Devising Face Authentication System and Performance Evaluation Based on Statistical Models

Sinjini Mitra[1] *, Anthony Brockwell[1], Marios Savvides[2], Stephen E Fienberg[1]

[1]Department of Statistics, Carnegie Mellon University,

{smitra, abrock, fienberg}@stat.cmu.edu

[2] ECE Department, Carnegie Mellon University, msavvid@cs.cmu.edu

Abstract

The modern world has seen a rapid evolution of the technology of biometric authentication, prompted by an increaing urgency to ensure a system's security. The need for efficient authentication systems has skyrocketed since 9/11, and the proposed inclusion of digitized photos in passports shows the importance of biometrics in homeland security today. Based on a person's essentially unique biological traits, these methods are potentially more reliable than traditional identifiers like PINs and ID cards. This paper focuses on demonstrating the use of statistical models in devising efficient authentication systems today that are capable of handling real-life applications. First, we propose a novel Gaussian Mixture Model-based face authentication approach in the frequency domain by exploiting the well-known significance of phase in face identification and illustrate that our method is superior to the non-model based state-of-the-art

*Part of this research is supported by a grant from the Army Research Office (ARO) to CyLab, CMU.

system called the Minimum Average Correlation Energy (MACE) filter in terms of performance on a database of 65 people under extreme illumination conditions. We then introduce a general statistical framework for assessing the predictive performance of a biometric system (including watch-list detection) and show that our model-based system outperforms the MACE system in this regard as well. Finally, we demonstrate how this framework can be used to study the watch-list performance of a biometric system.

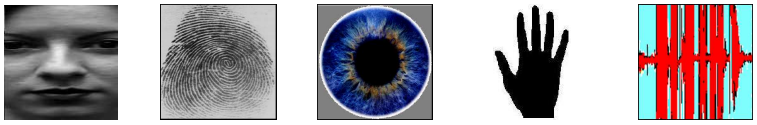
Keywords: authentication, biometrics, error rates, false alarms, frequency, Gaussian mixture model, phase, performance evaluation, random effects model, watch-list

1 Introduction

In the traditional statistical literature, the terms *biometrics* and *biometry* have been used since early in the 20th century to refer to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. Statistical methods for the analysis of data from agricultural field experiments to compare the yields of different varieties of wheat, for the analysis of data from human clinical trials evaluating the relative effectiveness of competing therapies for disease, or for the analysis of data from environmental studies on the effects of air or water pollution on the appearance of human disease in a region or country are all examples of problems that fall under the umbrella of "biometrics" as the term has been historically used.

Recently the term *biometrics* has also been used to denote the unique biological traits (physical or behavioral) of individuals that can be used for identification. *Biometric authentication* refers to the newly-emerging technology devoted to verification of a person's identity based on his/her biometrics. Typically-used biometric identifiers include face images, fingerprints, iris measurements, palm prints, hand geometry, hand veins (physical traits), and voice-print, gait and gesture (behavioral traits). They rely on "who you are" or "what you do" to make a positive personal identification and hence a biometric, in principle, cannot be lost or stolen or forgotten. It is thus inherently more reliable and more capable than knowledge-based (passwords, personal identification numbers or PINS) and token-based techniques (ID cards, drivers license) in differentiating between an authorized person and a fraudulent impostor, because many of the physiological or behavioral characteristics and traits are distinctive to each person. Some biometrics that are being popularly used today are shown in Figure 1.

Automated tools for biometric authentication are in ever increasing demand today, partly as a



face fingerprint iris scan palm-print voiceprint

Figure 1: Some popularly used biometrics.

result of efforts to improve security, especially following the deadly attacks of $9/11$. The recently adopted practice of recording photographs and fingerprints of foreign passengers at U.S. airports provides evidence towards the immense significance of biometrics in homeland security. Of all biometrics, the method of acquiring face images with the help of a digital camera is easy, non-intrusive and widely acceptable. However, while facial recognition is trivial for humans (an infant can discriminate his or her mother's face from a stranger's at the age of 45 hours (Voth, 2003)), it is an extremely challenging task to automate the process.

There are two broad approaches to devising face identification systems: (1) feature-based, and (2) model-based. Feature-based methods are more popular, and they make use of facial characteristics such as distance between eyes, nose, mouth, and their shapes and sizes (which are expected to be highly individualized) as the matching criteria. The model-based systems, on the other hand, use a statistical model to represent the pattern of some facial features (often, the ones mentioned above), and then some characteristics of the fitted model such as parameters or likelihood, are used as the matching criteria.

Although the importance of models is well-understood and has been exploited quite extensively in several aspects of image processing, such as image re-construction and segmentation, its use it devising face authentication systems has been relatively limited. Model-based approaches, such as Gaussian models (Turk and Pentland, 1991), deformable models (Yuille, 1991), and inhomogeneous Gibbs models (Liu et al., 2001) are more rigorous and flexible than feature-based ones,

having greater ability to capture the inherent variability in the data and offer greater reliability. One class of flexible statistical models is the *Mixture Models* (McLachlan and Peel, 2000), which represents complex distributions through an appropriate choice of its components to represent accurately the local areas of support of the true distribution. Apart from statistical applications, Gaussian mixture models (GMM), the most popular of the mixture models, have also been used in computer vision for modeling the shape and texture of face images (Zhu et al., 1997).

Most of the existing face recognition systems are based on spatial image intensities. Recently much research effort has focused on the frequency domain whose useful properties have been successfully exploited in many signal processing applications (Oppenheim and Schafer, 1989). The frequency domain representation of an image (the spectrum) consists of two components, the *magnitude* and the *phase*. In 2D images particularly, the phase captures more of the image intelligibility than magnitude and hence is very significant for performing image reconstruction (Hayes, 1982). Savvides et al. (2002) showed that correlation filters built in the frequency domain can be used for efficient face verification. Recently, the significance of phase has been utilized in identification problems also. Savvides and Kumar (2004) proposed correlation filters based only on phase, which performed as well as the original filters, and Savvides et al. (2004) demonstrated that performing PCA in the frequency domain using only the phase spectrum not only outperforms spatial domain PCA, but also has attractive features like illumination tolerance. These suggest that classification methods in the frequency domain, especially based on phase, may yield potentially good results. However, this has not been explored much yet, as per the authors' knowledge.

The other important component of biometric authentication is performance evaluation. Most of the face authentication systems that are developed today, are tested on databases with approximately thousands of people, which is not adequate to address bigger questions about the expected

performance of the system on large-scale real-world databases with millions of people to which the system has not been previously exposed to. This is very important for gauging the utility and validity of any biometric system for any practical application. For example, say a certain system yields a false alarm rate of 1%; this implies that a database of size $1,000,000$ will produce $10,000$ false alarms and this is quite undesirable in practice. It is known that there are about 500 million border crossings per year in the United States (one-way only), so this system will surely fail to provide a reliable means of authentication for that by resulting in a lot of innocent travelers being unnecessarily harassed and extra overhead (personnel, time) required to attend to them.

In 2002, the National Institute of Standards and Technology (NIST) carried out the *Face Recognition Vendor Test* (FRVT, NIST, 2002), where 10 commercial firms were tested on an extremely large dataset - $121,589$ facial images of $37,437$ individuals, which were henceforth unexposed to these systems. They (1) estimated the variability in performance for different groups of people, (2) characterized performance as a function of elapsed time between enrolled and new images, and (3) investigated the effect of demographics on performance. This was the first effort in the direction of an extensive performance evaluation of face authentication systems on a massive unseen database with images of diverse nature. It was an impressive undertaking with significant potential to be an useful testing protocol for all systems. But from a statistician's perspective, these are only observational studies and hence the results are at most empirical in nature - there is no statistical basis (e.g.,modeling) and scope for valid inference. Many system evaluations today are based on experiments like this, which despite being attractive, lack statistical rigor. Our goal in this paper is to propose a framework for performing such large-scale inference based on statistical models which have the potential to be more reliable in practice.

Another practical consideration of any authentication system is its performance with varying

watch-list size. A watch-list refers to the database of people who are being watched (by the FBI, for instance), that is, they are criminals who are on the "do not fly" list at the airports. The watch-list system that is currently in use matches names, and given that many individuals have same names, tends to produce a lot of false alarms. According to the Washington Post (August 20, 2004) *U.S. Sen. Edward M. "Ted" Kennedy said yesterday that he was stopped and questioned at airports on the East Coast five times in March because his name appeared on the government's secret "no-fly" list* (Goo, 2004). This shows the fragility of the present system and calls for other identifiers, such as biometrics like face, fingerprints, to be associated with the name for better and more reliable outcomes. For instance, if facial biometrics are employed in this task, a face recognition system will match an individual's face to the existing templates for the people on the watch-list for a possible identification. FRVT reported that the probability that a system correctly identifies an individual on the watch-list when presented to it usually deteriorates as the watch list size grows. Thus for effective results, they recommend that the list be kept as small as possible which is not helpful in practice.

The rest of the paper is organized as follows. Section 2 gives a brief description of the database used for the analysis and Section 3 introduces our GMM-based authentication scheme along with classification and verification results on the database at hand. Section 4 introduces an existing non model-based authentication system which will be treated as a baseline for comparing our results. The statistical framework for performance evaluation is presented in Section 5 and its application to our model-based scheme and comparison with the existing method appears in Section 6. Section 7 briefly addresses the "watch-list" problem and finally, a discussion appears in Section 8.

2 Data

The dataset used for developing our technique for facial identification is a subset of the publicly available "CMU-PIE Database" (Sim et al., 2002) which contains frontal images of 65 people under 21 different illumination conditions ranging from frontal to shadows. A small sample of images of 6 people under 3 different lighting effects is shown in Figure 2.

Figure 2: Sample images from the CMU-PIE database.

3 Gaussian Mixture Model-based System

As any continuous distribution can be approximated arbitrarily well by a finite mixture of Gaussian densities, mixture models provide a convenient semiparametric framework in which to model unknown distributional shapes. It can handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. The model framework is briefly described below.

Let $(\mathbf{Y_1}, \ldots, \mathbf{Y_n})$ be a random sample of size n where $\mathbf{Y_j}$ is a p-dimensional random vector with probability distribution $f(\mathbf{y_j})$ on $\mathcal{R}^p$, and let $\boldsymbol{\theta}$ denote a vector of the model parameters to be

estimated. A g-component mixture model can be written in parametric form as:

$$f(\mathbf{y_j}; \mathbf{\Psi}) \;=\; \sum_{i=1}^{g} \pi_i f_i(\mathbf{y_j}, \boldsymbol{\theta_i}), \tag{1}$$

where $\mathbf{\Psi} \;=\; (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\xi}^T)^T$ contains the unknown parameters and $\boldsymbol{\xi}$ is the vector of the parameters $\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_g}$ known *a priori* to be distinct. Here, $\boldsymbol{\theta_i}$ represents the model parameters for the i^{th} mixture component and $\boldsymbol{\pi} \;=\; (\pi_1, \ldots, \pi_g)^T$ is the vector of the mixing proportions with $\sum_{i=1}^{g} \pi_i = 1$. In case of Gaussian mixture models, the mixture components are multivariate Gaussian given by:

$$f(\mathbf{y_j}; \boldsymbol{\theta_i}) \;=\; \phi(\mathbf{y_j}; \boldsymbol{\mu_i}, \mathbf{\Sigma_i}) \;=\; (2\pi)^{-1} |\mathbf{\Sigma_i}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y_j} - \boldsymbol{\mu_i})^T \mathbf{\Sigma_i}^{-1} (\mathbf{y_j} - \boldsymbol{\mu_i})\} \tag{2}$$

so that the parameters in $\mathbf{\Psi}$ are the component means, variance and covariances, and the mixture model has the form:

$$f(\mathbf{y_j}; \mathbf{\Psi}) \;=\; \sum_{i=1}^{g} \pi_i \phi(\mathbf{y_j}; \boldsymbol{\mu_i}, \mathbf{\Sigma_i}). \tag{3}$$

Over the years several methods have been used to estimate mixture distributions. We use the MCMC-based Bayesian estimation method via posterior simulation (Gibbs sampler), which is now feasible and popular owing to the advent of computational power. According to Gelfand et al. (1990), the Gibbs sampler provides more refined numerical approximation for performing inference than EM. It yields a Markov chain $\{\Psi^{(k)},\ k = 1, 2, \ldots\}$ whose distribution converges to the true posterior distribution of the parameters. For our parameter estimates, we use the posterior mean, which could be estimated by the average of the first N values of the Markov chain. However, to reduce error associated with the fact that the chain takes time to converge to the correct distribution, we discard the first N_1 samples as *burn-in*. Thus our parameter estimates are

$$E\{\widehat{\mathbf{\Psi}} | \mathbf{y}\} \;=\; \sum_{k=N_1+1}^{N} \frac{\mathbf{\Psi}^{(k)}}{(N - N_1)}. \tag{4}$$

The parameter N_1 is chosen by inspection of plots of the components of the Markov chain. In particular, we choose it to be 2000 out of a total of $N = 5000$ iterations, since after this many iterations, visual inspection indicates that the chain has "settled down" into its steady-state behavior.

3.1 The Phase Model

Despite the significance of phase in face identification tasks, modeling the phase angle poses several difficulties, such as the circular or "wrapping around" property (it lies between $-\pi$ and π) and its sensitivity to distortions (such as illuminations) and transformations. This leads us to choose an alternative representation of phase for modeling purposes.

To this end, we first construct the "phase-only" images by removing the magnitude component from the frequency spectrum of the images. Since magnitude does not play as active a role in face identification, this is expected not to affect the system significantly. We then use the real and imaginary parts of these phase-only frequencies for modeling purposes. This is a simple and effective way of modeling phase, and at the same time does not suffer from the difficulties associated with direct phase modeling.

Let $R_{s,t}^{k,j}$ and $I_{s,t}^{k,j}$ respectively denote the real and the imaginary part at the $(s,t)^{th}$ frequency of the phase spectrum of the j^{th} image from the k^{th} person, $s,t = 1,2,\ldots$, $k = 1,\ldots,65$, $j = 1,\ldots,21$. We will model $(R_{s,t}^{k,j}, I_{s,t}^{k,j})$, $j = 1,\ldots,21$ as a mixture of bivariate Gaussians whose density is given by Eqn.(3), for each frequency (s,t) and each person k. We model only few low frequencies within a 50×50 grid around the origin of the spectral plane since they capture all the identifiability of any image (Lim, 1990), thus achieving considerable dimension reduction.

3.2 Classification Scheme

Classification of a new test image is done with the help of a MAP (maximum *á posteriori*) estimate based on the posterior likelihood of the data. For a new observation $Y = (R^j, I^j)$ extracted from the phase spectrum of a new image, if $f_k(\mathbf{y_j}; \mathbf{\Psi})$ denotes the GMM for person k, we can compute the likelihood under the model for person k as

$$g(Y|k) \;=\; \Pi_{\text{all freq.}} f_k(\mathbf{y_j}; \mathbf{\Psi}), \quad k = 1, \ldots, 65, \tag{5}$$

assuming independence among the frequencies. The convention is to use log-likelihoods for computational convenience in order to avoid numerical overflows/underflows in the evaluation of Equation 5. The posterior likelihood of the observed data belonging to a specific person is given by:

$$f(k|Y) \;\propto\; g(Y|k)p(k), \tag{6}$$

where $p(k)$ denotes the prior probability for each person which can be safely assumed to be uniform over all the possible people in the database. A particular image will then be assigned to class C if:

$$C = \arg\max_k f(k|Y). \tag{7}$$

3.3 Classification and Verification Results

We use $g = 2$, the components representing the illumination variations in the images of a person. A key step in the Bayesian estimation method consists of the specification of suitable priors for the unknown parameters in $\mathbf{\Psi}$. We choose conjugate priors (μ: Gaussian, Σ: Wishart, π: Dirichlet) to ensure proper posteriors and simplified computations.

Table 1 shows the classification results for our database using different number of training images. The training set in each case is randomly selected and the rest used for testing. This

selection of the training set is repeated 20 times (in order to remove selection bias) and the final errors are obtained by averaging over those from the 20 iterations. The results are fairly good,

# of Training images	# of test images	Error Rate	Standard Deviation
15	6	1.25%	0.69%
10	11	2.25%	1.12%
6	15	9.67%	2.89%

Table 1: Error rates for GMM. The standard deviations are computed over the 20 repetitions in each case.

which demonstrate that GMM is able to capture the illumination variation suitably. However, we notice that an adequate number of training images is required for the efficient estimation of the parameters; in our case, 10 is the optimal number of training images required. The associated standard errors in each case also proves the consistency of the results. Increasing the number of mixture components ($g = 3$ and $g = 4$) do not improve results significantly; hence a 2-component GMM represents the best parsimonious model in this case.

Verification is performed by imposing a threshold on the posterior likelihood of the test images, so that a person is deemed authentic if the likelihood is greater than that threshold. Figure 3 shows the ROC curve obtained by plotting the False Acceptance Rate (FAR) and False Rejection Rates (FRR) with varying thresholds on the posterior likelihood (for the optimal GMM with $g = 2$ and 10 training images). Satisfactory results are achieved with an Equal Error Rate (EER) of approximately 0.3% at a threshold value of -1700.

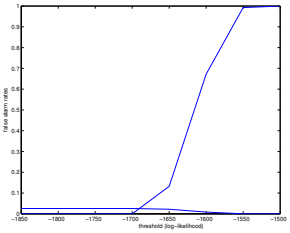


Figure 3: ROC curve for authentication based on the phase model. The lower curve is the FAR and the point of intersection of the two curves gives the EER.

4 An Existing System: The MACE Filter

The Minimum Average Correlation Energy (MACE) filter is based on a simple linear filter, easy to implement and has been reported to produce impressive results (Savvides et al., 2002). We treat this only as an example of an existing face authentication system to compare our results and point out the relative drawbacks from a statistical point of view. An interested reader is referred to Savvides et al., 2002 for more details.

Savvides et al. (2002) defines the MACE filter as:

$$\mathbf{h}_{MACE} = D^{-1}X(X^{+}D^{-1}X)^{-1}\mathbf{c}, \tag{8}$$

where X is a matrix of the vectorized 2D FFTs of the training images of a person (X^{+} denoting the conjugate transpose), D is a diagonal matrix of the average power spectrum of the training images and $\mathbf{c}$ is a column vector of ones. A filter is synthesized for each person in a database and applied to a test image via convolution. An inverse Fourier transform on the result yields the final output. If the test image belongs to an authentic person, a sharp spike occurs at the origin of the output plane indicating a match, while for an impostor, a flat surface is obtained suggesting a mismatch. A quantitative measure for authentication is the *peak-to-sidelobe ratio* (PSR), computed as PSR = $\frac{peak-mean}{\sigma}$, where $peak$ is the maximum value of the final output, and the $mean$ and the standard

deviation σ are computed from a 20×20 sidelobe region centered at the peak (excluding a 5×5 central mask). PSR values are high for authentics and considerably lower for impostors.

Figure 4 shows the MACE output for two images in the CMU-PIE database (using 3 training images per person). In both cases, the image has been been so shifted as to display the origin at the center of the plane, a convention in most engineering applications. Figure 5 shows the ROC curve obtained by plotting the FAR and the FRR for different thresholds on the PSR values for our dataset. We observe an EER of 0.9% for a threshold PSR value around 20.

PSR: 35.6599 PSR: 8.8697

(a) (b)

Figure 4: MACE filter output for an (a) authentic, and an (b) impostor.

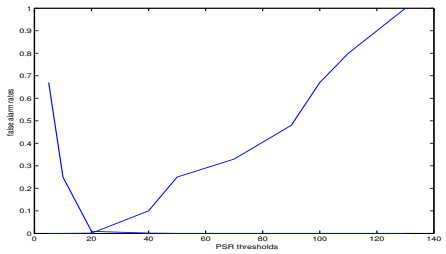


Figure 5: ROC curve for authentication. The descending curve represents the FAR and the ascending one represents the FRR. The point of intersection of the two curves is the EER.

4.1 Comparison with the Model-based Method

The MACE system is non model-based and thus suffers from some drawbacks. The success of MACE depends critically on selecting a suitable training set. A bigger N is often required to be

able to represent all possible distortions, which on the other hand, makes computations harder. So far, the choice of N has been solely based on experimental studies, and it is sensitive to the nature of the images in a database. No concrete guidelines exist to show how the number of training images affects the error rates in a given situation. Apart from this, the choice of the sidelobe dimension is based on experimentation and it is necessary to study its effect on PSR and hence on the authentication results. Similarly, no analysis has been reported so far on how the PSR values and the results vary with the nature of the images (e.g., levels of distortions, resolution). Moreover, some associated measures of the variability in the PSR estimates like standard errors and confidence intervals should be provided so as to assess their reliability.

Our authentication experiments indicate that the mixture model yields better results than the MACE system (EER=0.9% for MACE and EER=0.3% for mixture models). In applications as sensitive as authentication, even this little improvement is significant and this establishes that our approach is more efficient than the MACE system. Apart from the results themselves, our model-based method uses the posterior likelihood as the match score for the authentication procedure which is a deterministic statistical quantity having nice distributional properties (efficiency, consistency) for constructing probability intervals and hypothesis tests. This helps in assessing the reliability of these results. The MACE score PSR has no clear statistical interpretation of its own and this significantly limits its utility in inference type problems. Model-based methods also are better capable of accounting for the image variability and has a deterministic authentication process. Our method is also flexible and can be easily extended to model other distortions such as noise and expression changes by defining the mixture components to represent different levels of those. Such robustness is the primary advantage of model-based methods which is generally lacking in a non-model based framework.

One potential disadvantage of model-based techniques is that they usually require more training samples than non model-based methods. While MACE can yield satisfactory results with only 3 training images, the mixture model requires at least 10 for effective parameter estimation. So in case a sufficient number of images are not present, our model will not perform adequately. The training process is also time-consuming and is linear in the number of mixture components. However, as we have seen, in many cases we can obtain a sufficiently robust representation using as few mixture components as 2. On the other hand, the number of training images required by MACE in a given situation may vary from one dataset to another and there is no concrete way to determine the optimal number other than brute-force experimentation. For example, for a database with expression variations, a different number of training samples will be required to synthesize an effective filter than that required for images with illumination variations. For the model-based system, on the other hand, the number of training images required in a given scenario may not vary as much, and sufficiently robust models can be devised with say, 10 images in most cases. The testing process using the MAP estimate, however, is sufficiently fast can be done in real-time with no difficulties. This is what is crucial for practical application, say at an airport. This establishes that our model-based system is also useful from a practical point of view, and coupled with the statistical rigor it possesses, it proves to be much superior than any non-model based method.

5 Large-scale Performance: Random Effects Model

Traditional performance evaluation tools like linear regression models, ANOVA are *fixed effects models* and the inference obtained from them are confined only to the particular database at hand and cannot be generalized to people who do not belong to that database. This difficulty can be obvi-

ated by the use of *random effects models* (Gelfand et al., 1990) which provide a flexible framework for extending inference to a bigger population by assuming that the particular subset of subjects in the present database is a random sample from a bigger population. The regression framework allows the inclusion of any number of potential covariates representing image properties and system design parameters which are expected to influence the system performance, and quantitatively determine which of these factors significantly affect performance in a general population and to what extent. They are thus crucial for large-scale inference and also take into account the heterogeneity across individuals in their regression coefficients with the help of a probability distribution. Some particular questions of interest in this context are:

- What are the effects of certain image properties and system parameters on the score distribution in the population?
- What is the predicted score distribution for authentics and impostors in the population, based on observed image properties?
- What are the estimated error rates (both FAR and FRR) when a certain biometric system is applied to a large unknown database?

5.1 The Model Framework

Let Y_{ij} denote the outcome for the j^{th} observation on the i^{th} subject in the database, while $x_{ij}^{(m)}$ denotes the corresponding value for covariate m. We adopt the following model:

$$Y_{ij} \overset{ind.}{\sim} N(\alpha_i + \sum_{m=1}^{M} \beta_i^m x_{ij}^{(m)}, \sigma^2), \;\; i = 1, \ldots, k, \;\; j = 1, \ldots, n_i, \tag{9}$$

where M is the total number of covariates in the study. Thus the model supposes a linear dependency with homogeneous errors, but allows different slopes and intercepts for each individual. We

assume that the slope-intercept vectors are drawn from a common multivariate normal population:

$$\boldsymbol{\theta_i} \equiv \begin{pmatrix} \alpha_i \\ \beta_i^1 \\ \vdots \\ \beta_i^M \end{pmatrix} \sim MVN(\boldsymbol{\theta_0} \equiv \begin{pmatrix} \alpha_0 \\ \beta_0^1 \\ \vdots \\ \beta_0^M \end{pmatrix}, \Sigma), \quad i = 1, \ldots, k, \tag{10}$$

assuming conditional independence throughout. To effect a Bayesian analysis, we must complete the hierarchical structure at the second stage with a prior distribution for σ^2, and at the third stage with prior distributions for $\boldsymbol{\theta_0}$ and Σ. Conjugate priors for each are as follows:

$$\sigma^2 \sim IG(a, b), \;\; \boldsymbol{\theta_0} \sim N(\boldsymbol{\eta}, C), \;\; \Sigma^{-1} \sim Wishart((\rho R)^{-1}, \rho), \tag{11}$$

where R is a matrix and $\rho \geq 2$ is a scalar "degrees of freedom" parameter. All the hyperparameters in the model $a, b, \boldsymbol{\eta}, C, \rho, R$ are assumed known. The other unknown parameters are then estimated by using a Gibbs sampler to simulate from the respective full conditionals which are as follows:

$$\boldsymbol{\theta_i}|\mathbf{y}, \boldsymbol{\theta_0}, \Sigma^{-1}, \sigma^2 \sim N(D_i(\frac{1}{\sigma^2} X_i^T \mathbf{y_i} + \Sigma^{-1}\boldsymbol{\theta_0}), D_i), \;\; i = 1, \ldots, k, \;\; \text{where } D_i^{-1} = \frac{1}{\sigma^2} X_i^T X_i + \Sigma^{-1}, \tag{12}$$

$$\mathbf{y_i} = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad X_i = \begin{pmatrix} 1 & x_{i1}^1 \ldots & x_{i1}^M \\ \vdots & & \\ 1 & x_{in_i}^1 \ldots & x_{in_i}^M \end{pmatrix}. \tag{13}$$

Similarly, the full conditional for $\boldsymbol{\theta_0}$ is

$$\boldsymbol{\theta_0}|\mathbf{y}, \boldsymbol{\theta_i}, \Sigma^{-1}, \sigma^2 \sim N(V(k\Sigma^{-1}\bar{\boldsymbol{\theta}} + C^{-1}\boldsymbol{\eta}), V), \;\; \text{where } V = (k\Sigma^{-1} + C^{-1})^{-1} \text{ and } \bar{\boldsymbol{\theta}} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{\theta_i}. \tag{14}$$

The full conditional for Σ^{-1} is given by:

$$\Sigma^{-1}|\mathbf{y}, \boldsymbol{\theta_i}, \boldsymbol{\theta_0}, \sigma^2 \sim Wishart\left(\left[\sum_{i=1}^{k} (\boldsymbol{\theta_i} - \boldsymbol{\theta_0})(\boldsymbol{\theta_i} - \boldsymbol{\theta_0})^T + \rho R\right]^{-1}, \; k + \rho\right). \tag{15}$$

Finally, the full conditional for σ^2 is the updated inverse Gamma distribution:

$$\sigma^2|\mathbf{y}, \boldsymbol{\theta_i}, \boldsymbol{\theta_0}, \Sigma^{-1} \sim IG\left(\frac{n}{2} + a, \frac{1}{2}\sum_{i=1}^{k}(\mathbf{y_i} - X_i\boldsymbol{\theta_i})^T(\mathbf{y_i} - X_i\boldsymbol{\theta_i}) + b\right), \text{ where } n = \sum_{i=1}^{k} n_i. \tag{16}$$

Thus we obtain closed form full conditionals for each parameter.

We will make inferences from this model based on the marginal posteriors for the population parameters $\boldsymbol{\theta_0} = (\alpha_0, \beta_0^1, \ldots, \beta_0^M)$ and posterior predictive distributions $p(y_{ij}|\mathbf{y})$ for new data generated using Equation 9 but with the post-convergence values of the parameters. The predictive can then be estimated using a kernel density of the form

$$p(y_{ij}|\mathbf{y}) = \int p(y_{ij}|\boldsymbol{\theta_i}, \sigma^2)p(\boldsymbol{\theta_i}, \sigma^2|\mathbf{y})\partial\boldsymbol{\theta_i}\partial\sigma^2 \tag{17}$$

where g indexes the post-convergence replications from the sampler. We can use similar kernel functions to estimate the distribution of $\boldsymbol{\theta_0}$ using the parameter estimates from the Gibbs sampler after it has converged, and also compute point estimates and posterior probability intervals.

6 Application: Mixture Model-based Approach

For our model-based system, the authentication score is the posterior likelihood and we treat the logarithm of that as the response variable Y. Since the authentiction method does not involve any heuristic parameters, the only relevant covariate in this case is "authenticity" - whether the person considered is a genuine person or an impostor (denoted by X^0). There are much less number of factors involved in a model-based system which implies the use of a fairly simple model:

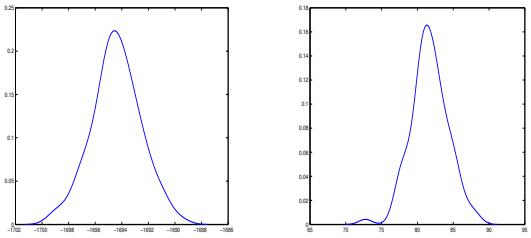
$$Y_{ij} \overset{iid}{\sim} N(\alpha_i + \beta_i x_{ij}^0, \sigma^2). \tag{18}$$

For the Gibbs sampler, we use vague hyperprior values as suggested in Gelfand et al. (1990) and trace plots reveal satisfactory convergence of the parameter chains after 2000 iterations.

The parameter estimates of the population parameters $\boldsymbol{\theta_0}$ are formed using the posterior means (after a burn-in of length 100), which appear in Table 2 along with the 95% probability intervals. α_0 denotes the mean log-likelihood value over the entire population and β_0 denotes the difference in the mean log-likelihood values between an authentic and an impostor person in the population. The estimated posterior marginals for $\boldsymbol{\theta_0}$ appear in Figure 6. Moreover, both the parameters are significantly different from zero (intervals do not include 0), hence authenticity has a statistically significant effect on the posterior likelihood in the population which is consistent with expectations.

Parameter	Estimate	Lower Limit	Upper Limit
α_0	-1694.4	-1698.0	-1690.8
β_0	81.7	76.7	86.7

Table 2: Parameter estimates and 95% probability intervals for the model-based system.

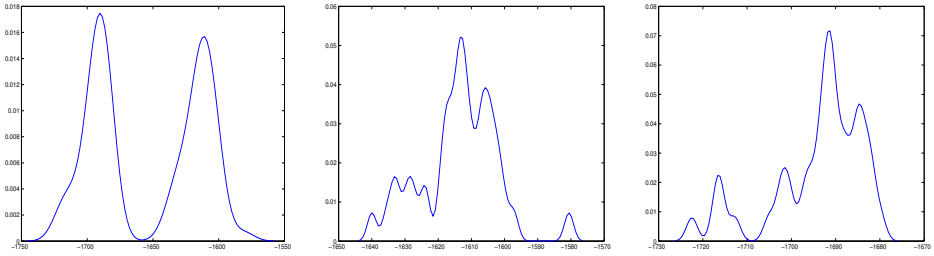


(a) α_0 (b) β_0

Figure 6: Estimated marginal posterior distributions of $\boldsymbol{\theta_0}$, using Gaussian kernel densities.

The posterior predictive distribution $p(y_{ij}|\mathbf{y})$ is computed using a Gaussian kernel and the parameter values of $\boldsymbol{\theta_i}$ and σ^2 from the post-convergence replications from the Gibbs sampler (iterations $1000 - 2000$). This is shown in Figure 7, along with separate plots for authentic and

impostor log-likelihood distributions. As can be seen, there exists a clear separation among the predicted log-likelihood values of authentic and impostor people; in fact, the distribution of log-likelihood appears to be a mixture of two distinct distributions. The amount of overlap in the tails of the authentic and impostor distributions also seem to be fairly negligible which indicates a reduced risk of false alarm.



(a) All values (b) Authentic (c) Impostor

Figure 7: Predictive posterior distribution of log-likelihood for the model-based system, using Gaussian kernel density functions.

The next step consists of estimating the predicted FAR and FRR for for predicted population. Owing to the Gaussian kernel density, it is possible to compute the exact closed-form expressions for the FAR and FRR as functions of the authentication threshold τ. This is done in the following manner. Let T denote the PSR values, and $f_A(\cdot)$ and $g_I(\cdot)$ respectively be the posterior predictive distributions of $\log(T)$ for the authentics and the impostors. If τ is the given threshold for authentication, the FAR and FRR are defined as:

$$FRR = P(T \leq \tau | T \in \text{Authentic}) = P(\log(T) \leq \log(\tau) | T \in \text{Authentic}) = \int_{-\infty}^{\tau} f_A(x)dx$$

$$FAR = P(T > \tau | T \in \text{Impostor}) = P(\log(T) > \log(\tau) | T \in \text{Impostor}) = \int_{\tau}^{\infty} g_I(y)dy \quad (19)$$

Now both f_A and g_I are Gaussian with means (μ_A, ν_I) and variances (σ_A^2, η_I^2), hence these can be

written in a closed form in terms of Φ (cumulative distribution function of standard normal) as:

$$FRR \;=\; \Phi\left(\frac{\log(\tau)-\mu_A}{\sigma_A}\right), \qquad FAR \;=\; 1-\Phi\left(\frac{\log(\tau)-\nu_I}{\eta_I}\right). \tag{20}$$

The mean and variance parameters for the two Gaussian densities for the authentic and the impostor log-likelihoods are: $\mu_A = -1610.3$, $\nu_I = -1701.2$, $\sigma_A = 19.83$ and $\eta_I = 14.7$. The resulting FAR and FRR for different selected thresholds are shown in the ROC curve in Figure 8. It has an EER value around $0.8 - 1\%$ at optimal threshold values of -1600 to -1650 for the log-likelihood function. Thus one can expect such verification results when applying the GMM-based system to a general large population, based on the results from the PIE database.

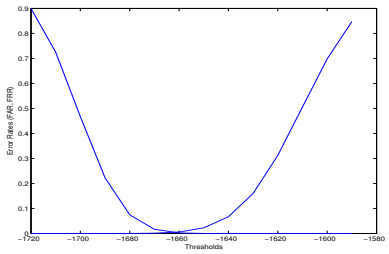


Figure 8: Predicted ROC curves for authentication using the GMM-based method. The curve on the left is the FAR (decreasing), and that on the right is the FRR (increasing).

6.1 Comparison with the MACE System

We now compare the predictive performance of the GMM-based approach with that of the MACE filter system. We apply the MACE system to the PIE database and use those results in the random effects model framework to make large-scale inference. The MACE system is a non-model based technology and hence involves more factors, which makes it necessary to include more covariates in the model to explain the variation in the score statistic called the *Peak-to-Sidelobe Ratio* (PSR, for short). For example, the PIE images have a lot of illumination variations in them, and hence

we include a covariate representing that effect. The random effects model is this case is:

$$Y_{ij} \overset{iid.}{\sim} N(\alpha_i + \beta_i x_{ij}^0 + \gamma_i x_{ij}^1, \sigma^2), \tag{21}$$

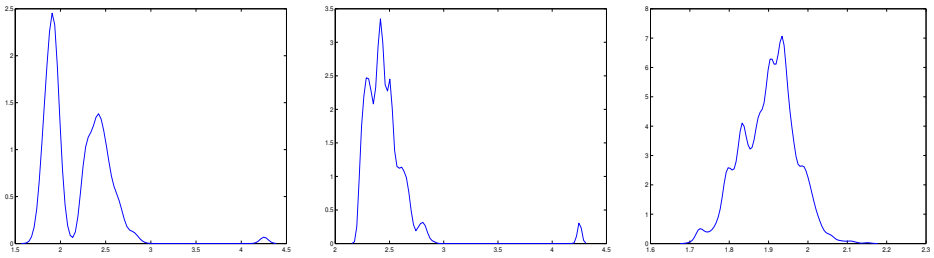
where Y is log(PSR), X^0 is the covariate representing authenticity and X^1 denotes the binary variable for the illumination level in an image (frontal balanced or shadows). We implement the model in exactly the same way as before. However, in this case, 5000 iterations were required to ensure satisfactory convergence of all the parameter chains. The parameter estimates and the respective probability intervals appear in Table 3.

Parameter	Estimate	Lower Limit	Upper Limit
α_0	1.9737	0.7504	3.1970
β_0	1.4634	1.2874	1.6394
γ_0	-0.0184	-0.1965	0.1597

Table 3: Parameter estimates and 95% probability intervals for the MACE system.

The posterior predictive distributions of log(PSR) estimated with Gaussian kernel density functions are shown in Figure 9. As can be seen clearly, the overlap between the predicted log(PSR) values of the authentics and the impostors is greater than that in the predicted log-likelihood values of the GMM-based approach (Figure 7).

Based on the kernel density estimates of the posterior predictive distributions, the parameters of the authentic and the impostor log(PSR) distributions are as: $\mu_A = 4.1331$, $\nu_I = 1.9265$, $\sigma_A = 0.6316$ and $\eta_I = 0.1471$. The resulting FAR and FRR computed according to Equation 20 are shown in Figure 10. It has an EER value around $1.2 - 1.5\%$ (higher than the GMM method) at an optimal threshold of $10 - 15$.



(a) All values (b) Authentic (c) Impostor

Figure 9: Predictive posterior distribution of log(PSR) for the MACE system, using Gaussian kernel density functions.

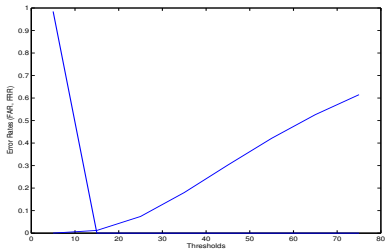


Figure 10: Predicted ROC curve for authentication using the MACE system.

7 The "Watch-list" Problem

Suppose a watch-list contains N individuals, that is, there are N stored templates in the database, one for each person on the relevant watch-list. Now if a person (randomly chosen from the general population) needs to be checked against the watch-list , his image will be tested against each of these N templates to see if a match occurs. We wish to compute the probabilities of the following two events:

- A *false match* - that is, the image matches one of the stored templates when the person tested is not actually on the watch-list.
- A *false non-match* - that is, the image does not match one of the stored templates when the person tested is actually a member of the watch-list.

Let p_0 denote the probability of a general incorrect match, and let p_1 denote the probability of a general correct match (that is, a match that should be). Then, p_0 is the FAR and p_1 is 1 - FRR. Then the probabilities of false match and false non-match for the watch-list can be computed as:

p_{FM} = Probability that an image will produce a false match with the watch-list database

= Probability that the image matches at least one of the N templates, given he is not on the list

= 1 - Probability that the image matches none of the N templates, given that he is not on the list

$= 1 - (1 - p_0)^N \approx Np_0$, if p_0 is small.

Hence this probability increases with N in a linear fashion for small p_0. Similarly,

p_{FNM} = Probability that an image will produce a false non-match with the watch-list database

= Probability that a watch-list person will not be identified

= Probability that it does not match its own template

$= N \times \frac{1}{N} \times (1 - p_1) = 1 - p_1$ (= FRR).

So this does not depend on the watch-list size in any way. This is reasonable, since regardless of the size of the watch-list database, there is only one template for each watch-list person which an image needs to be matched to. If FRR for any biometric system is small, so will be p_{FNM}. Now the probabilities of false match and non-match for a watch-list depend on the FAR and FRR and since the latter depend on a pre-specified threshold, say τ, so will the optimal watch-list probabilities. For Gaussian kernels, these probabilities of false match and mismatch can be written by using the expressions of FAR and FRR obtained in Equation 20 as:

$$p_{FM} = 1 - \left\{1 - \Phi\left(\frac{\log(\tau) - \nu_I}{\eta_I}\right)\right\}^N$$
$$p_{FNM} = 1 - \Phi\left(\frac{\log(\tau) - \mu_A}{\sigma_A}\right). \quad (22)$$

When interpreting these observations in practical terms, these results indicate that adding more terrorists to the watch-list leads to an increasing number of false positives (more travelers hassled at the airport, for instance), but at the same time, it does not affect the probability of capturing a terrorist who is on the list.

7.1 Application to the Model-based System

Figure 11 shows the variation in the probability of false match with the watch-list size, according to Equation 22. We see here that the increase in the false match error rate is lower than the linear rate and this is because p_0 or the FRR is not small enough in this case.

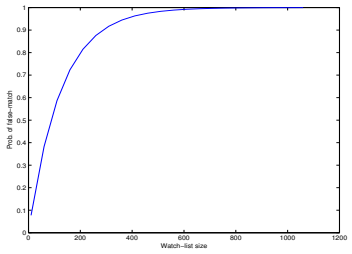


Figure 11: Variation in the probability of false match as watch-list size increases.

8 Discussion

This paper introduced a novel face identification and verification scheme based on phase and Gaussian mixture models. Although the importance of phase is well-known, this fact had not been utilized in building model-based classification techniques. This is partially because modeling phase requires an appropriate representation of its variability across different images of a person which is indeed a challenging task and our results show convincingly that our proposed model is able to handle it perfectly. Not only this, we demonstrated that our approach is tolerant to illumina-

tions; in fact, we believe that owing to its general framework, it should easily be used to model any kind of distortion, such as expression, noise, pose, by assigning different types of images to different components of mixture distributions. This proves the tremendous practical utility of this method for handling real life databases that are often subject to extraneous variations. In conclusion, harnessing the combined potential of mixture models and phase has indeed proved to be a grand success.

We then proposed a novel statistical framework based on random effects model to predict the performance of a biometric system on unknown large databases. We applied this to the MACE filter system and our model-based system, and established that the latter has a superior performance in terms of predictive performance. Development of such a rigorous evaluation protocol to assess the true potential of authentication systems in handling real-world applications is feasible only with the help of statistical models. This is the first of its kind and hence replaces the empirical and naive approaches based on observational studies that are being used currently. We could also use this method to study the "watch-list" performance of authentication systems, and estimates of the false match probabilities were obtained. This is also useful for practical applications in airports and other public places. Not only this, our method is fairly general and easily extends to other biometrics as such as fingerprints, iris, etc. as well. In conclusion, both our techniques have proved quite convincingly the significant role played by statistical modeling tools in the technology of biometric authentication in practice. Future directions include a more generalized framework for performance evaluation accounting for correlated data, application to images with other types of distortions such as expressions, and extension to other biometrics such as fingerprints, iris and multi-modal systems.

References

[1] NIST (2002). Face recognition vendor test (FRVT). http://www.frvt.org/.

[2] Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs Sampling. *Journal of the American Statistical Association*, 85(412):972–985.

[3] S.K. (2004) Goo. Sen. kennedy flagged by no-fly list. The Washington Post, Friday, August 20; Page A01.

[4] Hayes, M.H. (1982). The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier Transform. *ASSP*, 30(2):140–154.

[5] Lim, J.S. (1990). *Two-dimensional signal and image processing*. Prentice Hall, New Jersey.

[6] Liu, C., Zhu, S.C., and Shum, H.Y. (2001). Learning inhomogeneous gibbs model of faces by minimax entropy. *In Proceedings of ICCV*, pages 281–287.

[7] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons.

[8] Oppenheim, A.V. and Schafer, R.W. (1989). *Discrete-time Signal Processing*. Prentice Hall, NJ.

[9] Savvides, M., Kumar, B.V.K., and Khosla, P.K. (2004). Corefaces - robust shift invariant PCA based correlation filter for illumination tolerant face recognition. *CVPR*.

[10] Savvides, M. and Kumar, B.V.K. (2004). Eigenphases vs.eigenfaces. *In Proceedings of International Conference on Pattern Recognition (ICPR)*.

[11] Savvides, M., Vijaya Kumar, B.V.K., and Khosla, P. (2002). Face verification using correlation filters. In *3rd IEEE Automatic Identification Advanced Technologies*, pages 56–61, Tarrytown, NY.

[12] Sim, T., Baker, S., and Bsat, M. (2002). The CMU pose, illumination, and expression (PIE) database. In *Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition*.

[13] Turk, M.A. and Pentland, A.P. (1991). Face recognition using eigenfaces. *In Proceedings of Computer Vision and Pattern Recognition (CVPR)*.

[14] Voth, D. (2003). In the news: face recognition technology. IEEE magazine on intelligent systems, May-June. Vol. 18, Issue 3, pp. 4-7.

[15] Yuille, A. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1).

[16] Zhu, S., Wu, Y., and Mumford, D. (1997). Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8).