TOWARDS A MEASURE OF BIOMETRIC INFORMATION

Andy Adler
University of Ottawa,
Ontario, Canada
adler@site.uottawa.ca

Richard Youmaran
University of Ottawa,
Ontario, Canada
youmaran@site.uottawa.ca

Sergey Loyka
University of Ottawa,
Ontario, Canada
sloyka@site.uottawa.ca

Abstract

This paper addresses the issue of the information content of a biometric image or system. 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 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 ICA based face recognition, with biometric information calculated to be 45.0 *bits (PCA),* 39.0 *bits (ICA) and* 46.9 *bits (fusion of PCA and ICA features). Finally, we discuss general applications of this measure.*

Keywords *— Biometric features, Relative entropy, Face recognition, Information content.*

1 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 [10] 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 ([1], [12], [6], [10]). However, none of these methods directly address the measurement of information content of biometric data. 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 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. Based on the definition we introduced, this paper develops a mathematical framework to measure biometric 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.

2 Methods

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

2.1 Measure requirements

In order to elaborate the requirements that a good measure of biometric information measure 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 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 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.
3. If a person has unusual feature values (i.e. far from the population mean), they become more distinguishable, and their biometric 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 information is the relative entropy ($D(p\|q)$) [5] between the intra- ($q(\mathbf{x})$) and inter-person ($p(\mathbf{x})$) 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

$$D(p\|q) = \int\limits_{\mathbf{x}} p(\mathbf{x}) log_2 \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \qquad (1)$$

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, 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 $q(\mathbf{x})$ [5]. $D(p\|q)$ 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 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.

2.2 Distribution modeling

In a generic biometric system, F biometric features are measured, to create a biometric feature vector $\mathbf{x}$ ($F\times1$) for each person. For person p, we have N_p 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 $\mu_q = \underset{q}{E}[X]$ where the feature mean of person p, μ_p, is defined analogously, replacing q by p. The population feature covariance is $\mathbf{\Sigma}_q = \underset{q}{E}\left[(X-\mu_q)^t(X-\mu_q)\right]$. The individual's feature covariance, $\mathbf{\Sigma}_p$, is again defined analogously. Features are calculated from a set of N_q images using different component analysis methods such as Principle Component Analysis (PCA, also referred to as Eigenface features) [8][11] and Independent Component Analysis (ICA) [3][7][9]. μ_p and μ_q are $F\times1$ 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 a Gaussian model for p and q, $D(p\|q)$can be written as:

$$D(p\|q) \quad = \quad k\left(\alpha + trace\left((\mathbf{\Sigma}_p + \mathbf{T})\mathbf{\Sigma}_q{}^{-1} - \mathbf{I}\right)\right) \qquad (2)$$

where $\alpha = ln\frac{|\mathbf{\Sigma}_q|}{|\mathbf{\Sigma}_p|}$, $\mathbf{T} = (\mu_p - \mu_q)^t(\mu_p - \mu_q)$ 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 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 a 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 ∈ 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 increase the condition number (and thus reduce 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.

2.3 Regularization 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) [7][8] 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) of the feature covariance matrix, such that

$$\mathbf{U}\mathbf{S}_q\mathbf{U}^t = svd(cov(X)) = svd(\mathbf{\Sigma}_q) \tag{3}$$

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 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:

$$\mathbf{S}_p = \mathbf{U}^t\mathbf{\Sigma}_p\mathbf{U} \tag{4}$$

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.

Based on this regularization scheme, (2) may be rewritten in the PCA space as:

$$D(p\|q) = k\left(\beta + trace\ \mathbf{U}\left((\mathbf{S}_p + \mathbf{S}_t){\mathbf{S}_q}^{-1} - \mathbf{I}\right)\mathbf{U}^t\right) \tag{5}$$

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

2.4 Regularization 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 using 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

$$M = \begin{cases} 1, & \text{if } i = j \text{ or } (i < L \text{ and } j < L); \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Using (6), $[\mathbf{S}_p]_{i,j} = (\mathbf{M}_{i,j})[\mathbf{U}^t\mathbf{\Sigma}_p\mathbf{U}]_{i,j}$. 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 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).

2.5 Average information of a system

This section has developed a measure of biometric information content of a biometric feature representation of a single individual with respect to the feature distribution of the population. The biometric information will vary between people; those with feature values further from the mean have larger biometric information. Using this approach, the biometric information content of a biometric system is calculated as the average information across all people in the system at a specific L.

3 Face Recognition

Information in a feature representation of faces is calculated using our described method for different individuals. 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 [8] and the ICA face features components using the FastIca algorithm [9]. For PCA and ICA feature decompositions, 288 independent vectors were computed, and the most significant 100 features used for subsequent analysis.

3.1 Biometric information calculations

After fitting the distributions of $p(\mathbf{x})$ and $q(\mathbf{x})$ to a Gaussian model, we initially analyze the biometric information in each PCA and ICA feature separately. PCA features are shown in Fig. 1A, 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 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 information over the first 100 PCA features gives 40.5 bits. On the other hand, ICA features show no gradual decrease with feature number, as shown in Fig. 1B. Interestingly, ICA shows several features with large peaks, suggesting that these features are significantly more informative than the others. A sum of biometric information over the first 100 ICA features gives 13.4 bits.

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 ICA 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 6) with a cut-off point L. This scheme is motivated by the diagonal structure of $\mathbf{S}_p$. To ensure convergence, the mask size L is set to a value smaller than N_p. Results for $D(p\|q)$ for PCA features for each person as a function of L are shown in Fig. 2 for $N_p = 8$, 12 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 $\mathbf{\Sigma}_q$. In order to produce an unique and stable estimate for $D(p\|q)$, it is necessary to choose a compromise between having an under-estimated ($L \ll N_p$) or an over-estimated ($L \geq N_p$) solution. We therefore recommend choosing $L = \frac{3}{4}N_p$, since a larger value of L puts the estimate in an unstable region of Fig. 2. Using this algorithm and value of L, we calculate the overall biometric information for different face recognition algorithms. For PCA features, the average $D(p\|q)$ is 45.0 bits, and for ICA features $D(p\|q)$ is 39.0 bits. If PCA and ICA features are combined (making 200 features in all), average $D(p\|q)$ is 46.9 bits.

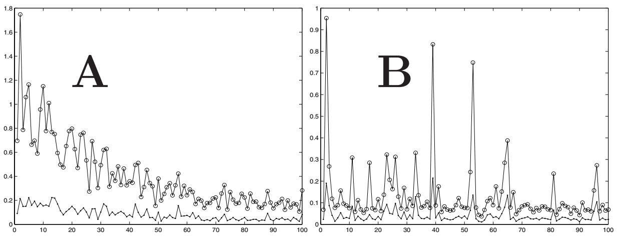

Figure 1. Biometric information as a function of feature number (circles) for (A) PCA (Eigenface) and (B) ICA face feature decomposition. The standard deviation for each value (line) is shown below the $D(p\|q)$ measure.

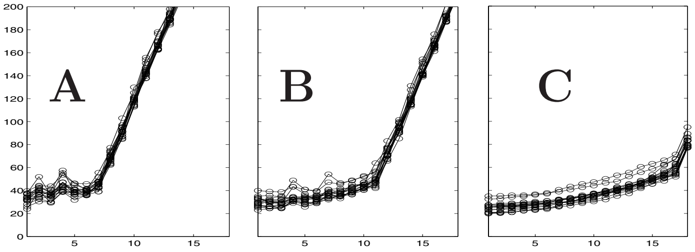

Figure 2. $D(p\|q)$ (y-axis) vs L (x-axis) for each person. Each subfigure represents a different value of N_p: (A) 8, (B) 12 and (C) 18. The curves show that $D(p\|q)$ diverges as Σ_p becomes singular ($L \geq N_p$).

4 Discussion

This paper has introduced a definition of biometric information and an algorithm to measure it from a set of population and individual biometric features, as measured by a biometric algorithm under test. Examples of its application were shown for two different face recognition algorithms based on PCA (Eigenface) and ICA feature decompositions. In a general biometric system, the following issues associated with biometric features must be considered: 1) 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, 2) Raw sample images need to be processed by alignment and scaling before features can be measured, 3) Feature dimensionality may not be constant. 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: uniqueness of biometric features, inherent limits to biometric template size requirements, feasibility of biometric encryption, performance limits of biometric matchers, biometric fusion and privacy protection.

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] Bartlett, M.S., Movellan, J.R., Sejnowski, T.J., Face recognition by independent component analysis *IEEE Trans. Neural Networks*, **13**:1450–1464, 2002.

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

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

[7] 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.

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

[9] Hyvärinen, A., "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis". *IEEE Trans. Neural Networks* **10**:626–634, 1999.

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

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

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