Discriminative Face Recognition

Florent Perronnin and Jean-Luc Dugelay

Institut Eurecom
Multimedia Communications Department
BP 193, 06904 Sophia Antipolis Cedex, France
{perronni, dugelay}@eurecom.fr

Abstract. A novel probabilistic deformable model of face mapping was recently introduced and successfully applied to automatic person identification. In this paper, we consider the use of discrimination to improve the performance of this system. It is possible to introduce discriminative information at two different levels: 1) in the deformable model used to match face images and 2) in the face representations. We explore both types of discrimination and compare them in terms of performance and computational complexity. Results are presented on the FERET face database and show that, in this framework and for the discriminative techniques that were considered, the discrimination of the deformable model should be preferred and can result in a 25-40% relative error rate reduction compared to the non-discriminative system.

1 Introduction

We recently introduced a novel probabilistic deformable model of face mapping [1] whose philosophy is similar to Elastic Graph Matching (EGM) [2]. The global face deformation, which is too complex to be modeled directly, is divided into a set of local transformations with the constraint that neighboring transformations must be consistent with each other. Local transformations and neighboring constraints are embedded within a probabilistic framework using two-dimensional Hidden Markov Models (2-D HMMs).

Given a template face $\mathcal{F}_T$, a query face $\mathcal{F}_Q$ and a deformable model of the face $\mathcal{M}$, for a face identification task we have to estimate $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$. The two major differences between EGM and the approach presented in [1] are 1) in the use of the HMM framework which provides efficient formulae to compute $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$ and train automatically all the parameters of $\mathcal{M}$, 2) in the use of a shared deformable model of the face $\mathcal{M}$ for all individuals, which is particularly useful when little enrollment data is available.

This recognition system can be improved through the use of discrimination. In this framework, it is possible to introduce discriminative information at two different levels: in the deformable model $\mathcal{M}$ or in the face representations $\mathcal{F}_T$ and $\mathcal{F}_Q$. Although both types of discrimination aim at reducing the error rate, it must be underlined that they are conceptually very different.

The remainder of this paper is organized as follows. Our probabilistic deformable model of face mapping is briefly reviewed in the next section. In sections

3 and 4, we describe the two techniques used for discrimination. To discriminate the deformable model, we perform discriminative training of HMM parameters. To discriminate face representations, we project Gabor features in discriminant sub-spaces. In section 5 we give experimental results for a face identification task on the FERET face database which show that, in this framework and for the discriminative techniques that were considered, the discrimination of the deformable model $\mathcal{M}$ should be preferred as it performs better and it is more computationally attractive.

2 A Deformable Model of the Face

As a global face transformation (deformation) may be too complex to be modeled directly, it should be approximated with a set of *local transformations*. The composition of all local transformations, i.e. the global transformation, should be rich enough to model a wide range of facial deformations. However, if we allow any combination of local transformations, the model could be over-flexible and may manage to patch together very different faces. Hence the introduction of a *neighborhood coherence constraint* whose purpose is to provide context information. To combine local transformations and consistency costs, we embed the system within a probabilistic framework using 2-D HMMs. At any position on the face the system is in one of a finite set of states where each state represents a local deformation. Emission probabilities model the cost of local transformations and transition probabilities relate states of neighboring regions and implement the consistency rules.

2.1 Local Transformations

Feature vectors are extracted on a sparse grid from the template image $\mathcal{F}_T$ and on a dense grid from the query image $\mathcal{F}_Q$ as is done in EGM [2]. Each vector summarizes local properties of the face. In our experiments, we used Gabor features (c.f. section 5). We then apply a set of local deformations (i.e. translations) at each position (i, j) of the sparse grid. Each transformation maps a feature vector of $\mathcal{F}_T$ with a feature vector in $\mathcal{F}_Q$.

Let $o_{i,j}$ be the observation extracted from $\mathcal{F}_T$ at position (i, j) and let $q_{i,j}$ be the associated state (i.e. local deformation). If τ is a translation vector, the probability that at position (i, j) the system emits observation $o_{i,j}$ knowing that it is in state $q_{i,j} = \tau$, is $b_\tau(o_{i,j}) = P(o_{i,j}|q_{i,j} = \tau, \lambda)$ where $\lambda = (\lambda_Q, \lambda_\mathcal{M})$. We clearly separate λ into Face Dependent (FD) parameters $\lambda_\mathcal{Q}$ which are extracted from $\mathcal{F}_Q$ (i.e. the feature vectors) and the Face Independent Transformation (FIT) parameters $\lambda_\mathcal{M}$, i.e. the parameters of the shared deformation model $\mathcal{M}$ that can be trained reliably by pooling together all training. The emission probability $b_\tau(o_{i,j})$ represents the cost of matching $o_{i,j}$ with the corresponding feature vector in $\mathcal{F}_Q$ that will be denoted $m^\tau_{i,j}$. $b_\tau(o_{i,j})$ is modeled with a mixture of Gaussians as linear combinations of Gaussians have the ability to approximate

arbitrarily shaped densities:

$$b_\tau(o_{i,j}) = \sum_k w_{i,j}^k b_{\tau,k}(o_{i,j}) \tag{1}$$

$b_{\tau,k}(o_{i,j})$'s are the component densities and the $w_{i,j}^k$'s are the mixture weights and must satisfy the following constraint: $\forall(i,j)$ and $\forall\tau$, $\sum_k w_{i,j}^k = 1$. Each component density is a N-variate Gaussian function of the form:

$$b_{\tau,k}(o_{i,j}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma_{i,j}^k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(o_{i,j} - \mu_{i,j}^{\tau,k})^T \Sigma_{i,j}^{k(-1)}(o_{i,j} - \mu_{i,j}^{\tau,k})\right\} \tag{2}$$

where $\mu_{i,j}^{\tau,k}$ and $\Sigma_{i,j}^k$ are respectively the mean and covariance matrix of the Gaussian, N is the size of feature vectors and $|.|$ is the determinant operator. This HMM is non-stationary as Gaussian parameters depend on the position (i,j). We use a bi-partite model which separates the mean into additive FD and FIT parts:

$$\mu_{i,j}^{k,\tau} = m_{i,j}^\tau + \delta_{i,j}^k \tag{3}$$

where $m_{i,j}^\tau$ is the FD part of the mean and $\delta_{i,j}^k$ is a FIT offset. Intuitively, $b_\tau(o_{i,j})$ should be approximately centered and maximum around $m_{i,j}^\tau$.

2.2 Neighborhood Consistency

The neighborhood consistency of the transformation is ensured via the transition probabilities of the 2-D HMM. If we assume that the 2-D HMM is first order Markovian in a 2-D sense, the transition probabilities are of the form $P(q_{i,j}|q_{i,j-1}, q_{i-1,j}, \lambda)$. However, as explained in the next section, a 2-D HMM can be approximated by a Turbo-HMM (T-HMM): a set of horizontal and vertical 1-D HMMs that "communicate" through an iterative process. As we want to be insensitive to global translations of face images, we choose the transition probabilities of the horizontal and vertical 1-D HMMs to be of the form:

$$P(q_{i,j} = \tau|q_{i,j-1} = \tau', \lambda) = a_{i,j}^{\mathcal{H}}(\delta\tau) \quad P(q_{i,j} = \tau|q_{i-1,j} = \tau', \lambda) = a_{i,j}^{\mathcal{V}}(\delta\tau) \tag{4}$$

where $\delta\tau = \tau - \tau'$. $a_{i,j}^{\mathcal{H}}$ and $a_{i,j}^{\mathcal{V}}$ model respectively the horizontal and vertical elastic properties of the face at position (i,j) and are part of the face transformation model $\mathcal{M}$.

We assume in the remainder that the initial occupancy probability of the 2-D HMM is uniform to ensure invariance to global translations of face images. To summarize, the parameters we need to estimate are the FIT parameters $\lambda_{\mathcal{M}}$, i.e. w's, δ's, Σ's and transition probabilities $a_{i,j}^{\mathcal{H}}$'s and $a_{i,j}^{\mathcal{V}}$'s.

2.3 Turbo-HMMs

While HMMs have been extensively applied to one-dimensional problems [3], the complexity of their extension to two-dimensions grows exponentially with

the data size and is intractable in most cases of interest. [1] introduced Turbo-HMMs (T-HMMs), in reference to the turbo error-correcting codes, to approximate the computationally intractable 2-D HMMs. A T-HMM consists of horizontal and vertical 1-D HMMs that "communicate" through an iterative process. The T-HMM framework provides efficient formulae to 1) compute efficiently $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$ and 2) train automatically all the parameters of $\mathcal{M}$.

The computation of $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$, i.e. of $P(O|\lambda)$, is based on a modified version of the forward-backward algorithm which is applied successively and iteratively on the rows and columns until the horizontal and vertical passes reach some kind of agreement. This algorithm is clearly linear in the size of the data. It must be underlined that we obtain one horizontal and one vertical scores. As experiments showed that they were generally close, to obtain one unique score we simply averaged them.

The *Maximum Likelihood Estimation* (MLE) formulae for HMM parameters can be derived directly by maximizing Baum's auxiliary function $Q(\lambda|\bar{\lambda})$ in the 1-D case [3]. In the T-HMM framework we obtain one horizontal and one vertical functions $Q(\lambda^{\mathcal{H}}|\bar{\lambda}^{\mathcal{H}})$ and $Q(\lambda^{\mathcal{V}}|\bar{\lambda}^{\mathcal{V}})$ that may be incompatible in the case where horizontal and vertical passes do not converge. So a simple idea is to maximize:

$$Q(\lambda|\bar{\lambda}) = Q(\lambda^{\mathcal{H}}|\bar{\lambda}^{\mathcal{H}}) + Q(\lambda^{\mathcal{V}}|\bar{\lambda}^{\mathcal{V}}) \tag{5}$$

At training time, we present pairs of pictures (a template and a query image) that belong to the same person and optimize the transformation parameters $\lambda_{\mathcal{M}}$.

3 A Discriminative Deformable Model of the Face

MLE formulae for HMM parameters can be shown to be optimal when certain conditions hold, including model correctness and infinite training data. However, as generally the true data source is not an HMM and as training data is sparse, other training criteria, especially discriminative criteria, should be considered. While MLE adjusts model parameters $\lambda_{\mathcal{M}}$ to increase the value of $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$ without taking into account competing faces, discriminative criteria take into account competitors. Hence, a new possible objective function could be:

$$\mathcal{F}(\lambda) = \log P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M}) - \log P(\mathcal{F}_T|\bar{\mathcal{F}}_Q, \mathcal{M}) \tag{6}$$

where $\bar{\mathcal{F}}_Q$ is a competitor for $\mathcal{F}_Q$. The choice of such competitors will be explicited in section 5. We should note that a similar criterion is often used in speech recognition to train HMM parameters discriminatively and is referred as *Maximum Mutual Information Estimation* (MMIE) [5, 6].

Analogous to the Baum-Welch algorithm for MLE, the Extended Baum-Welch (EBW) for rational objective functions was introduced in [4] and extended in [5] to continuous density HMMs. The update equations for the delta-offset and the variance, assuming diagonal covariance matrices, can be adapted to our problem as follows (the update for a single dimension is shown):

$$\hat{\delta}^k_{i,j} = \frac{\left\{\theta^k_{i,j}(\mathcal{O}) - \bar{\theta}^k_{i,j}(\mathcal{O})\right\} + D\delta^k_{i,j}}{\sum_\tau \left\{\gamma_{i,j}(\tau,k) - \bar{\gamma}_{i,j}(\tau,k)\right\} + D} \tag{7}$$

$$(\hat{\sigma}^k_{i,j})^2 = \frac{\left\{\theta^k_{i,j}(\mathcal{O}^2) - \bar{\theta}^k_{i,j}(\mathcal{O}^2)\right\} + D\left\{(\sigma^k_{i,j})^2 + (\delta^k_{i,j})^2\right\}}{\sum_\tau \left\{\gamma_{i,j}(\tau,k) - \bar{\gamma}_{i,j}(\tau,k)\right\} + D} - (\hat{\delta}^k_{i,j})^2 \qquad (8)$$

where

$$\theta^k_{i,j}(\mathcal{O}) = \sum_\tau \gamma_{i,j}(\tau,k)(o_{i,j} - m^\tau_{i,j}) \quad \theta^k_{i,j}(\mathcal{O}^2) = \sum_\tau \gamma_{i,j}(\tau,k)(o_{i,j} - m^\tau_{i,j})^2 \quad (9)$$

and $\gamma_{i,j}(\tau,k)$ is the probability of being in state $q_{i,j} = \tau$ at position (i,j) with the k-th mixture component accounting for $o_{i,j}$. $\gamma_{i,j}(\tau,k)$'s, $\theta^k_{i,j}(\mathcal{O})$'s and $\theta^k_{i,j}(\mathcal{O}^2)$'s are the accumulators estimated for $\mathcal{F}_Q$ and $\bar{\gamma}_{i,j}(\tau,k)$'s, $\bar{\theta}^k_{i,j}(\mathcal{O})$'s and $\bar{\theta}^k_{i,j}(\mathcal{O}^2)$'s are estimated for $\bar{\mathcal{F}}_Q$. The choice of a proper learning step $1/D$ is of utmost importance as a large value of D would result in slow convergence while a small value may lead to instability. We chose a strategy similar to the one used in [6]: D is set on a per Gaussian level (i.e. a Gaussian specific $D^k_{i,j}$ is used) to the maximum of (1) $\bar{\gamma}_{i,j}(\tau,k)$ and (2) twice the minimum value that guarantees a positive update of variances for all dimensions.

As for the update of mixture weights and transition probabilities, we did not apply the formulae for discrete output probabilities derived in [4] but the update formulae proposed in [7] which are generally preferred for their convergence properties.

It must be underlined that there is no modification of the algorithm at test time.

4 A Discriminative Representation of the Face

Although it should be possible to train a discriminative representation of the face using equations similar to (7), this would require abundant enrollment data. As we generally have very limited data to learn the representation of the face, we implemented another approach based on discriminant sub-spaces.

While Gabor features may be useful for representing local properties of the face, there is no guarantee they are optimal for discriminating between different faces. Hence, the idea is to build a sub-space $P_{i,j}$ which is optimal for discriminating between the features of different individuals at each position (i,j) of the sparse grid, and to project Gabor features in these spaces.

The construction of the spaces is done as follows. We assume we have pairs of template and query images $(\mathcal{F}^p_T, \mathcal{F}^p_Q)$ for building the space (one pair of pictures per person). For each pair we perform the modified version of the forward-backward algorithm and estimate the following quantities (we index each quantity γ, o and μ with p to show it corresponds to the p-th pair of pictures):

$$\mu^p_{i,j} = \sum_{\tau,k} \gamma^p_{i,j}(\tau,k)\mu^{\tau,k,p}_{i,j} \qquad \phi^p_{i,j} = (o^p_{i,j} + \mu^p_{i,j})/2 \qquad (10)$$

Then, we compute the within- and between-scatter matrices:

$$S^w_{i,j} = \sum_p (o^p_{i,j} - \mu^p_{i,j})(o^p_{i,j} - \mu^p_{i,j})^T \qquad S^b_{i,j} = \sum_p (\phi^p_{i,j} - \phi_{i,j})(\phi^p_{i,j} - \phi_{i,j})^T \quad (11)$$

where $\phi_{i,j}$ is the average of the $\phi^p_{i,j}$'s. The optimal projection matrix $P_{i,j}$ is chosen such that it maximizes the following ratio [8]:

$$P_{i,j} = \arg\max_P \frac{|P^T S_b P|}{|P^T S_w P|} \tag{12}$$

The columns of the optimal $P_{i,j}$ are the generalized eigenvectors that correspond to the largest eigenvalues e_k in:

$$S^b_{i,j} e_k = \lambda_k S^w_{i,j} e_k \tag{13}$$

This is known as Fisher's Linear Discriminant (FLD).

The feature vectors $o_{i,j}$'s extracted from $\mathcal{F}_T$ can be directly projected using the corresponding $P_{i,j}$'s. However, this is not possible for $\mathcal{F}_Q$ as, for each feature m, we generally do not know beforehand with which feature $o_{i,j}$ in $\mathcal{F}_T$ it will be matched, and hence, we do not know in which space $P_{i,j}$ it should be projected. Therefore, for $\mathcal{F}_Q$ the projection of features has to be done on-line, at both training and test time, which is computationally intensive.

5 Experimental Results

The following experiments were carried out on a subset of the FERET face database [9]. 1,000 individuals were extracted: 500 for training the face deformation model and the projection matrices and 500 for testing the performance. We use two images (one target and one query image) per training and test individual. It means that test individuals are enrolled with one unique image. Target images are extracted from the gallery (FA images) and query images from the FB probe. FA and FB images are frontal views of the face that exhibit large variabilities in terms of facial expressions. Images are pre-processed to extract 128x128 normalized facial regions.

We used Gabor features that have been successfully applied to face recognition [2]. They have desirable properties of spatial locality and orientation selectivity and are optimally localized in the space and frequency domain. Gabor wavelets can be characterized by the following equation:

$$\psi_{\mu,\nu}(z) = \frac{||k_{\mu,\nu}||^2}{\sigma^2} \exp\left(-\frac{||k_{\mu,\nu}||^2 ||z||^2}{2\sigma^2}\right) \left[\exp(ik_{\mu,\mu} z) - \exp(-\sigma^2/2)\right] \tag{14}$$

where $k_{\mu,\nu} = k_\nu \exp(i\phi_\mu)$. $k_\nu = k_{max}/f^\nu$ with $\nu \in [1, N]$ and $\phi_\mu = \pi\mu/M$ with $\mu \in [1, M]$. μ and ν define respectively the orientation and scale of $k_{\mu,\nu}$. After preliminary experiments, we chose the following set of parameters: $N = 5$, $M = 8$ (which makes 40-dimensional features), $\sigma = 2\pi$, $k_{max} = \pi/4$ and $f = \sqrt{2}$. For each image we normalized the feature coefficients to zero mean and unit variance. A feature vector is extracted every 16 pixels of the template images in both horizontal and vertical directions (sparse grid) and every 4 pixels of the query images (dense grid).

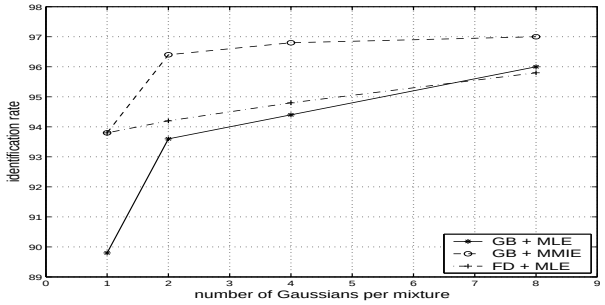

Fig. 1. Performance of the baseline system (GB + MLE), the system which discriminates on the deformable model (GB + MMIE) and the system which discriminates on the face representations (FD + MLE).

For MLE, we first perform a rigid matching of each couple $(\mathcal{F}_T, \mathcal{F}_Q)$ and initialize the parameters of single Gaussian mixtures. Transition probabilities are initialized uniformly. Then the $\lambda_{\mathcal{M}}$ parameters are re-estimated using the modified Baum-Welch (deformable matching). To train multiple Gaussians per mixture we implemented an iterative splitting/re-training strategy.

Before MMIE training we estimate the $\lambda_{\mathcal{M}}$ parameters using standard MLE for the desired number of Gaussians per mixture (GpM). Then identification is performed on the training set to find for each query image $\mathcal{F}_Q$ the best competitors $\bar{\mathcal{F}}_Q$. We only consider the competitors $\bar{\mathcal{F}}_Q$ that satisfy $P(\mathcal{F}_T|\mathcal{F}_Q, \lambda) - P(\mathcal{F}_T|\bar{\mathcal{F}}_Q, \lambda) < \Theta$ where Θ is a parameter that has to be set by hand. Moreover, to reduce the amount of computation, for each $\mathcal{F}_Q$ the number of competitors is limited to a maximum of 5. Once competitors are selected, MMIE training can be carried out.

As for the Fisher Discriminant features (FD), after preliminary experiments we decided to project Gabor features (GB) into 20-dimensional sub-spaces.

Results are presented on Figure 1. While both types of discrimination clearly outperform the baseline system for 1 GpM (approximately 40% relative error rate reduction), for a larger number of Gaussians the discrimination of the transformation model (GB + MMIE) outperforms the discrimination of the face representations (FD + MLE). For 8 GpM, GB + MMIE still manages to outperform the baseline system (25% relative error rate reduction) while the performance of FD + MLE is similar to the baseline (the difference can be considered insignificant). We also tried to combine the discriminations of the deformable model and the representations of the face (FD + MMIE) but we did not manage to outperform GB + MLE.

We should note that the discriminative system based on the face representation is the most computationally intensive. To perform identification on the whole test set ($500 \times 500 = 250,000$ comparisons) using 8 GpM models, it takes approximately 50 min for GB + MLE and GB + MMIE and 1 h 30 min for FD

+ MLE on a Pentium IV 2 GHz with 1 GB RAM. This is due to the on-line projection of Gabor features in the discriminant sub-spaces.

Finally, it must be underlined that the difference in performance and computation time may be due to the framework, i.e. we may draw different conclusions if we used another deformable model of face mapping, but also to the choice of the techniques used for discrimination.

6 Conclusion

In this paper, we considered the use of discrimination to enhance the performance of an automatic person identification system based on a probabilistic deformable model of face mapping. Two types of discrimination were considered: 1) on the deformable model $\mathcal{M}$ with discriminative training of HMM parameters or 2) on the face representations $\mathcal{F}_T$ and $\mathcal{F}_Q$ by projecting features into discriminant sub-spaces. It was shown that, in this framework and for the discriminative techniques that were considered, the discrimination of the deformable model was superior to the discrimination of the face representations, in terms of performance and computational complexity.

As discussed in section 4, the same technique based on discriminative training of HMM parameters could theoretically be used for both types of discrimination and this should be tested in the case where abundant enrollment data is available.

References

1. Perronnin, F., Dugelay, J.-L. and Rose, K., "A Probabilistic Model of Face Transformation Applied to Person Identification", submitted to the Special Issue on Biometric Signal Processing of the EURASIP Journal.
2. Lades, M., Vorbrüggen, J. C., Buhmann J., Lange, J., von der Malsburg, C., Würtz, R. and Konen, W., "Distortion Invariant Object Recognition in the Dynamic Link Architecture", IEEE Trans. on Computers, Vol. 42, No. 3, 1993, March.
3. Rabiner L. R., "A Tutorial on Hidden Markov Models and Selected Applications", Proc. of the IEEE, 1999, Vol. 77, No. 2, Feb.
4. Gopalakrishnan, P.S., Kanevsky, D., Nadas, A. and Nahamoo, D., "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems", IEEE Trans. on Information Theory, Vol. 37, No. 1, 1991, pp. 107-113.
5. Normandin, Y., "An Improved MMIE Training Algorithm for Speaker Independent, Small Vocabulary, Continuous Speech Recognition", ICASSP'91, pp. 537-540.
6. Woodland, P. C. and Povey, D., "Large Scale MMIE Training for Conversational Telephone Speech Recognition", Proc. Speech Transcription Workshop, College Park, 2000.
7. Merialdo, B., "Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information Training", ICASSP'88, Vol. 1, pp. 111-114.
8. Duda, R. O., Hart, P. E. and Stork D. G., "Pattern Classification", John Wiley & Sons, 2nd edition, 2001.
9. Phillips, P. J., Wechsler, H., Huang, J. and Rauss, P. "The FERET Database and Evaluation Procedure for Face Recognition Algorithms," Image and Vision Computing Journal, Vol. 16, No. 5, 1998, pp. 295-306.