North Carolina State University

ECE 763 Computer Vision

Project Report:

Face Image Classification

Author: Nazneen Kotwal (200205655)
Instructor: Dr. Tianfu Wu

INDEX

Sr. No.	Title	Pg. No.
1.	Objective	2
2.	Dataset Preparation	2
3.	Single Gaussian Model	3
4.	Mixture of Gaussian Model	4
5.	T Distribution Model	6
6.	Mixture of T Distribution Model	8
7.	Factor Analysis	10

Objective:

In face detection, we attempt to infer a discrete label $\omega \varepsilon \{0, 1\}$ indicating whether a face is present or not based on observed image data. Through this project I aim to achieve face image classification using Gaussian model, Mixture of Gaussian model, t-distribution, Mixture of t-distribution, and Factor Analysis. For each classification model the results are reported as follows:

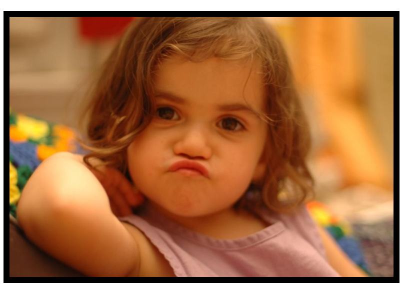
1. Visualization of the the estimated mean(s) and covariance matrix for face and non-face respectively.
2. Evaluation of the learned model on the testing images using 0.5 as threshold for the posterior by computing the false positive rate, false negative rate and the misclassification rate.
3. Plot of the Receiver Operating Characteristic Curve (ROC Curve) to evaluate the diagnostic ability of the binary classifier.

Dataset Preparation:

The Image Database used for this project is a subset of the Helen Dataset (http://www.ifp.illinois.edu/~vuongle2/helen/) which was created using annotated Flickr images [1]. The Database consists of RGB images with annotations for the face bounding boxes. I extracted 1000 training images and 100 testing images from the dataset through careful filtering to avoid anomalies such as spectacles. The annotations provided in the form of rectangular coordinates are used for the extraction of the face boundary from the images. The non-face image data set is created using the same extracted images by carefully cropping the background while ensuring no or minimal overlap exists between the annotated face and background section. All the testing and training images are resized to a 60 x 60 dimension. It is ensured that training face images and testing face images are separate, that is, no face testing images are from the same person in the training set of face images. The resulting dataset consists of 2000 training and 200 test images. A sampling of the dataset is depicted below. To recognize if a provided image patch contains a face or not we first concatenate the RGB values to form a 1 x 10800 vector x.

Sample:

a)



b)

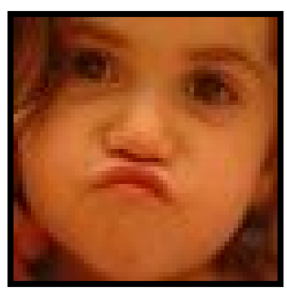


Figure 1. a) Sample Image from Helen Database b) Extracted image patch from annotation.

Model 1: Single Gaussian Model:

The single Gaussian model also known as the multivariate gaussian mode is a generative approach to face detection in which we calculate the probability of the data x and parameterize this by the world state ω. The data is described with a multivariate normal distribution as given in equation (1).

$$Pr(x/\omega) = Norm_x[\mu, \Sigma] \quad (1)$$

The parameters of this model $\Theta = \{\mu_0, \Sigma_0, \mu_1, \Sigma_1\}$ include the mean and covariance for the face image and non face image normal distributions. Since the full covariance matrix contains $D(D+1)/2$ parameters which we cannot uniquely specify with a training data set of just 1000 images, we use the diagonal form of the covariance matrix. We use the maximum likelihood approach as given in equation (2) to obtain the mean μ_0 and covariance Σ_0 concerned with the background regions from the subset of training data with non-face images (S_0). Similarly. μ_1 and covariance Σ_1 are concerned exclusively with faces and can be learned from the subset of training data which contains faces (S_1) using equation (3). Figure 2 shows the maximum likelihood estimates of the parameters with the diagonal form of the covariance matrix.

$$\widehat{\mu_0}, \widehat{\Sigma_0} = armax\left[\prod_{i \varepsilon s_0} Pr(x_i/\mu_{0,}\Sigma_0)\right] \quad (2)$$

$$\widehat{\mu_1}, \widehat{\Sigma_1} = armax\left[\prod_{i \varepsilon s_1} Pr(x_i/\mu_{1,}\Sigma_1)\right] \quad (3)$$

The mean of the face model clearly captures classified information. The covariance of the face is larger at the edges of the image, which usually contains hair or background. On evaluating the learned model on the testing images using 0.5 as the threshold for the posterior we obtain the false positive rate, the false negative rate and the misclassification rate as given in Table 1.

I was able to implement this model without running into computation hiccups such as Singular Matrix and Determinant calculation opposed to the later models. Hence the single gaussian model was tested on RGB images whereas the latter models were tested on images converted to the grey scale.

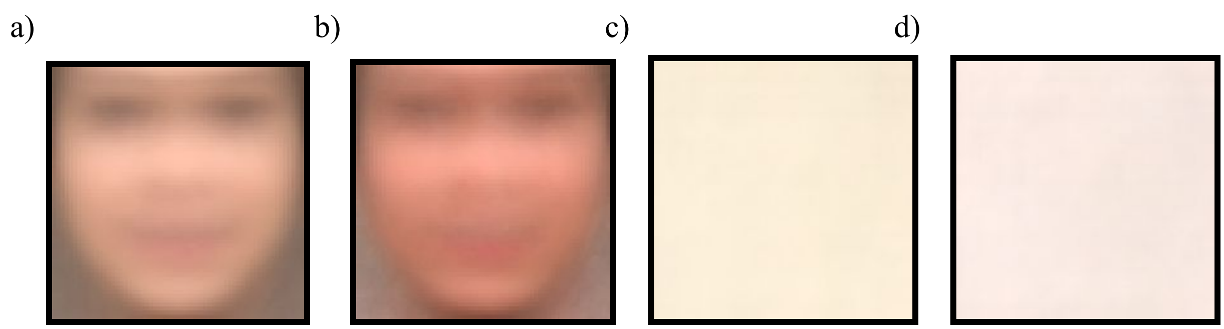


Figure 2. Maximum Likelihood fit based on 1000 training examples per class. a) Mean from face data b) Covariance from face data c) Mean from non-face data d) Covariance from non-face data. The background model has little structure: The mean is uniform and the variance is high everywhere.

Evaluation Criteria	
False Positive Rate	0.2
False Negative Rate	0.11
Misclassification Rate	0.18

Table 1. Evaluation of Single Gaussian Model

The Receiver Operating Characteristic Curve (ROC Curve) is a graphical plot that illustrates the diagnostic ability of a binary classifier system .The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity. The false positive rate, also known as fall-out can be calculated as (1- specificity). The ROC curve for the simple gaussian model is given Figure 3.

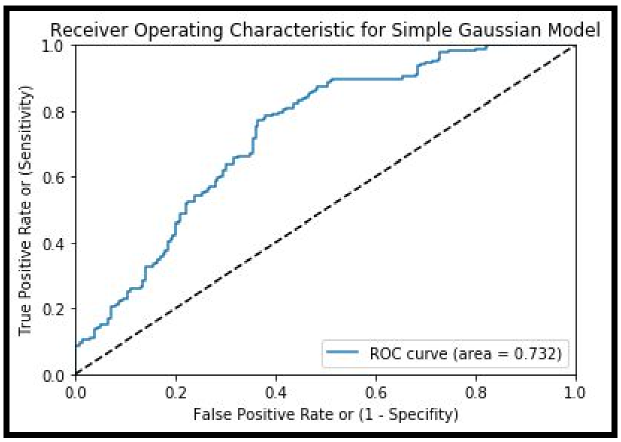


Figure 3. ROC Curve for the Single Gaussian Model

<u>Model 2: Mixture of Gaussian Model</u>:

The normal distribution is unimodal and neither the faces nor the background regions are well represented by a pdf with a single peek. To make the density multimodal we introduce mixture models. In the Mixture of Gaussians (MoG) model the dataset is described as a weighted sum of K normal distributions.

$$Pr(x/\omega) = \sum_{k=1}^{K} \lambda_k \, Norm_x[\mu_k , \Sigma_k] \qquad (4)$$

It is not possible to learn the parameters of the mode $\Theta = \{\mu_k, \Sigma_k, \lambda_k\} \; for \; k = 1...K$ using the maximum likelihood approach since we cannot solve the resulting equations in a closed form. Instead, we express the observed density as a marginalization and use the EM algorithm to learn the parameters. In the E-step we maximize the bound with respect to the posterior probability distribution of each hidden variable give the observation and current parameter setting. We compute the probability $Pr(h_i = k / x_i, \theta^{(t)})$ that the k^{th} normal

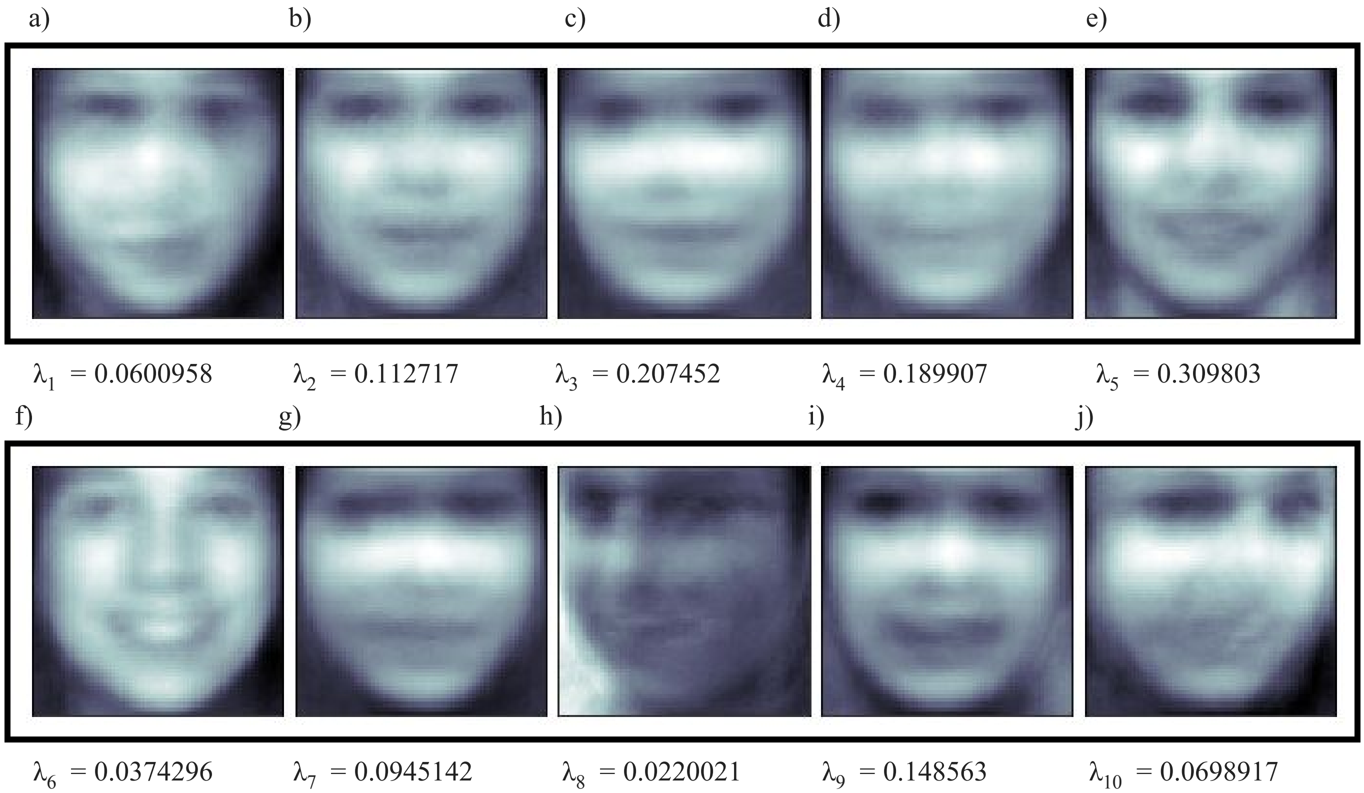


Figure 4. (I) Mixture of Gaussian Model for face data. a-j) Mean vectors for the mixture of ten Gaussians fitted to the face data set. The model has captured variation in the mean luminance and chromaticity of the face and other factors such as the pose and background color. Numbers indicate the weight of each component.

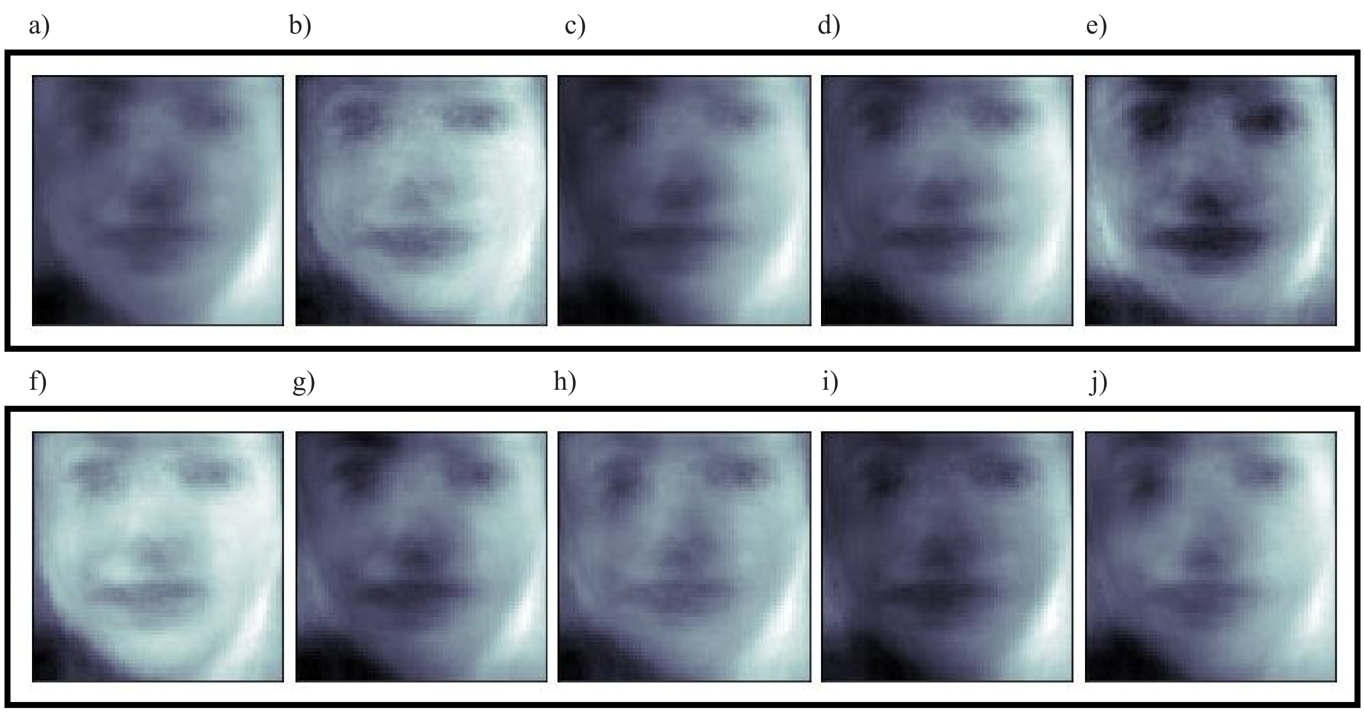


Figure 4. (II) a-j) Covariance vectors for the mixture of ten Gaussians fitted to the face data set.

distribution was responsible for the i^{th} data point. In the M -step we maximize the bound with respect to the model parameters. The E am M steps are alternated until the bound on the data no longer increases and the parameters no longer change. On applying the MoG model on the face data we obtain the mean vectors as shown in Figure 4. The weight λ_k for each component is indicated by the number.

The model captures variation in orientation of the face (pose) as well as the mean luminance. On evaluating the learned model on the testing images using 0.5 as the threshold for the posterior we obtain the false positive rate, the false negative rate and the misclassification rate as given in Table 2.

Evaluation Criteria	
False Positive Rate	0.07
False Negative Rate	0.08
Misclassification Rate	0.075

Table 2. Evaluation of Mixture of Gaussian Model

We notice that the misclassification rate for the MoG model is much lower in comparison to the single gaussian model. The ROC curve for the simple gaussian model is given Figure 5.

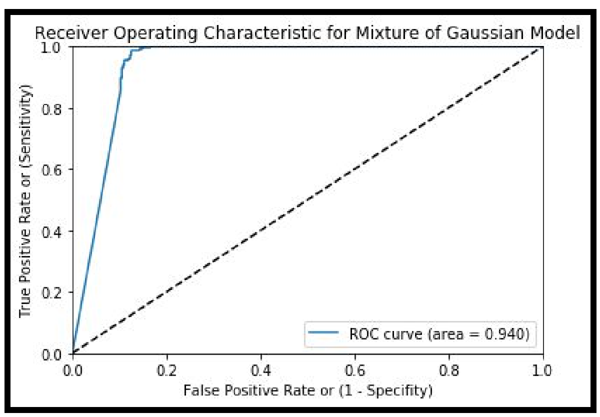


Figure 5. ROC Curve for the Mixture of Gaussian Model

Model 3: T - Distribution Model:

The normal distribution is not robust. The height of the normal pdf falls off very rapidly as we move into the tail due to which a single outlier can dramatically affect the estimates of the mean and the covariance of the classifier model. To make the density robust, we replace the normal distribution with the t distribution since the length of the tail is parameterized. The data is described using the probability density function of a univariate t distribution which has the parameters $\Theta = \{\mu, \Sigma, \nu\}$ where the degree of freedom ν controls the length of the tails: when ν is small there is considerable weight in the tails. As ν tends to infinity the distribution approximates a normal more closely and there is less weight in the tails.

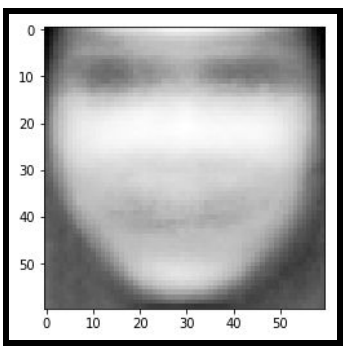
$$Pr(x/\omega) = Stud_x[\mu, \Sigma, \nu] \qquad (5)$$

The t distribution is the marginalisation of the joint distribution $Pr(x, h)$ between the observed variable x and a hidden variable h. The prior distribution over the hidden variable h has a gamma distribution. The conditional distribution $Pr(x/h)$ with a variance that depends on h. So the t distribution can be considered as an infinite weighted sum of normal distributions with variance determined by the gamma prior. We use the EM algorithm to learn the parameters from the training data. In the E step we treat each data point x_i as if it were generated from one of the normals in the mixture where the hidden variable h_i determines which normal.

The E-step computes a distribution over h_i which in turn determines which normal created the data. Outliers in the data set will be explained best by the normal distribution with large covariances: for these distributions h is small. There is no closed form solution for the degree of freedom ν and hence we perform a one dimensional line search to maximize the bound.

When we fit the face data set to the t distribution with a diagonal scale matrix, the mean and covariance look visually similar to those for the normal model as shown in figure 6.

a)



b)

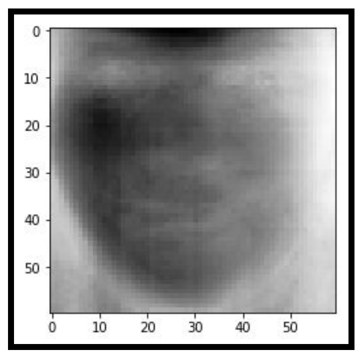


Figure 6. T distribution model for face dataset fitted on 1000 training examples per class. a) Mean from face data b) Covariance from face data

On evaluating the learned model on the testing images for face classification using 0.5 as the threshold for the posterior we obtain the false positive rate, the false negative rate and the misclassification rate as given in Table 3. The ROC curve for the simple gaussian model is given Figure 7.

Evaluation Criteria	
False Positive Rate	0.05
False Negative Rate	0.13
Misclassification Rate	0.09

Table 3. Evaluation of T - Distribution Model

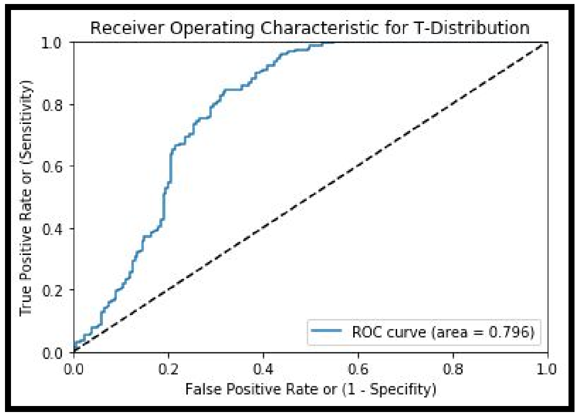


Figure 7. ROC Curve for the T-Distribution Model

Model 4: Mixture of T - Distribution:

In many applied problems, fitting the data to the normal distribution is not appropriate since the tails of the normal distribution are often shorter than required. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. The problem of providing protection against outliers in multivariate data is a very difficult problem and increases with the difficulty of the dimension of the data. We consider the fitting of mixtures of (multivariate) t distributions. The data is described using the distribution given in equation (6)

$$Pr(x/\omega) = \sum_{k=1}^{K} \lambda_k \, Stud_x[\mu_k, \Sigma_k, \nu_k] \qquad (6)$$

The t distribution provides a longer tailed alternative to the normal distribution. Hence it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters. Also, the use of t components gives less extreme estimates of the posterior probabilities of component membership of the mixture model. With this t mixture model-based approach, the normal distribution for each component in the mixture is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the degrees of freedom ν. As ν tends to infinity, the t distribution approaches the normal distribution. Hence this parameter ν may be viewed as a robustness tuning parameter. The use of a mixture model of t distributions provides a sound mathematical basis for a robust method of mixture estimation and hence clustering [2]. On training the face data with mixture of t distribution having three mixture components, I obtained the mean images as shown in Figure 8 which seem to capture the pose of the faces.

On evaluating the learned model on the testing images for face classification using 0.5 as the threshold for the posterior we obtain the false positive rate, the false negative rate and the misclassification rate as given in Table 4. The ROC curve for the simple gaussian model is given Figure 9.

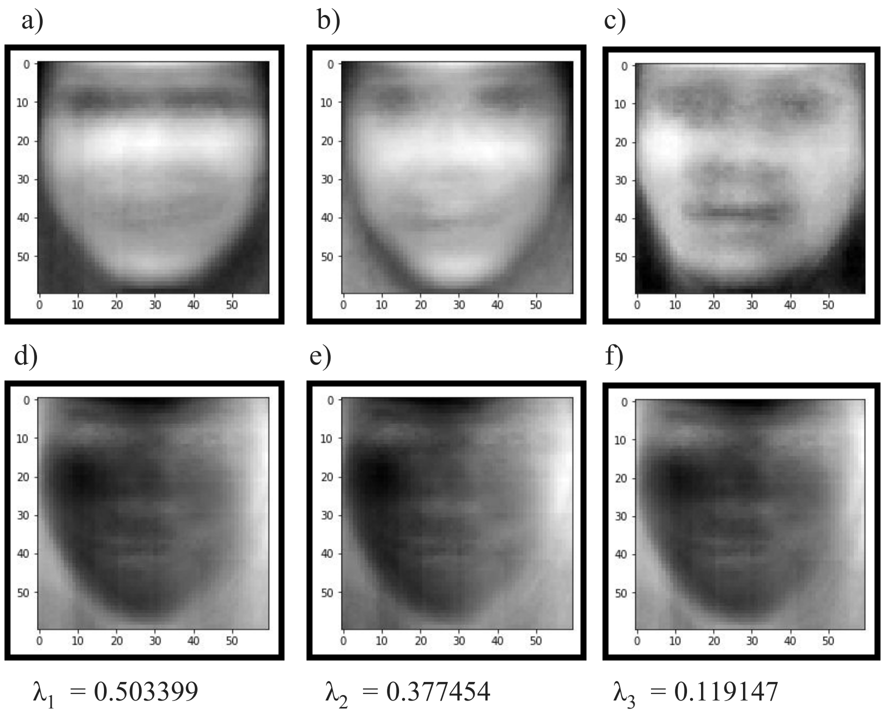


Figure 8. Mixture of T-Distribution Model for face data. a-d) Mean vectors for the mixture of three T-Distributions fitted to the face data set. d-f) Covariance vectors for the mixture of three T-Distributions fitted to the face data set.

Evaluation Criteria	
False Positive Rate	0.01
False Negative Rate	0.26
Misclassification Rate	0.14

Table 4. Evaluation of Mixture of T Distribution Model

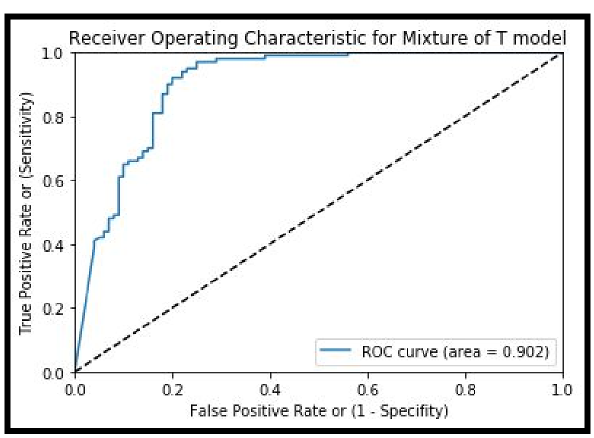


Figure 9. ROC Curve for the Mixture of T Distribution Model

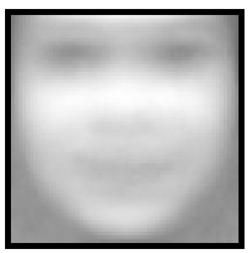
Model 5: Factor Analysis Model:

Visual data are often very high dimensional. In the face detection task, data comes in the form of a 60 x 60 x 3 = 10800 dimensional vector. To model the data with a full multivariate normal distribution we require the covariance matrix of dimension 10800 x 10800 for which we would need a very large number of training examples to get a good estimate of all these parameters in the absence of prior information. Memory for storing the parameters as well as inversion of the full covariance matrix are additional problems. Factor analysis provides a compromise in which the covariance matrix is structured so that it contains fewer parameters than the full matrix but more than the diagonal form. It describes a linear subspace with a full covariance model. The data is given a probability density given by equation (6).

$$Pr(x/\omega) = Norm_x[\mu, \phi\phi^T + \Sigma] \qquad (7)$$

The covariance matrix $\phi\phi^T + \Sigma$ contains a sum of two terms where $\phi\phi^T$ describes a full covariance model over the subspace and Σ is a diagonal matrix that accounts for all remaining variations. Replacing the MoG model with a infinite sum over a continuous family of Gaussians, each of which is determined by a certain value of h. If we choose the prior over the hidden variable to be a normal we obtain the equation for factor analyser. We describe the variance in a set of directions $\phi = \{\phi_{1....}\phi_k\}$ in a high dimensional space. The different factors encode the different modes of the variation of the data set which has real world interpretations such as changes in pose or lighting.

a)



b)

Figure 10. a) Mean μ for face model. b) Diagonal covariance component Σ for face model.

Factor analyzer is an efficient model for capturing the covariance in high dimensional data. Figure 11 represents the mean μ for face model and the diagonal covariance component Σ for face model. To visualize the effect of the first factor we add or subtract ϕ_1 or a multiple of it from the mean μ.

On evaluating the learned model on the testing images for face classification using 0.5 as the threshold for the posterior we obtain the false positive rate, the false negative rate and the misclassification rate as given in Table 5. The ROC curve for the simple gaussian model is given Figure 12.

Evaluation Criteria	
False Positive Rate	0.14
False Negative Rate	0.01
Misclassification Rate	0.075

Table 5. Evaluation of Factor Analysis Model

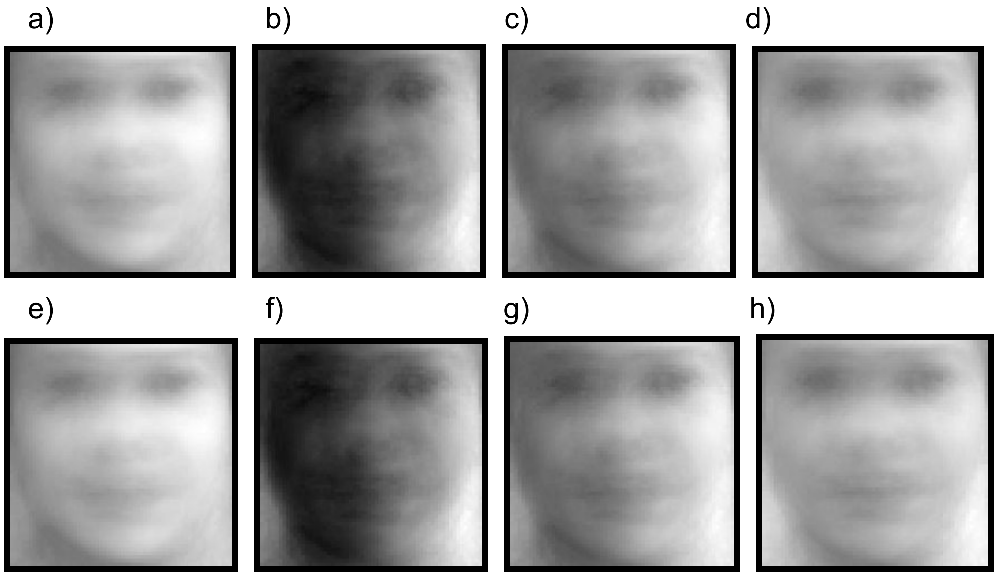


Figure 11. a-h) To visualise the effects of the factors we add and subtract multiples of the factors from the mean. We are moving along one axis of the 8D subspace that

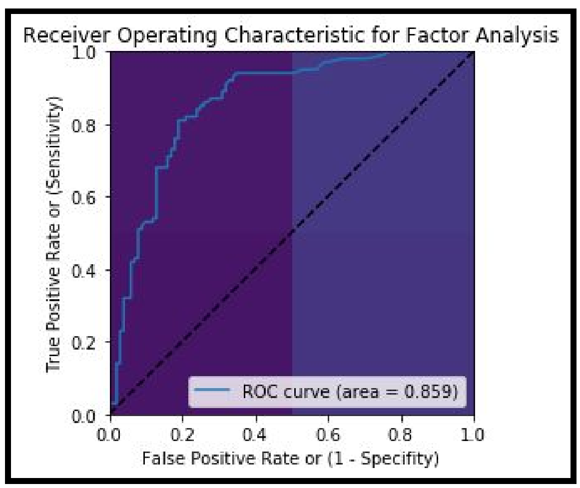


Figure 12. ROC Curve for the Factor Analysis Model

References

[1] Vuong Le, Jonathan Brandt, Lubomir Boudev, Zhe Lin, Thomas S. Huang, Interactive Facial Feature Localization, European conference on computer vision ECCV 2012.

[2] McLachlan, GJ and Peel, D. (2000) Robust mixture modelling using the t distribution. Statistics and Computing, 10 4: 335-344.

[3] Cambridge University Press 978-1-107-01179-3 - Computer Vision: Models, Learning, and Inference Simon J. D. Prince.