stands true but in certain conditions, the large training set is unlikely to help, as it is dependent upon features and attributes selection. Simultaneously the features are also interrelated and connected with each other and converse. For instance, if a feature does not contain enough information to predict y accurately it moreover leads to the problem of high bias (under fitting) and variance (overfitting). PCA works upon the principle of reducing the data from n-dimensions to k-dimensions. In other words, dimension usually refers to the number of attributes, and feature vectors spanned in these; say N dimensions by these basis vectors is partitioned by the decision boundaries that ultimately define the different classes in the multiclass problem of face recognition.

For example given a training set $X = \{x^{(i)}; i = 1, \ldots, m\}$having attributes of m different types of sports bikes, such as their maximum speed, acceleration etc. Two different attributes x_i and x_j respectively give a bike's maximum speed measured in miles per hour and kilometers per hour are unknown to us. These two attributes are almost linearly dependent, upon each other with differences introduced by rounding off to the nearest mph or kph. Thus, here actually data lies approximately on an $n-1$ dimensional subspace. The thing we want to depict by this example in here, is PCA compares the different attributes which are closely similar and merges them to reduce the no dimensions. Therefore, the selection of features and attributes play crucial role with the initial development of the system.

A. Developing the Database

The interesting part about the database is the elements comprising it. The database that has been developed for experiments is a mixture of almost all prominent types of images that are RGB, YCbCr or gray-scale image. The source of images were databases of AT&T labs, CVL Face Database, Face 94, Essex database and real times images taken by us with a standard professional camera. The images included in the database are 10,000 of people from various racial origins, it spans evenly over the age group that is from 18 to 60 years old. Some individuals are wearing glasses and beards. The lightning and illumination conditions were natural, artificial, a mixture of tungsten and fluorescent. The reason of this diversity was to monitor the performance of our algorithm on new training sets i.e. our goal is to reduce the cross-validation error instead of the test set error which is more important for performance enhancement of algorithm when new training set is tested.

B. Experiments

While selecting the parameters for learning algorithm we think about choosing the parameters to minimize training error.On the other hand, getting a really low value of training error might be a good thing, but just because a hypothesis has low training error, that doesn't mean it is necessarily a good hypothesis. A recommended way to overcome this is segregation of the dataset i.e. splitting it into three groups namely training set, cross-validation set and test set. Out database contains 10,000 images and in it, one person has almost 10-15 images in general of different moods. It sums up to approximately 1000 images of the distinct person. So by above discussion, we split the dataset between the training set (600 images), cross-validation set (200 images) and test set (200 images). We implemented the set to learn the training error and computed the test set error. There also exist a misclassification error in case the algorithm mislabels the sample data on the basis of threshold assumption.The misclassification error can be defined as;

$$Error\ (x,y) = \begin{cases} 1 & if\ x \geq 0.5\ ,\ y = 0 \\ & or\ if\ x < 0.5\ ,\ y = 1 \\ 0 & otherwise \end{cases} \quad (8)$$

This gives us a binary 0 or 1 error result based on a misclassification. The average test error for the test set is shown in 'Equation 9' which gives the proportion of the test data that was misclassified.

$$\text{Test Error} = \frac{1}{m_{test}} \sum_{i=1}^{m} Error\ (x,y) \quad (9)$$

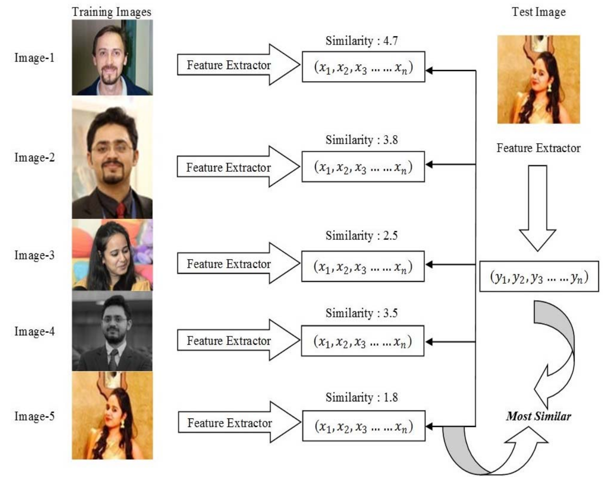
Similarly, the training, cross-validation and test errors for the observed parameter θare given by;

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (x^{(i)} - y^{(i)})^2 \quad (10)$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (x_{cv}^{i} - y_{cv}^{i})^2 \quad (11)$$

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (x_{test}^{i} - y_{test}^{i})^2 \quad (12)$$

It should be noted that for low threshold the feature selections will be more acute and hence algorithm has to face difficulty to get a right match. While for high threshold it will be rather easier but it will be more prone to misclassification error. Snapshots of the experimental results are included which show that performance of the face recognition system with all type of existing images is accomplished with fair accuracies. Although there is the scope for improvement in recognition accuracy for the image with least feature details by means of finding the right threshold and reducing the cross-validation error by exposing it to new training examples.



V. IMPLEMENTATION AND RESULT

Having trained the data set of test images with PCA we observed that the maximum deviation in recognizing the faces has been observed for the images with features not very prominent and understandable like side view and bad illumination condition.

The output is categorized in two categories match or no match. If the test image could not find the match from the database. It says no match. The reason behind it is the threshold set of the algorithm.

For higher values of the threshold, it is likely to happen to the test image with a not clear face also yields the output, but It will be not a right match, as it will match with the another face of the databasse. For the same image varying the threshold can give aconcrete idea about its suitability. It should be noted that for low threshold the feature selections will be more acute and hence algorithm has to face difficulty to get a right match. While for high threshold it will be rather easier but it will be more prone to misclassification error.

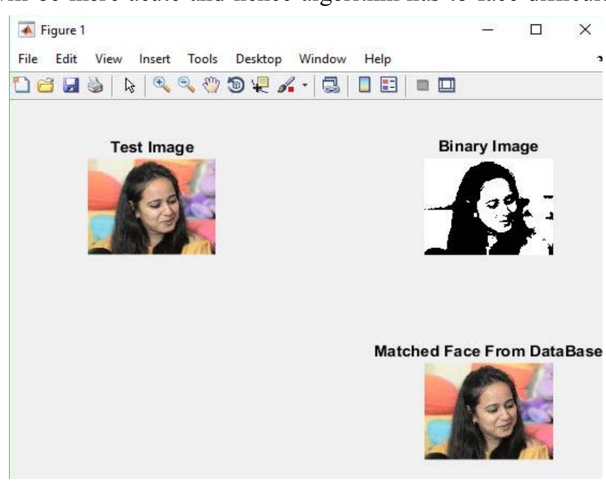


Fig. 3. Figure showing the simplified recognition process.

Fig. 4. Snapshot showing the matched results.

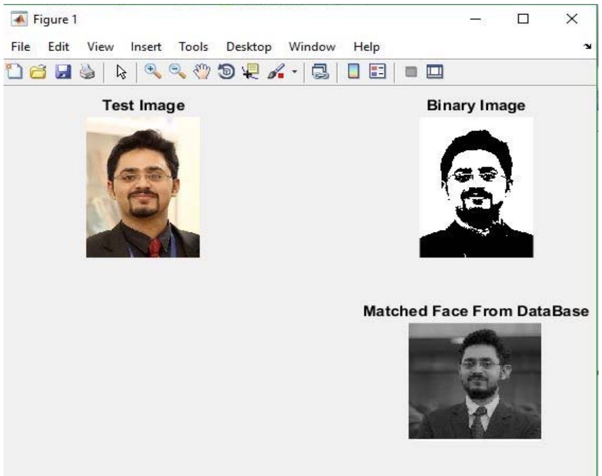


Fig. 5. Snapshot showing the matched results.

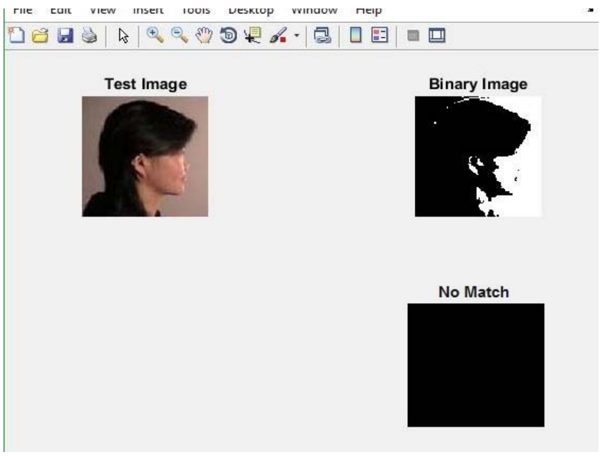


Fig. 6. Snapshot showing no match result for side view facing image.

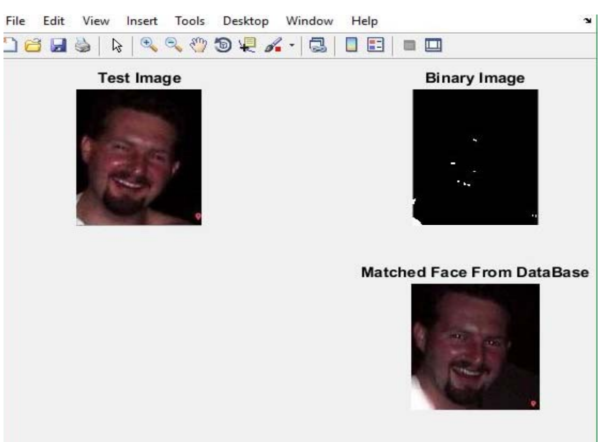


Fig. 7. Snapshot showingmatched result in bare illumination condition.

Snapshots of the experimental results are included which show that the performance of face recognition system is accomplished with fair accuracies for all type of exisiting test images. Although there isscope for improvement in recognition accuracy for the image with least feature details by means of finding the right threshold and reducing the cross-validation error by exposing it to new training examples.

VI. Conclusion and Future Work

In this research work, PCA is applied to find the matching image from the database for the test image. There are most common problems encountered were because, lighting variations, head pose, and different faces dimensions. The results can be altered by varying the threshold. This variation declares the selectivity of user on the basis of details of images. Harsher the threshold precise would be the result. For future work, we are going to implement PCA for 'Sign Language and Gesture Recognition' through neural network. Once having fair results with PCA we will implement other algorithms like Support Vector Machine and Stochastic Gradient Descent for increasing the system accuracy [11]. Lastly, we will implement deep convolution neural networks with Python as a scripting language for further improvements [12],[13].

References

[1] The template will number citations consecutively within brS. Huang, M. O. Ward, and E. A. Rundensteiner. Exploration of dimensionality reduction for text visualization. In CMV '05: Proceedings of the Coordinated and Multiple Views in Exploratory Visualization, page 63-74, Washington, DC, USA, 2005. IEEE Computer Society.

[2] Y. Koren and L. Carmel. Visualization of labeled data using linear transformations. InfoVis, 00:16, 2003.

[3] A. Safonova, J. K. Hodgins, and N. S. Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. ACM Trans. Graph., 23(3):514{521, 2004.

[4] M. Turk and A. Pentland. Face recognition using eigenfaces. In Computer Vision and Pattern Recognition, pages 586-591. IEEE Computer Society, 1991.

[5] Lalita Kumari, Swapan Debbarma, Nikhil Debbarma, Suman Deb, Image pattern matching using principal component analysis method. International Journal of Advanced Engineering & Application, June 2011.

[6] I. T. Jolliffe. Principal component analysis. Springer, second edition, 2002. Ahmad, T., Jameel A. and Ahmad, B.(2011).

[7] Pattern recognition using statistical and neural techniques.Computer Networks and Information Technology OCCNIT), 2011 International Conference.

[8] Sacco, E. "Image preprocessing methods for face recognition applications".University Of Genova. 2010.

[9] HaiYan, C. and JunHong, L.(2010). "The image process technologies in face recognition".Information Science and Engineering (ICISE), 2nd International Conference.

[10] E. Rasmus Bro and T. Kolda. Resolving the sign ambiguity in the singular value decomposition. Journal of Chemometrics, 22(2):135-140, 2008

[11] Abdulaziz A. Alorf, "Performance Evaluation of the PCA VersusImproved PCA (IPCA) in Image Compression, and in Face Detection and Recognition", page 537-546, Future Technologies Conference, San Francisco, USA, 2016.

[12] Shanshan Guo, Shiyu Chen and Yanjie Li, "Face recognition based on convolutional neural network and support vector machine".Proceedings of the IEEE International Conference on Information and Automation Ningbo, China, page 1787-1792, 2016.

[13] Y.-H. Kim, H. Kim, S.-W. Kim, H.-Y. Kim and S.-J. Ko, "Illumination normalisation using convolutional neural network with application to face recognition." Electronics Letters ,Vol. 53 No. 6 pp. 399–401, 2017.