Towards a Measure of Biometric Feature Information

Andy Adler[1], *Richard Youmaran[2], Sergey Loyka[2]

[1]Carleton University, Ottawa, Ontario,K1S 5B6, Canada

[2]University of Ottawa, Ontario, Canada

Tel: $(613)520-2600$ ext. 8785

Fax: $(613)520-5727$

adler@sce.carleton.ca

{youmaran,loyka}@site.uottawa.ca

Abstract

This paper develops an approach to measure the information content of a biometric feature representation. We define biometric information as the decrease in uncertainty about the identity of a person due to a set of biometric measurements. We then show that the biometric feature information for a person may be calculated by the relative entropy $D(p\|q)$ between the population feature distribution q and the person's feature distribution p. The biometric information for a system is the mean $D(p\|q)$ for all persons in the population. In order to practically measure $D(p\|q)$ with limited data samples, we introduce an algorithm which regularizes a Gaussian model of the feature covariances. An example of this method is shown for PCA and Fisher linear discriminant (FLD) based face recognition, with biometric feature information calculated to be 45.0 bits (PCA), 37.0 bits (FLD) and 55.6 bits (fusion of PCA and FLD features). Finally, we discuss general applications of this measure.

Index Terms

Biometric features, Relative entropy, Face recognition, Information content

EDICS: BIO-THEO, WAT-THEO

I. INTRODUCTION

How much information is there in a face, or a fingerprint? This question is related to many issues in biometric technology. For example, one of the most common biometric questions is that of uniqueness, eg. to what extent are fingerprints unique? From the point of view of identifiability, one may be interested in how much identifying information is available from a given technology, such as video surveillance. In the context of biometric fusion [19] one would like to be able to quantify the biometric information in each system individually, and the potential gain from fusing the systems. Additionally, such a measure is relevant to biometric cryptosystems and privacy measures. Several authors have presented approaches relevant to this question. For example, Wayman et al. [23] introduced a set of statistical approaches to measure the separability of Gaussian feature distributions using a "cotton ball model". Another approach was developed by Daugman [7] to measure the information content of iris images based on the discrimination entropy [6], calculated directly from the match score distributions. Also, Golfarelli et al. [11] showed that the most commonly used feature representations of handgeometry and face biometrics have a limited number of distinguishable patterns, on the order of 10^5 and 10^3, respectively, as measured by a theoretical estimate of the equal error rate. Other authors have used information theoretic approaches, such as the approach of Ross and Jain [19] to biometric fusion. However, none of these methods approach measurement of information content of biometric data from an information theoretic point of view.

In this paper we elaborate an approach to address this question based on definitions from information theory [2]. We define the term "biometric information" as follows:

biometric information: the decrease in uncertainty about the identity of a person due to a set of biometric features measurements

In order to interpret this definition, we refer to two instants: 1) before a biometric measurement, t_0, at which time we only know a person p is part of a population q, which may be the whole planet; and 2) after receiving a set of measurements, t_1, we have more information and less uncertainty about the person's identity.

In order to motivate our approach, we initially consider the properties that such a measure should have. Consider a soft biometric system which measures height and weight; furthermore,

assume all humans are uniformly and independently distributed in height between 100–200 cm and weight between 100–200 lb. If a person's features were completely stable and could be measured with infinite accuracy, people could be uniquely identified from these measurements, and the biometric features could be considered to yield infinite information. However, in reality, repeated biometric measurements give different results due to measurement inaccuracies, and to short- and long-term changes in the biometric features themselves. If this variability results in an uncertainty of $\pm$ 5 cm and ± 5 lb, one simple model would be to round each measure to $105, 115, ..., 195$. In this case, there are 10×10 equiprobable outcomes, and an information content of $log_2(100) = 6.6$ bits.

Such an analysis is intrinsically tied to a choice of biometric features. Thus, our approach does not allow us to answer "how much information is in a fingerprint?", but only "how much information is in the position and angle data of fingerprint minutiae?". Furthermore, for many biometrics, it is not clear what the underlying features are. Face images, for example, can be described by image basis features or landmark based features. To overcome this, we may choose to calculate the information in all possible features. In the example, we may provide height in inches as well as cm; however, in this case, a good measure of information must not increase with such redundant data.

Based on the definition of introduced, this paper develops a mathematical framework to measure biometric feature information for a given system and set of biometric features. In practice, there are limited numbers of samples of each person, which makes our measure ill-conditioned. In order to address this issue, we develop a stable algorithm based on a distribution modeling and regularization. We then use this algorithm to analyze the biometric information content of two different face recognition algorithms.

II. METHODS

In this section we develop an algorithm to calculate biometric information based on a set of features, using the relative entropy measure [6]. We explain our method in the following steps: 1) measure requirements, 2) relative entropy of biometric features, 3) Gaussian models for biometric features and relative entropy calculations, 4) regularization methods for degenerate features, 5) regularization methods for insufficient data.

A. *Requirements for biometric feature information*

In order to elaborate the requirements that a good measure of biometric feature information must have, we consider system that measures height and weight. These values differ within the global population, but also vary for a given individual, both due to variations in the features themselves and to measurement inaccuracies. We now wish to consider the properties a measure of biometric feature information should have:

1) If an intra-person distribution p is exactly equal to the inter-person q distribution, then there is no information to distinguish a person, and biometric feature information is zero.
2) As the feature measurement becomes more accurate (less variability), then it is easier to distinguish someone in the population and the biometric information increases (this criterion may appear counterintuitive, but a biometric system is being evaluated, not the raw feature values).
3) If a person has unusual feature values (i.e. far from the population mean), they become more distinguishable, and their biometric feature information will be larger.
4) The biometric information of uncorrelated features should be the sum of the biometric information of each individual feature.
5) Features that are unrelated to identity should not increase biometric information. For example, if a biometric system accurately measured the direction a person was facing, information on identity would be unchanged.
6) Correlated features such as height and weight are less informative. In an extreme example consider the height in inches and in cm. Clearly, these two features are no more informative than a single value.

Based on this definition, the most appropriate information theoretic measure for the biometric feature information is the relative entropy ($D(p\|q)$) [6] between the intra- ($p(\mathbf{x})$) and inter-person ($q(\mathbf{x})$) biometric feature distributions. $D(p\|q)$, or the Kullback-Leibler distance, is defined to be the "extra bits" of information needed to represent $p(\mathbf{x})$ with respect to $q(\mathbf{x})$. $D(p\|q)$ is defined to be

$$D(p\|q) = \int_{\mathbf{x}} p(\mathbf{x}) log_2 \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \tag{1}$$

where the integral is over all feature dimensions, $\mathbf{x}$. $p(\mathbf{x})$ is the probability mass function or distribution of features of an individual and $q(\mathbf{x})$ is the overall population distribution. A comment

on notation: we use p to refer to both an individual person, and the distribution of the person's features, while q represents the population and the distribution of its features.

This measure can be motivated as follows: the relative entropy, $D(p\|q)$, is the extra information required to describe a distribution $p(\mathbf{x})$ based on an assumed distribution $q(\mathbf{x})$ [6]. $D(p\|q)$ differs from the entropy, $H(p)$, which is the information required, on average, to describe features $\mathbf{x}$ distributed as $p(\mathbf{x})$. H is not in itself an appropriate measure for biometric feature information, since it does not account the extent to which each feature can identify a person p in a population q. An example of a feature unrelated to identity is the direction a person is facing. Measuring this quantity will increase H of a feature set, but not increase its ability to identify a person. The measure $D(p\|q)$ corresponds to the requirements: given a knowledge of the population feature distribution q, the information in a biometric feature set allows us to describe a particular person p.

B. Distribution modeling

In a generic biometric system, F biometric features are measured, to create a biometric feature vector $\mathbf{x}$ ($F\times 1$) for each person. For person p, we have N_p features samples, while we have N_q samples for the population. For convenience of notation, we sort p's measurements to be the first grouping of the population. Defining $\mathbf{x}$ as an instance of random variable X, we calculate the population feature mean μ_q

$$\mu_q = \underset{q}{E}[X] = \frac{1}{N_q}\sum_{i=1}^{N_q}\mathbf{x}_i \tag{2}$$

where the feature mean of person p, μ_p, is defined analogously, replacing q by p. The population feature covariance $\mathbf{\Sigma}_q$ is

$$\mathbf{\Sigma}_q = \underset{q}{E}\left[(X-\mu_q)^t(X-\mu_q)\right] = \frac{1}{N_q-1}\sum_{i=1}^{N_q}(\mathbf{x}_i-\mu_q)^t(\mathbf{x}_i-\mu_q) \tag{3}$$

where the feature correlation of person p. The individuals feature convariance, $\mathbf{\Sigma}_p$, is again defined analogously.

Features are calculated from a set of N_q images using different component analysis methods such as Principal Component Analysis (PCA, also referred to as Eigenface features) [12][21] and Fisher linear discriminant (FLD) [13]. μ_p and μ_q are $F\times 1$ vectors of the population and individual

mean distributions, while $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_q$ are $F \times F$ matrices of the individual and population covariance matrices.

One important general difficulty with direct information theoretic measures is that of data availability. Distributions are difficult to estimate accurately, especially at the tails; and yet $log_2\left(p(\mathbf{x})/q(\mathbf{x})\right)$ will give large absolute values for small $p(\mathbf{x})$ or $q(\mathbf{x})$. Instead, it is typical to fit data to a model with a small number of parameters. The Gaussian distribution is the most common model; it is often a good reflection of the real world distributions, and is analytically solvable in entropy integrals. Another important property of the Gaussian is that it gives the maximum entropy for a given standard deviation, allowing such models to be used to give an upper bound to entropy values. Based on the Gaussian model, which seems to be the simplest and appropriate for p and q, we write:

$$p(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}_p|}} exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_p\right)^t {\mathbf{\Sigma}_p}^{-1}\left(\mathbf{x}-\mu_p\right)\right) \tag{4}$$

$$q(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}_q|}} exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_q\right)^t {\mathbf{\Sigma}_q}^{-1}\left(\mathbf{x}-\mu_q\right)\right) \tag{5}$$

From which we can calculate $D(p\|q)$.

$$D(p\|q) = \int p(\mathbf{x})\left(log_2 p(\mathbf{x}) - log_2 q(\mathbf{x})\right) d\mathbf{x} \tag{6}$$

$$= -k\left(ln|2\pi\mathbf{\Sigma}_p| - ln|2\pi\mathbf{\Sigma}_q| + 1 - \underset{p}{E}\left[\left(\mathbf{x}-\mu_q\right)^t {\mathbf{\Sigma}_q}^{-1}\left(\mathbf{x}-\mu_q\right)\right]\right) \tag{7}$$

$$= k\left(ln\frac{|2\pi\mathbf{\Sigma}_q|}{|2\pi\mathbf{\Sigma}_p|} + trace\left((\mathbf{\Sigma}_p+\mathbf{T}){\mathbf{\Sigma}_q}^{-1} - \mathbf{I}\right)\right) \tag{8}$$

where $\mathbf{T} = (\mu_p - \mu_q)^t(\mu_p - \mu_q)$ and $k = log_2\sqrt{e}$.

This expression calculates the relative entropy in bits for Gaussian distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. This expression corresponds to most of the desired requirements for a biometric feature information measure introduced in the previous section:

1) If person's feature distribution matches the population, p = q; this yields $D(p\|q) = 0$, as required.
2) As feature measurements improve, the covariance values, $\mathbf{\Sigma}_p$, will decrease, resulting in a reduction in $|\mathbf{\Sigma}_p|$, and an increase in $D(p\|q)$.
3) If a person has feature values far from the population mean, $\mathbf{T}$ will be larger, resulting in an larger value of $D(p\|q)$.

4) Combinations of uncorrelated feature vectors yield the sum of the individual $D(p\|q)$ measures. Thus, for uncorrelated features f_1 and f_2, where $\{f_1, f_2\}$ represents concatenation of the feature vectors, $D(p(f_1)\|q(f_1)) + D(p(f_2)\|q(f_2)) = D(p(\{f_1, f_2\})\|q(\{f_1, f_2\}))$
5) Addition of features uncorrelated to identity will not change $D(p\|q)$. Such a feature will have an identical distribution in p and q. If U is the set of such uncorrelated features, $[\mathbf{\Sigma}_p]_{ij} = [\mathbf{\Sigma}_q]_{ij} = 0$ for i or $j \in U$, and $i \neq j$, while $[\mathbf{\Sigma}_p]_{ii} = [\mathbf{\Sigma}_q]_{ii}$ and $[\mu_q]_i = [\mu_p]_i$. Under these conditions, $D(p\|q)$ will be identical to its value when excluding the features in U. One way to understand this criterion is that if the distributions for q and p differ for features in U, then those features can be used as a biometric to help identify a person.
6) Correlated features are less informative than uncorrelated ones. Such features will decrease the condition number (and thus the determinant) of both $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_q$. This will decrease the accuracy of the measure $D(p\|q)$. In the extreme case of perfectly correlated features, $\mathbf{\Sigma}_p$ becomes singular with a zero determinant and $D(p\|q)$ is undefined. Thus, our measure is inadequate in this case. In the next section, we develop an algorithm to deal with this effect.

C. Regularization Methods for degenerate features

In order to guard against numerical instability in our measures, we wish to extract a mutually independent set of G "important" features ($G \leq F$). To do this, we use the principal component analysis (PCA) [10][12] to generate a mapping ($\mathbf{U}^t : X \rightarrow Y$), from the original biometric features X ($F \times 1$) to a new feature space Y of size $G \times 1$. The PCA may be calculated from a Singular Value Decomposition (SVD) [3] of the feature covariance matrix, such that

$$\mathbf{U}\mathbf{S}_q\mathbf{U}^t = svd(cov(X)) = svd(\mathbf{\Sigma}_q) \tag{9}$$

Since $\mathbf{\Sigma}_q$ is positive definite, $\mathbf{U}$ is orthonormal and $\mathbf{S}_q$ is diagonal. We choose to perform the PCA on the population distribution q, rather than p, since q is based on far more data, and is therefore likely to be a more reliable estimate. The values of $\mathbf{S}_q$ indicate the significance of each feature in PCA space. A feature j, with small $[\mathbf{S}_q]_{j,j}$ will have very little effect on the overall biometric feature information. We use this analysis, in order to regularize $\mathbf{\Sigma}_q$, and to reject degenerate features by truncating the SVD. We select a truncation threshold of j where $[\mathbf{S}_q]_{j,j} < 10^{-10}[\mathbf{S}_q]_{1,1}$. Based on this threshold, $\mathbf{S}_q$ is truncated to be $G \times G$, and $\mathbf{U}$ is truncated

to $F \times G$. Using the basis $\mathbf{U}$ calculated from the population, we decompose the individual's covariance into feature space Y:

$$\mathbf{S}_p = \mathbf{U}^t \mathbf{\Sigma}_p \mathbf{U} \tag{10}$$

where $\mathbf{S}_p$ is not necessarily a diagonal matrix. However, since p and q describe somewhat similar data, we expect $\mathbf{S}_p$ to have a strong diagonal component, as seen in Fig. 4.

Based on this regularization scheme, (8) may be rewritten in the PCA space as:

$$D(p\|q) = k\left(\beta + trace\ \mathbf{U}\left((\mathbf{S}_p + \mathbf{S}_t)\mathbf{S}_q{}^{-1} - \mathbf{I}\right)\mathbf{U}^t\right) \tag{11}$$

where $\beta = ln\frac{|\mathbf{S}_q|}{|\mathbf{S}_p|}$ and $\mathbf{S}_t = \mathbf{U}^t\mathbf{T}\mathbf{U}$

D. Regularization Methods for insufficient data

The expression developed in the previous section solves the problem of ill-poseness of $\mathbf{\Sigma}_q$. However, $\mathbf{\Sigma}_p$ may still be singular in the common circumstance in which only a small number of samples of each individual are available. Given N_p images of an individual from which G features are calculated, $\mathbf{\Sigma}_p$ will be singular if $G \geq N_p$, which will result in $D(p\|q)$ diverging to ∞. In practice, this is a common occurrence, since most biometric systems calculate many hundreds of features, and there are only rarely more then ten of samples of each person. In order to address this issue, we develop an estimate which may act as a lower bound. In order to do this, we make the following assumptions:

1) Estimates of feature variances are valid $[\mathbf{S}_p]_{i,i}$ for all i.
2) Estimates of feature covariances $[\mathbf{S}_p]_{i,j}$ for $i \neq j$ are only valid for the most important L features, where $L < N_p$.

Features which are not considered valid based on these assumptions, are set to zero by multiplying $\mathbf{S}_q$ by a mask $\mathbf{M}$, where

$$M = \begin{cases} 1, & \text{if } i = j \text{ or } (i < L \text{ and } j < L); \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Using (12), $[\mathbf{S}_p]_{i,j} = (\mathbf{M}_{i,j})[\mathbf{U}^t\mathbf{\Sigma}_p\mathbf{U}]_{i,j}$.

This expression regularizes the intra-person covariance, $\mathbf{\Sigma}_p$, and assures that $D(p\|q)$ does not diverge. To clarify the effect of this regularization on $D(p\|q)$, we note that intra-feature

covariances will decrease $|\mathbf{\Sigma}_p|$ toward zero, leading a differential entropy estimate diverging to ∞. We thus consider this regularization strategy to generate a lower bound on the biometric feature information. The selection of L is a compromise between using all available measurements (by using large L) and avoiding numerical instability when $\mathbf{S}_p$ is close to singular (by using small L).

E. Average information of a biometric system

This section has developed a measure of biometric feature information content of a biometric feature representation of a single individual with respect to the feature distribution of the population. As discussed, the biometric feature information will vary between people; those with feature values further from the mean have larger biometric feature information. In order to use this approach to measure the biometric feature information content of a biometric system, we calculate the average biometric feature information for each individual in the population (weighted by the probability of needing to identify that person, if appropriate).

III. FACE RECOGNITION

Fig. 1. An example of PCA (Eigenface) face features. From left to right, PCA features number 2, 5, 25, 60 are shown. The PCA features are othonormal and fit the data in a least squares sense.

Information in a feature representation of faces is calculated using our described method for different individuals. In order to test our algorithm, it is necessary to have multiple images of the same individual.For this reason, using the Aberdeen face database [4], we chose 18 frontal images of 16 persons, from which we calculate the PCA (eigenface) features using the algorithm of [12] and the FLD face features components using the algorithm described in [24]. Initially, all face images were registered by rotation and scaling to have eye positions at $(50, 90)$ and

Fig. 2. An example of FLD face features. From left to right, FLD features number 2, 5, 25, 60 are shown. FLD attempts to maximize class separation while minimizing the within class scatter.

$(100, 90)$. Images were then cropped to 150×200 pixels and histogram equalized to cover the intensity range 0–255. The same set of operations is applied to all images using the same thresholds. This results with the same effect on all images when computing the biometric feature information.

The feature decomposition process was conducted on 18 images of each of 16 persons, giving 288 total images. For PCA and Fisher feature decompositions, 288 seperate vectors were computed, and the most significant 100 features used for subsequent analysis. Fig. 1 and Fig. 2 illustrate PCA and FLD features, respectively. From this, $D(p\|q)$ is computed for each of 16 persons using (11), which assumes that p and q have Gaussian distributions. In order to test the validity of the Gaussian model for our data, we use the following normality tests:

- Kolmogorov-Smirnov test: compares the distributions of values in the two data vectors X_1 and X_2, where X_1 represents random samples from the underlying distribution and X_2 follows an ideal Gaussian with zero mean and variance. The null hypothesis is that X_1 and X_2 are drawn from the same continuous normal distribution. We reject the null hypothesis at $p < 0.01$.
- The Lilliefors test [5]: evaluates the hypothesis that $\mathbf{x}$ has a normal distribution with unspecified mean and variance, against the alternative that $\mathbf{x}$ does not have a normal distribution. This test compares the empirical distribution of X with a normal distribution having the same mean and variance as X. We reject the null hypothesis at $p < 0.01$.

Using these tests, on average 89% of the marginal distribution of all the FLD and PCA computed features is normally distributed.

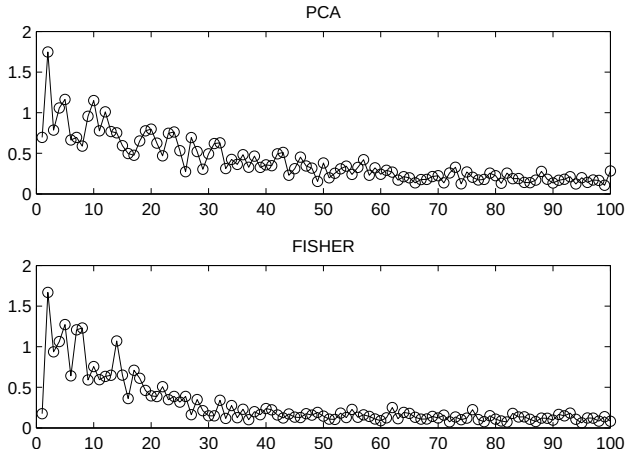

Fig. 3. Biometric information as a function of feature number (circles) for (A) PCA (Eigenface) and (B) FLD (bottom) face feature decomposition.

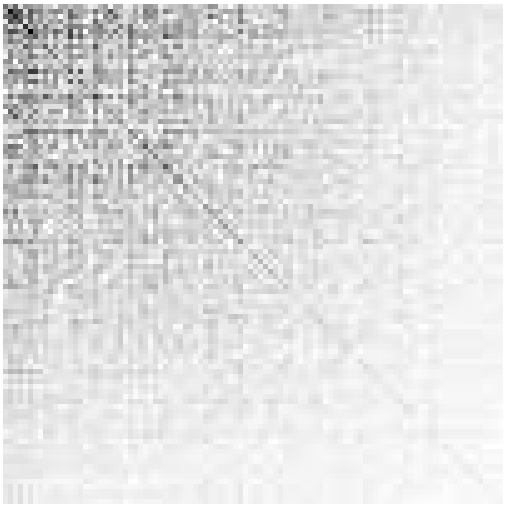

Fig. 4. The regularized intra-person covariance matrix $\mathbf{S}_p$ showing dominant components along its diagonal. Since $\mathbf{\Sigma}_p$ represents similar information to $\mathbf{\Sigma}_q$ it is reasonable to expect the matrices have similar eigenvectors, resulting in strong diagonal components in $\mathbf{\Sigma}_p$.

A. Biometric information calculations

After fitting the distributions of $p(\mathbf{x})$ and $q(\mathbf{x})$ to a Gaussian model, we initially analyse the biometric feature information in each PCA and Fisher feature separately. PCA features are shown in Fig. 3, and show a gradual decrease from an initial peak at feature 2. The form of the curve can be understood from the nature of the PCA decomposition, which tends to

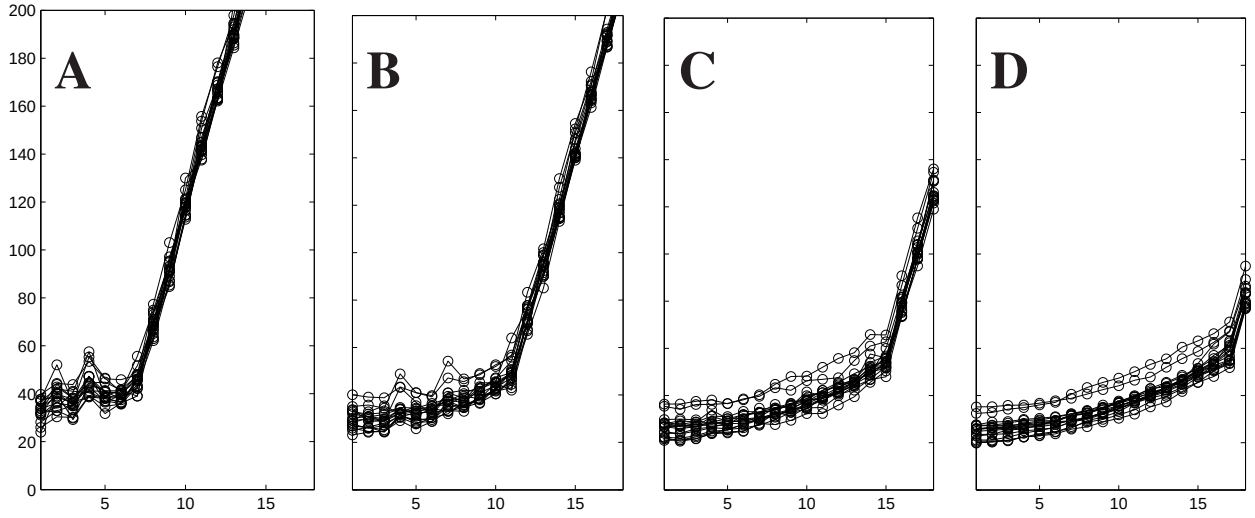

Fig. 5. Biometric information (in bits) (y-axis) vs. the mask size (L) (x-axis) for each person. Each subfigure represents a different value of N_p (images of the same person): (A) 8, (B) 12, (C) 16 and (D) 18. The curves show that $D(p\|q)$ diverges as Σ_p becomes singular ($L \geq N_p$). The relative entropy increases with the size of the mask.

place higher frequency details in higher number features. Since noise tends to increase with frequency, the biometric information in these higher numbered PCA features will be less. A sum of biometric feature information over the first 100 PCA features gives 40.5 bits. This does not assume statistical independence nor uncorrelatedness of PCA coefficients. Biometric feature information calculated using FLD features seems to be similar to PCA features such that most biometric feature information is computed for the most dominant fisherfaces.

In order to calculate $D(p\|q)$ for all features, we are limited by the available information. Since $N_p = 18$ images are used to calculate the covariances, attempts to calculate $D(p\|q)$ for more than 17 features will fail, because $\mathbf{\Sigma}_p$ is singular. This effect is seen in the condition number (ratio of the largest to the smallest singular value) which was 4.82×10^3 for $\mathbf{S}_q$ and 1.32×10^{20} for $\mathbf{S}_p$. The relatively small condition number of $\mathbf{S}_q$ indicates that no features are degenerate for PCA and FLD face recognition features. However, $\mathbf{S}_p$ is severely ill-conditioned. To overcome this ill-conditioning, we introduced a regularization scheme based on a mask (equation 12) with a cut-off point L. This scheme is motivated by the diagonal structure of $\mathbf{S}_p$, as shown in Fig. 4. To ensure convergence, the mask size L is set to a value smaller than N_p.

We solve this singularity of (11) using a mask for $\mathbf{S}_p$ based on a parameter L. To further

explore the effect of parameters L and N_p, we artificially reduce the N_p by randomly eliminating some images from individuals. Results for $D(p\|q)$ for PCA features for each person as a function of L are shown in Fig. 5 for N_p = 8, 12, 16 and 18. In these curves, we observe a "hockey stick" shape. The relative entropy measure remains stable when $L < N_p$, but if $L \geq N_p$, we observe a dramatic increase in $D(p\|q)$ as the algorithm approaches a singularity of $\mathbf{\Sigma}_p$ and the ill-conditioning of $\mathbf{\Sigma}_q$. When $L < N_p$, $D(p\|q)$ is stable with a lower and upper bounds between 35 to 50 bits. However, when $L \geq N_p$, $D(p\|q)$ estimates start diverging and reach very large values.

Clearly, points for L greater than the knee in the hockey stick do not represent accurate estimates of $D(p\|q)$. We also argue that when L approaches N_p, the inherent ill-conditioning of $\mathbf{\Sigma}_p$ makes the our algorithm over-estimate $D(p\|q)$. On the other hand, small values of L will under-estimate $D(p\|q)$, since these values will mask inter-feature correlations. This effect increases $|\mathbf{S}_p|$ as L decreases. However, the results suggest that this effect is minor, especially in Fig. 4A and 4B, where the "base" of the hockey stick is more flat. In order to produce an unique and stable estimate for $D(p\|q)$, it is necessary to choose a compromise between these effects. We recommend choosing $L = \frac{3}{4}N_p$, since a larger value of L puts the estimate in an unstable region of Fig. 4.

Using this algorithm and value of L, we calculate the overall biometric feature information for different face recognition algorithms. For PCA features, the average $D(p\|q)$ is 45.0 bits, and for FLD features $D(p\|q)$ is 37.0 bits. If PCA and FLD features are combined (making 200 features in all), average $D(p\|q)$ is 55.6 bits. This combination of features illustrates that a biometric fusion of similar features may offer very little information above that of the individual underlying features. It is intially somewhat surprising that FLD feature information is measured to be lower than that from PCA. This results may be understood becuase PCA features retain unwanted information due to variations in facial expression and lighting, while FLD "projects away" variations in lighting and facial expression while maintaining the discriminant features. . In addition, feature decomposition using independent component analysis (ICA) [10] was also conducted on the same set of faces. ICA has the advantage that it does not only decorrelate the signals but also reduces higher-order statistical dependencies in order to make the signals as statistically independent as possible [14]. Since ICA maximizes non-gaussianity, it fits less well to the assumptions of our model. For ICA features, an average of 39.0 bits was computed for

$D(p\|q)$.

IV. Discussion

This paper has introduced a definition of biometric feature information and an algorithm to measure it from a set of population and individual biometric features, as measured by a biometric algorithm under test. Biometric information is defined in terms of the reduction in uncertainty of the identity of a person resulting from a set of biometric feature measurements. Based on this definition, we show that this concept matches the information theoretic concept of *relative entropy* $D(p\|q)$, where p is the probability distribution of the persons's features, and q is the distribution of features of the population. Examples of its application were shown for two different face recognition algorithms based on PCA (Eigenface) and FLD feature decompositions. Clearly, the framework developed in this paper depends on accurate estimates of the population distributions q. Developing a good estimate of the "world model" is known to be a hard problem; in this work, we use the typical approach of assuming our database is an adequate representation of the population.

The result of our calculations (approximately 40 bits per face) is reasonable when compared with previous analysis of face recognition accuracy. From the FRVT results, we extrapolate the gallery size for an identification rate of 0.5 ([17], [18]). This is taken to be a rough model of the population for which the algorithm can reduce the identity uncertainty to 50%. For the top three algorithms, the gallery sizes were 1.67×10^8, 3.53×10^7, and 2.33×10^6, corresponding to 27.3, 25.1, and 21.2 bits. This value is over half that calculated here, and is reasonable since the FRVT database appears to be significantly more difficult than the one used here [4], and current face recognition algorithms are not yet considered to be close to optimal since they seem to use approximately 1/2 to 2/3 of the available feature information. Similarly, our work on biometric encryption [1] seemed to suggest an upper limit of approximately 20 bits of key into face images, using the algorithm of [20].

As an exploration of the implications of this work, an analogy can be made between a biometric system and a traditional communication system in terms of information capacity [6]. The signal source transmits one symbol from an alphabet; this corresponds to one person from a population to be identified. The symbol is encoded and sent across a channel and is subject to channel noise; similarly, biometric features from a person are measured, and are subject to variability due to

noise in the measurement system and to inherent feature variability. Thus the biometric feature measurement system corresponds to the communication channel. The communications system receiver detects a signal and must decide which symbol was sent, corresponding to the role of the biometrics identification process. In this context, $D(p\|q)$ is the differential information of a single signal, and the average $D(p\|q)$, weighted by the probability of each signal p, is the channel capacity. Based on this analogy, we can say that biometric feature information is the channel capacity of a biometric measurement system.

In a general biometric system, the following issues associated with biometric features must be considered:

- Feature distributions vary. Features, such as minutiae ridge angles may be uniformly distributed over 0–2π, while other features may be better modeled as Gaussian. In this paper, all features are modeled as Gaussian. This is valid model for most PCA and FLD features, but is not valid for any ICA features (since ICA is designed to maximize non-Gaussianity). On the other hand, a Gaussian model may be considered to estimate an upper bound for the entropy.
- Raw sample images need to be processed by alignment and scaling before features can be measured. Any variability in registration will dramatically increase the variability in measured features and decrease the biometric feature information measure.
- Feature dimensionality may not be constant. For example, the number of available minutiae points varies. The method presented in this paper does not address this issue, since the dimensions of $p(\mathbf{x})$ and $q(\mathbf{x})$ must be the same. Generalized Entropy measures exist which may allow an extension of this approach to non-constant dimensional features.

It is interesting to note that the biometric entropy is larger for some faces. Fig. 5 shows a range of biometric information (from 32 to 47 bits) for different individuals, which may help explain why some people are potentially easier to recognize than others. This is perhaps some evidence for the "biometrics zoo" hypothesis [8]. In order to explore this effect, we plot the biometric feature information as a function of average feature variance for each person (Fig. 6). A significant correlation ($p < .01$) is detected, indicating that features are less variable in those subjects with higher biometric feature information.

While we have introduced a measure in the context of face recognition, we anticipate that such a measure may help address many questions in biometrics technology, such as the following:

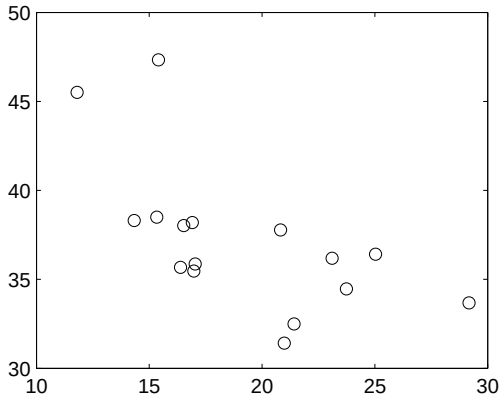

Fig. 6. Average $D(p\|q)$ in bits (y-axis) as a function of the mean feature variance (arbitrary units) (x-axis) for 16 different persons. The mean feature variance is computed by summing all the diagonal components of $\mathbf{S}_p$ matrix for each person. The correlation coefficient is -0.62, which is significant at $p < 0.01$.

- Uniqueness of biometric features: A common question is "are fingerprints really unique?". While Pankanti et al. [16] have recently provided a sophisticated analysis of this problem based on biometric feature distributions directly, a general approach based on information content would help address this question for other biometric modalities.
- Inherent limits to biometric template size requirements. A maximum compression of biometric features will be limited to the biometric feature information. This theoretical lower limit may be of use for ID card applications with limited data density.
- Feasibility of biometric encryption: Proposed biometric encryption systems use biometric data to generate keys [22], and thus the availability of biometric feature information limits the security of cryptographic key generation. In fact, the original motivation for this work was from Smith [9] who wanted to quantify the cryptographic security of biometric encryption.
- Performance limits of biometric matchers: While some algorithms outperform others, it clear that there are ultimate limits to error rates, based on the information available in the biometric features. In this application, the biometric feature information is related to the discrimination entropy [7].
- Biometric fusion: Systems which combine biometric features are well understood to offer increased performance [19]. It may be possible to use the measure of biometric feature

information to quantify whether a given combination of features offers any advantage, or whether the fused features are largely redundant. The example of fusion of FLD and PCA (200 features) given here clearly falls into the latter category, since it does not necessarily offer double the amount of information.

- Novel biometric features: Many novel biometric features have been suggested, but it is often unclear whether a given feature offers much in the way of identifiable information. Biometric information measurement may offer a way to validate the potential of such features.
- Privacy protection: It would be useful to quantify the threat to privacy posed by the release of biometric feature information, and also to be able to quantify the value of technologies to preserve privacy, such as algorithms to de-identify face images [15].

REFERENCES

[1] Adler, A., "Vulnerabilities in biometric encryption systems" *Audio- and Video-based Biometric Person Auth.* Tarrytown, NY, USA, Jul. 20–22, 2005

[2] Adler, A., Youmaran, R., Loyka, S., "Information content of biometric features" *Biometrics Consortium Conference* Washington, DC, USA, Sep. 19-21, 2005.

[3] Alter O, Brown PO, Botstein D., "Singular value decomposition for genome-wide expression data processing and modeling", *Proc Natl. Acad. Sci.*, *97*:10101–10106, 2000.

[4] Craw, I., Costen, N.P., Kato, T., Akamatsu, S., "How should we represent faces for automatic recognition?", *IEEE Trans. Pat. Anal. Mach. Intel.* **21**725–736, 1999

[5] Conover, W.J., *Practical Nonparametric Statistics*, Wiley, 1980.

[6] Cover, T.M., Thomas, J.A., *Elements of Information Theory* New York: Wiley, 1991

[7] Daugman, J., "The importance of being random: Statistical principles of iris recognition." *Pattern Recognition*, **36**:279–291, 2003.

[8] Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D., "Sheep, Goats, Lambs and Wolves: An Analysis of Individual Differences in Speaker Recognition Performance", Proc. Int. Conf. Auditory-Visual Speech Processing, Sidney, Australia, Nov. 1998

[9] Dodis, Y., Reyzin, L., Smith, A., "Fuzzy Extractors and Cryptography, or How to Use Your Fingerprints", Proc. Eurocrypt'04, (2004) `http://eprint.iacr.org/2003/235/`

[10] Draper, B.A., Baek, K., Bartlett, M.S., Beveridge, J.R., "Recognizing faces with PCA and ICA", *Computer Vision and Image Understanding*, **91**:115-137, 2003.

[11] Golfarelli, M., Maio, D., Maltoni, D., "On the Error-Reject Tradeoff in Biometric Verification Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**:786-796, 1997.

[12] Grother, P., "Software Tools for an Eigenface Implementation" National Institute of Standards and Technology, (2000) `http://www.nist.gov/humanid/feret/`

[13] Li, S., Jain A. (Eds.). Handbook of Face Recognition. Springer, 2005.

[14] Lee, T.W., "Nonlinear approaches to independent component analysis", Proc. American Institute of Physics, 1999.

[15] Newton, E.M., Sweeney, L., Malin, B., "Preserving Privacy by De-Identifying Face Images", *IEEE Trans. Knowledge Data Eng.* **17** 232–243, 2005.

[16] Pankanti, S., Prabhakar, S., Jain, A.K., "On the Individuality of Fingerprints", *IEEE Trans. Pat. Anal. Mach Intel.*, **24**:1010–1025, 2002

[17] Phillips, P.J., Grother, P., Micheals, R.J., Blackburn, D.M., Tabassi, E., Bone, J.M., *FRVT 2002: Evaluation Report*, NIST, March 2003, `http://www.frvt.org/DLs/FRVT_2002_Evaluation_Report.pdf`

[18] Phillips, P.J., Scruggs, T.W., OToole, A.J., Flynn, P.J., Bowyer, K.W., Svhott, C.L., Sharpe, M., *FRVT 2006: Evaluation Report*, NIST, March 2007, `http://www.frvt.org/FRVT2006/docs/FRVT2006andICE2006LargeScaleReport.pdf`

[19] Ross, A., Jain, A., "Information Fusion in Biometrics", *Pattern Recognition Letters*, **24**:2115-2125, 2003

[20] Soutar, C., Roberge, D., Stoianov, A., Gilroy, R., Vijaya, B.: "Biometric Encryption using image processing", Proc. SPIE Int. Soc. Opt. Eng., **3314** 178-188 (1998)

[21] Turk, M., Pentland, A., "Eigenfaces for recognition", *J. Cognitive Neuroscience*, **3**:71-86, 1991.

[22] Uludag, U., Pankanti, S., Prabhakar, S., Jain, A.K.: "Biometric Cryptosystems: Issues and Challenges", *Proc. IEEE* **92**:948–960, 2004

[23] Wayman, J.S., "The cotton ball problem", *Biometrics Conference*, Washington DC, USA, Sep. 20-22, 2004.

[24] Xiang, C., Fan, X.A., Lee, T.H., "Face recognition using recursive Fisher linear discriminant", *Communications, Circuits and Systems international Conference on*, June 27-29, 2004.