ROBUST SCORE NORMALIZATION FOR RELATIONAL APPROACHES TO FACE AUTHENTICATION

Florent Perronnin and Jean-Luc Dugelay

Institut Eurécom
Multimedia Communications Department
BP 193, 06904 Sophia Antipolis Cedex, France
{florent.perronnin, jean-luc.dugelay}@eurecom.fr

ABSTRACT

Relational approaches to pattern recognition consist in modeling the relationship between observations. In this paper, we consider two score normalization strategies based on a Bayesian framework for relational approaches to face authentication. The first one is specific to relational approaches and models the relationship between face images of different persons. The second one, which is very general and can be applied to any face authentication system, models directly impostors. These two techniques are compared from a theoretical and an experimental point of view and both comparisons hint at a superiority of the general approach.

1. INTRODUCTION

A biometrics authentication/verification system accepts or rejects a person based on a claimed identity and a sample of the considered biometrics. Hence, authentication is a two-class decision problem and the success of an authentication system is based on the accurate modeling of the client and impostor distributions. Although this framework has long been applied to biometrics such as speaker verification [1, 2], surprisingly, the issue of impostor modeling seems to have drawn very little attention from the face verification community [3].

While most approaches to face recognition consist in estimating for each person a model of his/her face, relational approaches, which are considered in this paper, model directly the relationship between observations. The probabilistic matching [4, 5] which models the distribution of difference images is one of the most famous examples of such an algorithm. Another relational approach which will also be considered in this paper is the probabilistic model of face mapping which models the transformations between face images due to different facial expressions [6] or illumination conditions [7].

Let o_q and o_t denote respectively a query and a template image. We also introduce $\mathscr{R}$ and $\overline{\mathscr{R}}$, respectively the relationship between observations of the same class (i.e. between face images of the same person) and between observations of different classes (i.e. between face images of different persons). In the following, we assume that $\mathscr{R}$ and $\overline{\mathscr{R}}$ are shared by all the clients of the system. The approach to authentication introduced for the probabilistic matching algorithm in [4, 5] can be slightly generalized and is based on the following likelihood ratio:

$$P(\mathscr{R}|o_q,o_t)/P(\overline{\mathscr{R}}|o_q,o_t)$$

This work was supported in part by France Telecom Research and Development.

While this approach to authentication is specific to relational approaches, there exists a more general approach which can be applied to any face authentication system and which consists in estimating an impostor model [1, 2]. Which of these two approaches to score normalization is the more robust is not obvious and this question is the focus of this paper.

The remainder of this article is organized as follows. In the next section, we describe with more details the two possible strategies to verification for relational approaches. In section 3, we compare them from a theoretical point of view. In section 4 we present an experimental comparison carried out on the FERET face database [8] with the probabilistic matching and probabilistic model of face mapping algorithms. Both the theoretical and experimental comparisons indicate that the general approach results in a better performance, especially in the challenging case where variabilities that were not learned during the training phase are observed at test time.

2. TWO AUTHENTICATION STRATEGIES FOR RELATIONAL APPROACHES

To illustrate the two authentication strategies for relational approaches, we first introduce the simple Gaussian classifier considered in [4, 5]. The difference between face images of the same person is supposed to be a normally distributed random variable. If we denote $\delta = o_q - o_t$, then:

$$P(\delta|\mathscr{R}) = \frac{\exp\left\{-\frac{1}{2}\delta^T S^{-1}\delta\right\}}{(2\pi)^{N/2}|S|^{1/2}}$$

where N is the dimension of the image space, i.e. the number of pixels in o_q or o_t and $|.|$ is the determinant operator. The covariance matrix S is the only parameter and is estimated with pairs of images of the same person. Although this classifier has little practical value due to the high dimensionality of the image space, the theoretical comparison of both strategies to authentication on this classifier is simple. Interestingly, $P(\delta|\mathscr{R}) \equiv P(o_q|o_t,\mathscr{R})$ with:

$$P(o_q|o_t,\mathscr{R}) = \frac{\exp\left\{-\frac{1}{2}(o_q-o_t)^T S^{-1}(o_q-o_t)\right\}}{(2\pi)^{N/2}|S|^{1/2}}$$

The difference is that the notation $P(\delta|\mathscr{R})$ assumes that δ is emitted by a Gaussian with zero mean while the notation $P(o_q|o_t,\mathscr{R})$ assumes that o_q is emitted by a Gaussian with mean o_t. In the remainder, we will use the notation $P(o_q|o_t,\mathscr{R})$ as it is more general and can be applied to any relational approach.

2.1 A relational strategy

Generalizing the approach to authentication introduced in [4, 5], acceptance/rejection should be based on the following test ratio:

$$P(\mathscr{R}|o_q,o_t)/P(\overline{\mathscr{R}}|o_q,o_t) \gtrless \theta$$

where θ is an application dependent threshold which is set according to the desired level of security. However, as $P(\mathscr{R}|o_q,o_t)$ and $P(\overline{\mathscr{R}}|o_q,o_t)$ are difficult to estimate, one uses Bayes' formula to rephrase the previous test ratio as follows:

$$P(o_q|o_t,\mathscr{R})/P(o_q|o_t,\overline{\mathscr{R}}) \gtrless \theta' \quad (1)$$

where θ' now incorporates also $P(\mathscr{R}|o_t)$ and $P(\overline{\mathscr{R}}|o_t)$, respectively the probabilities of a client or an impostor trial on the template o_t. If $P(o_q|o_t,\overline{\mathscr{R}})$ is also assumed to be Gaussian:

$$P(o_q|o_t,\overline{\mathscr{R}}) = \frac{\exp\left\{-\frac{1}{2}(o_q-o_t)^T\overline{S}^{-1}(o_q-o_t)\right\}}{(2\pi)^{N/2}|\overline{S}|^{1/2}} \quad (2)$$

where $\overline{S}$ is estimated on pairs of images of different persons.

The relational approach to score normalization will be later referred to as R-norm.

2.2 A general strategy

If o_t is the template image for client C, then in the expression $P(o_q|o_t,\mathscr{R})$, $(o_t,\mathscr{R})$ can be seen as a model of C: $M_C \equiv (o_t,\mathscr{R})$. Note that grouping o_t and $\mathscr{R}$ would not have been possible if we had kept the notation δ. Let $\overline{M_C}$ denote the anti-model of C, i.e. the model of all the impostors that could try to gain access to the system by claiming the identity of C. Then the classical approach to verification in this case is [1, 2]:

$$P(M_C|o_q)/P(\overline{M_C}|o_q) \gtrless \theta$$

Using one more time Bayes' formula, we get:

$$P(o_q|M_C)/P(o_q|\overline{M_C}) = P(o_q|o_t,\mathscr{R})/P(o_q|\overline{o_t,\mathscr{R}}) \gtrless \theta' \quad (3)$$

where θ' now incorporates $P(M_C)$ and $P(\overline{M_C})$, respectively the probabilities of a client and an impostor trials on the model M_C. There exists two traditional approaches to model $\overline{M_C}$: the background model set approach (BMS) [1] and the universal background model (UBM) [2]. The BMS uses one $\overline{M_C}$ per client C while the UBM makes use of one unique model U for all clients C. As the focus of this paper is not on the comparison between the BMS and the UBM (see [2] for such a comparison in the case of speaker authentication and [3] in the case of face authentication), and as the UBM approach is simpler, we choose the UBM.

In practice, if $P(o_q|\overline{o_t,\mathscr{R}}) = P(o_q|U)$ is also a Gaussian, the parameters of U, which include this time both the mean and the covariance matrix, are simply estimated with training data from a large number of people. Let μ and Σ denote respectively the mean and covariance of this distribution:

$$P(o_q|\overline{o_t,\mathscr{R}}) = \frac{\exp\left\{-\frac{1}{2}(o_q-\mu)^T\Sigma^{-1}(o_q-\mu)\right\}}{(2\pi)^{N/2}|\Sigma|^{1/2}} \quad (4)$$

The general approach to score normalization will be later referred to as G-norm.

3. THEORETICAL COMPARISON

While the likelihood ratios of both strategies to authentication have the same numerator (see equations (1) and (3)), i.e. while the client distribution is modeled in the same manner, denominators are different as $\overline{M_C} = \overline{(o_t,\mathscr{R})} \neq (o_t,\overline{\mathscr{R}})$ and thus the impostor distribution is modeled differently. Since $\Sigma \approx \overline{S}$ (e.g. see [9]), for this Gaussian classifier the main difference between $P(o_q|o_t,\overline{R})$ and $P(o_q|\overline{o_t,\mathscr{R}})$ is in the means of the Gaussians.

In effect, R-norm and G-norm are very different: R-norm measures the amount of inter-class variability between two images and G-norm measures a sort of distance between the query image and the data used to train the universal model U.

We will now show with two arguments that, from a theoretical point of view, G-norm is superior to R-norm. The first argument holds for any relational approach. The validity of the second one is limited to the Gaussian classifier and, incidentally, to the probabilistic matching algorithm which is directly derived from this classifier.

3.1 First argument

If T is the set of all possible template images, then $\overline{o_t}$ is defined as $T-\{o_t\}$. We now rewrite $P(o_q|\overline{o_t,\mathscr{R}})$ as follows:

$$\begin{aligned}
P(o_q|\overline{o_t,\mathscr{R}}) &= \frac{P(o_q,\overline{o_t,\mathscr{R}})}{P(\overline{o_t,\mathscr{R}})} \\
&= \frac{P(o_q,o_t,\overline{\mathscr{R}})+P(o_q,\overline{o_t},\mathscr{R})+P(o_q,\overline{o_t},\overline{\mathscr{R}})}{P(\overline{o_t,\mathscr{R}})} \\
&= \frac{P(o_q,o_t,\overline{\mathscr{R}})+P(o_q,\overline{o_t})}{P(\overline{o_t,\mathscr{R}})} \\
&= P(o_q|o_t,\overline{\mathscr{R}})\frac{P(o_t,\overline{\mathscr{R}})}{P(\overline{o_t,\mathscr{R}})}+P(o_q|\overline{o_t})\frac{P(\overline{o_t})}{P(\overline{o_t,\mathscr{R}})}
\end{aligned}$$

Hence $P(o_q|\overline{o_t,\mathscr{R}})$ takes into account $P(o_q|o_t,\overline{\mathscr{R}})$, the normalization score of R-norm, and an additional term $P(o_q|\overline{o_t})$. Note that $P(o_q|o_t,\overline{\mathscr{R}})$ is maximum when $o_q = o_t$ (c.f. equation (2)) which intuitively is not satisfying as we would like the normalization score to be as small as possible when $o_q = o_t$. The additional term in R-norm prevents this unwanted effect (c.f. equation (4)). This first argument favors the choice of G-norm over R-norm.

3.2 Second argument

Until now, we have always assumed a shared model $\overline{\mathscr{R}}$ of anti-relationship. As $\overline{\mathscr{R}}$ is supposed to model all the possible transformations between face images of different persons, it should be described with a very large number of parameters and, for a robust estimation, these parameters should be estimated with a large amount of training data.

However, when comparing o_t and o_q one does not need to know the whole distribution of the difference between images that belong to two arbitrary persons. Instead, as we have access to the identity of the client C to be verified, we could concentrate on the distribution of the difference between o_t and all the images that do not belong to C. This would require an $\overline{\mathscr{R}_t}$, i.e. a specific anti-relationship model, for each template image o_t. Intuitively, using an $\overline{\mathscr{R}_t}$ should yield a

better performance than a $\overline{\mathscr{R}}$ as we would then focus on the region of interest of the distribution.

Let $\overline{\mathbf{O_C}}$ denote the random variable which describes the emission of the query images o_q that do not belong to C. Theoretically, there is one such distribution for each client C and it should be estimated with all the available images that do not belong to C. However, in practice, this distribution will be estimated on an independent training set that contains none of the template images of the evaluation set. Moreover, even if this training set contained one or a few images from C, their influence would be negligible compared to all the other images. Hence, for all clients, we use a shared random variable denoted $\overline{\mathbf{O}}$. If $\overline{\mathbf{O}}$ is assumed normally distributed with mean μ and covariance Σ (assuming that we use the same data to train this distribution as in 2.2), then $\overline{\mathbf{O}} - \mathbf{o_t}$ is also normally distributed with mean $\mu - o_t$ and covariance Σ. Hence $(o_q - o_t) - (\mu - o_t) = o_q - \mu$ and we get $P(o_q|o_t,\overline{\mathscr{R}_t}) = P(o_q|U) = P(o_q|\overline{o_t,\mathscr{R}})$. Keeping in mind that $P(o_q|o_t,\overline{\mathscr{R}_t})$ should yield a better performance than $P(o_q|o_t,\overline{\mathscr{R}})$, G-norm should theoretically outperform R-norm.

Note that this argument is not exact for the probabilistic model of face mapping algorithm which does not work directly on difference images.

4. EXPERIMENTAL COMPARISON

In this section, we first present the database used for our experiments. We then describe the experimental setup. Finally we present results for two relational approaches to face recognition: the probabilistic matching (PM) [4, 5] and the probabilistic model of face mapping (PMFM) [6, 7]. Due to space limitation, these algorithms will not be reviewed in this paper and the interested reader can refer to the previously cited articles.

4.1 The database

Experiments were carried out on the FERET face database [8]. To train our systems, we used 500 individuals. For each individual, we extracted one image from the FA set and one image from the FB set. The FA and FB sets contain frontal views that exhibit large variations in facial expression but almost no other variability.

To assess the performance of R-norm and G-norm on the PM and PMFM we used 200 persons (we made sure that no person present in this evaluation set was also present in the training set). For each of these persons, we extracted one image in each of the following sets: BA, BD, BE, BF, BG, BJ and BK. The BA images, which are similar to the FA images, are used as enrollment images which means that each client is enrolled with one unique image. All other images are used as test images and are split into four subsets:

- BJ images are similar to the FB images.
- BK images exhibit large variations in illumination.
- BE & BF images correspond respectively to rotations of the head of 15^o to the left and right.
- BD & BG images correspond respectively to rotations of the head of 25^o to the left and right.

The total number of test images was hence 1,200. It is especially interesting to test the two strategies to score normalization on the BK, BE & BF and BD & BG sets as, when training a system, one seldom has access to the exact testing conditions and in practice, there is always a mismatch between the training and test data.

All the FERET images were pre-processed to extract 128x128 pixels normalized facial regions.

4.2 Experimental setup

To train the client distribution, i.e. the parameters of $\mathscr{R}$, we used the 500 pairs of images of the FA and FB sets. For each of the 500 persons, the FA and FB images were successively used as template and query images and the parameters of $\mathscr{R}$ were thus estimated with 1,000 pairs of images. To train the impostor distribution in the case of R-norm, i.e. to estimate the parameters of $\overline{\mathscr{R}}$, for each image in FA and FB we chose randomly another image in FA or FB that belonged to a different person. Hence, the parameters of $\overline{\mathscr{R}}$ were also estimated with 1,000 pairs of images. To train the impostor distribution in the case of G-norm, i.e. to estimate the parameters of U, we used all the FA and FB images (in this case we do not consider pairs of images).

For the PM, we used 25 features for the client distribution, but also for the impostor distributions for both approaches to score normalization.

For the PMFM, the different systems were trained exactly as described in [6] up to 16 Gaussians per mixture. We did not make use of the illumination compensation algorithm described in [7].

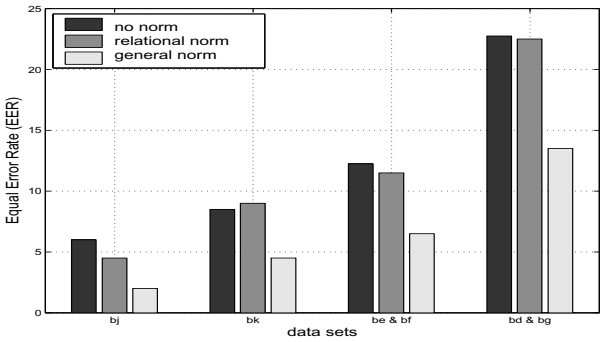

Figure 1: EER of the probabilistic matching (PM) on four subsets of FERET.

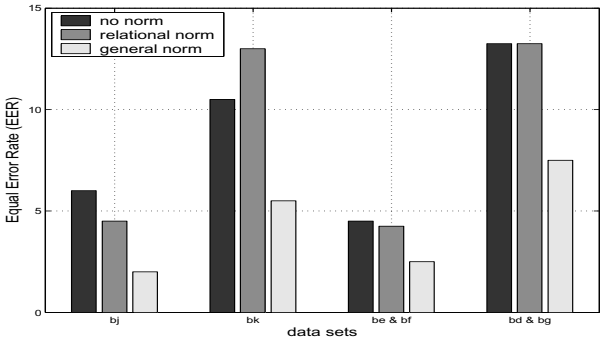

Figure 2: EER of the probabilistic model of face mapping (PMFM) on four subsets of FERET.

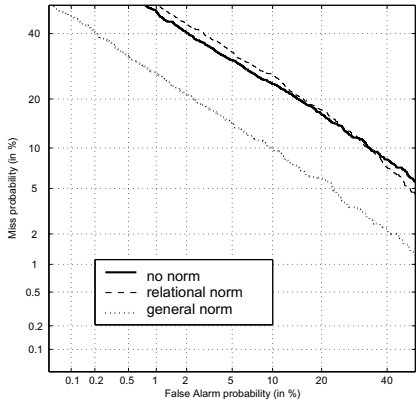

Figure 3: DET curve for the probabilistic matching (PM) merging the four considered subsets of FERET.

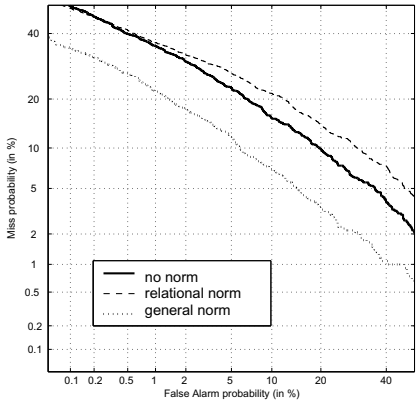

Figure 4: DET curve for the probabilistic model of face mapping (PMFM) merging the four considered subsets of FERET.

4.3 Results

We present results for systems 1) with no score normalization, 2) with R-norm and 3) with G-norm. We first show the performance (equal error rate or EER) for the PM on figure 1 and the PMFM on figure 2 on the four separate data sets described in 4.1. We would like to underline that the focus of this paper is not on a comparison of the PM and the PMFM.

For both algorithms, R-norm only seems to have a significantly positive impact on the EER in the case where there is no mismatch between the training and test conditions (BJ set). In the case of a mismatch, R-norm has at best no significant impact on the performance and can even result in an increase of the EER (see the performance of the PMFM on set BK on figure 2). This seems to indicate that, when facing new conditions, R-norm is unable to distinguish between inter- and intra-class variabilities. On the other hand, G-norm results in a large and consistent decrease of the EER for both matched and mismatched conditions.

If we merge the four data sets into one set, the EER for the PM (resp. PMFM) is, with a confidence interval of 95%, $17.9\% \pm 2.2\%$ (resp. $13.6\% \pm 1.9\%$) when there is no score normalization, $18.3\% \pm 2.2\%$ (resp. $16.3\% \pm 2.1\%$) for R-norm and $9.9\% \pm 1.7\%$ (resp. $8.1\% \pm 1.5\%$) for G-norm.

As the EER only represents the performance for a specific threshold θ, we also plotted on figures 3 and 4 the Detection Error Trade-off (DET) curves for the PM and PMFM. For both algorithms and all false alarm or miss probabilities, G-norm performs significantly better than the systems without normalization or with R-norm, thus validating the intuition we got from the theoretical comparison.

Finally, we would like to outline that, in addition to its better performance, G-norm requires significantly less computation than R-norm in the case where one query image has to be scored against multiple template images as R-norm computes the same normalization score $P(o_q|\overline{o_t,\mathscr{R}}) = P(o_q|U)$ for each o_t while G-norm has to compute a $P(o_q|o_t,\overline{\mathscr{R}})$ per o_t.

5. SUMMARY

In this paper, we considered two strategies to score normalization for the class of relational approaches to face recognition. The first strategy, which is specific to relational approaches and which consists in modeling the relationship between face images of different persons is a direct extension of the work of [4, 5]. The second one, which consists in building directly an impostor model, is very general and can be applied to any face authentication system. These two techniques were first compared from a theoretical and then from an experimental point of view on two very different relational approaches to face authentication. Both comparisons indicated that the general approach to score normalization results in a better performance, especially in the challenging but realistic case where there is a mismatch between the training and test conditions.

REFERENCES

[1] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[2] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[3] C. Sanderson and K. Paliwal, "Likelihood normalization for face authentication in variable recording conditions," in *IEEE ICIP*, 2002, vol. 1, pp. 301–304.

[4] B. Moghaddam, and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 696–710, 1997.

[5] B. Moghaddam, W. Wahid and A. Pentland, "Beyond eigenfaces: probabilistic matching for face recognition," *IEEE AFGR*, pp. 30–35, 1998.

[6] F. Perronnin, J.-L. Dugelay and K. Rose, "Deformable face mapping for person identification," *IEEE ICIP*, vol. 1, pp. 661–664, 2003.

[7] F. Perronnin and J.-L. Dugelay, "A model of illumination variation for robust face recognition," in *MMUA workshop*, 2003, pp. 157–164.

[8] P. J. Phillips, H. Moon, S. A. Rizvi and P. J. Rauss, "The feret evaluation methodology for face recognition algorithms," *IEEE Trans. on PAMI*, vol. 22, no. 10, pp. 1090–1104, Oct 2000.

[9] X. Wang and X. Tang, "Unified subspace analysis for face recognition," in *IEEE ICCV*, 2003, pp. 679–686.