DEFORMABLE FACE MAPPING FOR PERSON IDENTIFICATION

Florent Perronnin[*]*, Jean-Luc Dugelay*

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

Kenneth Rose[†]

Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106-9560
rose@ece.ucsb.edu

ABSTRACT

This paper introduces a novel deformable model for face mapping and its application to automatic person identification. While most face recognition techniques directly model the face, our goal is to model the *transformation* between face images of the same person. As a global face transformation may be too complex to be modeled in its entirety, it is approximated by 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 the probabilistic framework of a two-dimensional Hidden Markov Model (2-D HMM). Experimental results on a face identification task show that the new approach compares favorably to the popular Fisherfaces algorithm.

1. INTRODUCTION

Realistic models of the face are required for a wide variety of applications including facial animation, classification of facial expressions or face recognition. Face recognition is a challenging pattern classification problem as face images of the same person are subject to variations in facial expressions, pose, illumination conditions, presence or absence of eyeglasses or facial hair, etc. The focus of this paper will be on the first source of variability.

While most face recognition algorithms attempt to build for each person a *face model* which is intended to describe as accurately as possible his/her intra-face variability, in this paper we model a *transformation* between face images of the same person. To avoid the excessive complexity of direct modeling of the global face transformation, we propose to split it into a set of local transformations and to impose neighborhood consistency of these local transformations. Local transformations and neighboring constraints are naturally embedded within the flexible probabilistic framework of the 2-D HMM.

Deformable models of the face have already been applied to face recognition. The basic approach of Elastic Graph Matching (EGM) [1] is to match two face graphs in an elastic manner. The quality of a match is evaluated with a cost function $\mathcal{C} = \mathcal{C}_v + \rho\mathcal{C}_e$ where $\mathcal{C}_v$ and $\mathcal{C}_e$ are respectively the costs of local matchings and local distortions and ρ controls the rigidity of the matching. [2] elaborated on the idea with the Elastic Bunch Graph Matching (EBGM) and both algorithms were later improved, especially to weight the different parts of the face according to their discriminatory power [3, 4]. The two major differences between the above elastic approaches and the new approach presented in this paper are:

- in the use of the HMM framework which provides efficient formulae to 1) compute the likelihood that a template image $\mathcal{F}_T$ and a query image $\mathcal{F}_Q$ belong to the same person given the face transformation model $\mathcal{M}$, i.e. $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$ and 2) train automatically all the parameters of $\mathcal{M}$,
- 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.

The remainder of this paper is organized as follows. A high-level description of the 2-D HMM as a probabilistic model for face transformation is given in the next section. Sections 3 and 4 provide a quantitative formulation for local transformations and neighborhood consistency, respectively. In section 5 we briefly introduce Turbo-HMMs (T-HMMs) to approximate the computationally intractable 2-D HMMs [5]. Section 6 summarizes experimental results for a face identification task on the FERET face database [6] showing that the proposed approach can significantly outperform the popular Fisherfaces technique [7].

2. FRAMEWORK

A global face transformation is too complex for direct modeling. We hence propose to approximate it with a set of *local transformations*. These transformations should be as simple as possible for an efficient implementation, while the composition of all local transformations, i.e. the resulting global transformation, should be rich enough to model a wide range of facial deformations. However, if we allow all possible combinations of local transformations, the model might become over-flexible and "succeed" to patch together very different faces. This naturally leads to the second component of our framework: a *neighborhood coherence constraint* whose purpose is to provide context information. It must be emphasized that such neighborhood consistency rules produce dependence in the local transformation selection and the optimal solution must therefore involve a global decision. To combine the local transformation and consistency costs, we embed the system within a probabilistic framework using 2-D HMMs.

At any location on the face, the system is in one of a finite set of states. If we assume that the 2-D HMM is first-order Markovian, the probability of the system to enter a particular state at a given location, i.e. the *transition probability*, depends on the state

[*]This work was supported in part by France Telecom Research.

[†]This work was supported in part by the NSF under grants EIA-9986057 and EIA-0080134, the University of California MICRO program, Dolby Laboratories, Inc., Lucent Technologies, Inc., Mindspeed Technologies, Inc., and Qualcomm, Inc.

© IEEE 2003

of the system at the horizontally and vertically adjacent locations. At each position, an observation is emitted by the state according to an *emission probability distribution*. In our framework, local transformations correspond to the states of the 2-D HMM and the target/template image data is the collection of emitted observations. Emission probabilities model the cost associated with a local mapping. These transformations or states are "hidden" and information on them can only be extracted through the observations. Transition probabilities relate states of neighboring regions and implement the consistency rules.

3. 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 [1]. Each vector summarizes local properties of the face. We then apply a set of local geometric transformations to the vectors extracted from $\mathcal{F}_T$. Each transformation maps a feature vector of $\mathcal{F}_T$ with a feature vector in $\mathcal{F}_Q$. Translation, rotation and scaling are examples of simple geometric transformations and may be useful to model local deformations of the face. In the remainder of this paper, we restrict the set of geometric transformations to translations, as a small global affine transformation can be approximated by a set of local translations.

We now formulate the emission probabilities. Let $o_{i,j}$ be the observation extracted from $\mathcal{F}_T$ at position (i,j) (c.f. Fig. 1) and let $q_{i,j}$ be the associated state (i.e. the translation). If $\tau = (\tau_x, \tau_y)$ is a translation vector, the probability that at position (i,j) the system emits observation $o_{i,j}$ given 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 $\lambda = (\lambda_\mathcal{M}, \lambda_\mathcal{Q})$. We clearly separate HMM parameters into Face Dependent (FD) parameters $\lambda_\mathcal{Q}$ that are extracted from $\mathcal{F}_Q$ and Face Independent Transformation (FIT) parameters $\lambda_\mathcal{M}$, i.e. the parameters of the shared transformation $\mathcal{M}$ that can be trained reliably by pooling together the training images of all individuals.

Let $z_{i,j} = (x_{i,j}, y_{i,j})$ denote the coordinates of observation $o_{i,j}$ in $\mathcal{F}_T$. Let $z^\tau_{i,j}$ be the coordinates of the matching feature in $\mathcal{F}_Q$: $z^\tau_{i,j} = z_{i,j} + \tau$. The emission probability $b_\tau(o_{i,j})$ represents the cost of matching these feature vectors. $b_\tau(o_{i,j})$ is modeled with a mixture of Gaussians as linear combinations of Gaussians have the ability to approximate arbitrarily shaped densities:

$$b_\tau(o_{i,j}) = \sum_k w^k_{i,j} b_{\tau,k}(o_{i,j})$$

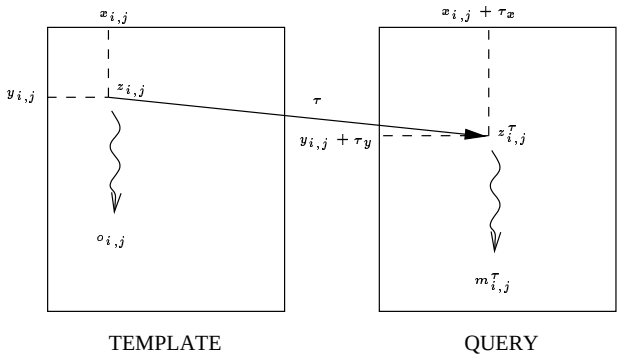

Fig. 1. Local matching

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

$$b_{\tau,k}(o_{i,j}) = \frac{\exp\left\{-\frac{1}{2}(o_{i,j} - \mu^{\tau,k}_{i,j})^T \Sigma^{k(-1)}_{i,j}(o_{i,j} - \mu^{\tau,k}_{i,j})\right\}}{(2\pi)^{\frac{N}{2}}|\Sigma^k_{i,j}|^{\frac{1}{2}}}$$

where $\mu^{\tau,k}_{i,j}$ and $\Sigma^k_{i,j}$ are respectively the mean and covariance matrix of the Gaussian, N is the size of feature vectors and $|\cdot|$ denotes the determinant operator. This HMM is non-stationary as the Gaussian parameters depend on the position (i,j).

Let $m^\tau_{i,j}$ be the feature vector extracted from the matching block in $\mathcal{F}_Q$. We use a bi-partite model which separates the mean into additive FD and FIT parts:

$$\mu^{k,\tau}_{i,j} = m^\tau_{i,j} + \delta^k_{i,j}$$

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

4. 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 a first order Markov process, the transition probabilities are of the form $P(q_{i,j}|q_{i,j-1}, q_{i-1,j}, \lambda)$. However, we show in the next section that 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. So the transition probabilities of the corresponding horizontal and vertical 1-D HMMs are respectively:

$$\begin{aligned} a^\mathcal{H}_{i,j}(\tau;\tau') &= P(q_{i,j} = \tau | q_{i,j-1} = \tau', \lambda) \\ a^\mathcal{V}_{i,j}(\tau;\tau') &= P(q_{i,j} = \tau | q_{i-1,j} = \tau', \lambda) \end{aligned}$$

Invariance to global shift in face images is a desirable property. Hence we choose $a^\mathcal{H}$ and $a^\mathcal{V}$ to be of the form:

$$a^\mathcal{H}_{i,j}(\tau;\tau') = a^\mathcal{H}_{i,j}(\delta\tau) \qquad a^\mathcal{V}_{i,j}(\tau;\tau') = a^\mathcal{V}_{i,j}(\delta\tau)$$

where $\delta\tau = \tau - \tau'$. $a^\mathcal{H}_{i,j}$ and $a^\mathcal{V}_{i,j}$ model respectively the horizontal and vertical elastic properties of the face at position (i,j) and are part of the face transformation model $\mathcal{M}$. If we assume that $\mathcal{F}_T$ and $\mathcal{F}_Q$ have the same scale and orientation, then $a^\mathcal{H}_{i,j}$ and $a^\mathcal{V}_{i,j}$ should have two properties: they should preserve both *local distance*, i.e. τ and τ' should have the same norm, and *ordering*, i.e. τ and τ' should have the same direction. An horizontal separable parametric transition probability that satisfies the two previous constraints is:

$$a^{\mathcal{H}x}_{i,j}(\delta\tau_x) = c(\sigma^{\mathcal{H}x}_{i,j})e^{-\frac{\delta\tau_x^2}{2\sigma^{\mathcal{H}x\,2}_{i,j}}} \qquad a^{\mathcal{H}y}_{i,j}(\delta\tau_y) = c(\sigma^{\mathcal{H}y}_{i,j})e^{-\frac{\delta\tau_y^2}{2\sigma^{\mathcal{H}y\,2}_{i,j}}}$$

where c is a normalization factor such that $\sum_{\delta\tau_x} a^{\mathcal{H}x}_{i,j}(\delta\tau_x) = 1$ and $\sum_{\delta\tau_y} a^{\mathcal{H}y}_{i,j}(\delta\tau_y) = 1$. Similar formulae can be derived for vertical transition probabilities.

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 $a^\mathcal{H}_{i,j}$'s and $a^\mathcal{V}_{i,j}$'s.

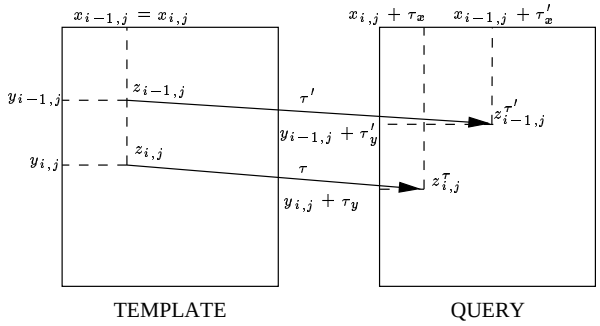

Fig. 2. Neighborhood consistency

5. TURBO-HMMS

While HMMs have been extensively applied to one-dimensional problems [8], the complexity of their extension to two-dimensions grows exponentially with the data size and is intractable in most cases of interest. [5] introduced Turbo-HMMs (T-HMMs), in reference to the celebrated 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.

T-HMMs rely on the following approximation of the joint-likelihood of observations O and states Q given the HMM parameters λ:

$$P(O,Q|\lambda) \approx \prod_j \left[P(o_j^{\mathcal{V}}, q_j^{\mathcal{V}}|\lambda_j^{\mathcal{V}}) \prod_i P(q_{i,j}|o_i^{\mathcal{H}}, \lambda_i^{\mathcal{H}}) \right]$$

where $o_i^{\mathcal{H}}$ and $o_j^{\mathcal{V}}$ are respectively the i-th row and j-th column of observations, $\lambda_i^{\mathcal{H}}$ and $\lambda_j^{\mathcal{V}}$ are the i-th row and j-th column of model parameters and $q_j^{\mathcal{V}}$ is the j-th column of states. Each term $P(o_j^{\mathcal{V}}, q_j^{\mathcal{V}}|\lambda_j^{\mathcal{V}})$ corresponds to a 1-D vertical HMM and $\prod_i P(q_{i,j}|o_i^{\mathcal{H}}, \lambda_i^{\mathcal{H}})$ is in effect a horizontal prior for column j. We can derive the dual formula where 1-D horizontal HMMs communicate through the use of a vertical prior.

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 priors reach some kind of agreement [5]. This algorithm is clearly linear in the size of the data. It must be underlined that we do not obtain one unique score but one horizontal and one vertical score. Combining these two scores is a classical problem of decision fusion. As experiments showed that these scores were generally close, we simply averaged the log-likelihoods. Although this simple heuristic may not be optimal it provided good results. While EGM only takes into account the best transformation during the score computation, we take into account all possible transformations weighted according to their probability, which should yield a more robust score.

During training, we present pairs of pictures (a template and a query image) that belong to the same person and optimize the transformation parameters $\lambda_{\mathcal{M}}$, to increase the likelihood value $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$ (Maximum Likelihood Estimation). This is another advantage of the proposed approach as we can train all model parameters while, to the best of our knowledge, EGM's rigidity parameter (which has the same function as our transition probabilities) must be hand-tuned.

6. EXPERIMENTAL RESULTS

6.1. The Database

All the following experiments were carried out on a subset of the FERET face database [6]. 1,000 individuals were extracted: 500 for training the face deformation model and 500 for testing the performance. We use two images (one target and one query image) per training or 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. For this purpose, we used the coordinates of the eyes and the tip of the nose provided with each image.

6.2. Gabor Features

We used Gabor features that have been successfully applied to face recognition [1, 2, 3, 9] and facial analysis [10]. Gabor wavelets are plane waves restricted by a Gaussian envelope and can be characterized by the following equation:

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

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 that yielded better results with both our Fisherfaces baseline and the proposed algorithm: $N = 5$, $M = 8$, $\sigma = 2\pi$, $k_{max} = \pi/4$ and $f = \sqrt{2}$. For each image we normalized the feature coefficients to zero mean and unit variance which performed a divisive contrast normalization [10].

6.3. The Baseline: Fisherfaces

While Principal Component Analysis (PCA) is a dimension reduction technique which is optimal with respect to data compression, in general it is sub-optimal for recognition. For such a task, Fisher's Linear Discriminant (FLD) should be preferred to PCA. The idea of FLD is to select a subspace that minimizes the ratio of the inter-class variability and the intra-class variability. However, the straightforward application of this principle to face recognition is often impossible due to the high dimensionality of the feature space. A method called Fisherfaces was developed to overcome this issue [7]: one first applies PCA to reduce the dimension of the feature space and then performs the standard FLD.

For fair comparison, we did not apply directly Fisherfaces on the gray level images but on the Gabor features as done for instance in [9]. A feature vector was extracted every four pixels in the horizontal and vertical directions and the concatenation of all these vectors formed the Gabor representation of the face. In [9] various metrics were tested: the L_1, L_2 (Euclidean), Mahalanobis and cosine distances. We chose the Mahalanobis metric which consistently outperformed all other distances. The best Fisherfaces identification rates is 93.2% with 300 Fisherfaces.

6.4. Performance of the Novel Algorithm

To reduce the computational load, and for a fair comparison with Fisherfaces, the precision of a translation vector τ was limited to

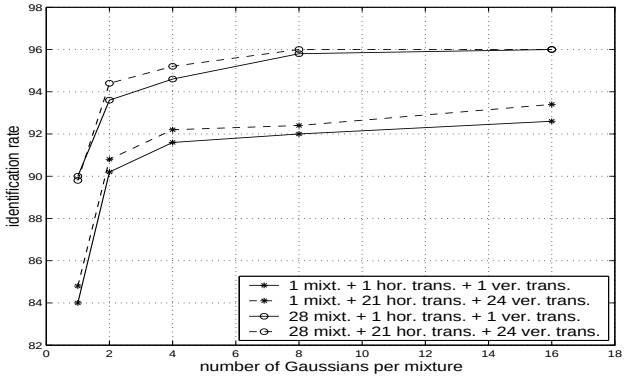

Fig. 3. Performance of the proposed algorithm.

4 pixels in both horizontal and vertical directions. Therefore, a vector m was extracted every 4 pixels of the query images as was done for Fisherfaces (dense grid). For each template image, a feature vector o was extracted every 16 pixels in both horizontal and vertical directions (sparse grid) which resulted in $7 \times 7 = 49$ observations per template image. We tried a smaller step size for template images but this resulted in marginal improvements of the performance at the expense of a much higher computational load.

To train single Gaussian mixtures, for each training couple $(\mathcal{F}_T, \mathcal{F}_Q)$ we first align approximately $\mathcal{F}_T$ and $\mathcal{F}_Q$, match each block in $\mathcal{F}_T$ with the corresponding block in $\mathcal{F}_Q$ and initialize Gaussian parameters. Transition probabilities are initialized uniformly. Then $\lambda_{\mathcal{M}}$ parameters are re-estimated using the modified Baum-Welch. To train multiple Gaussians per mixture we implemented an iterative splitting/re-training strategy.

We measured the impact of using multiple Gaussian mixtures to weight the different parts of the face and using multiple horizontal and vertical transitions matrices to model the elastic properties of the various parts of the face. In both cases, we used face symmetry to reduce the number of parameters to estimate. Hence, we tried one mixture for the whole face ($\Sigma^k_{i,j} = \Sigma^k$, $\delta^k_{i,j} = \delta^k$ and $w^k_{i,j} = w^k$) and one mixture for each position (using face symmetry, it resulted in $4 \times 7 = 28$ mixtures). We tried one horizontal and one vertical transition matrices for the whole face and one horizontal and one vertical transition matrices at each position (using face symmetry, it resulted in $3 \times 7 = 21$ horizontal and $4 \times 6 = 24$ vertical transition matrices). This made four test configurations. The performance was drawn on Fig. 3 as a function of the number of Gaussians per mixture.

While applying weights to different parts of the face provides a significant increase of the performance, modeling the various elasticity properties of the face had a limited impact and resulted in small consistent improvements. The best performance is 96.0% identification rate. Applying a simple Mc Nemar's test of significancy [11], we hence guarantee with more than 99% confidence that our approach performs significantly better than Fisherfaces.

We should underline that the approximation of $P(\mathcal{F}_T|\mathcal{F}_Q, \mathcal{M})$ based on T-HMMs is very efficient as, once the Gabor features are extracted from $\mathcal{F}_T$ and $\mathcal{F}_Q$, it takes only 15 ms to our best system with 16 GpM to compute the score on a Pentium IV 2 Ghz .

7. CONCLUSION

We presented a novel deformable model of the face and applied it successfully to face recognition. In our framework, the shared face deformation is approximated with 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 an approximation of the intractable 2-D HMMs: the Turbo-HMMs.

As the objective of this work was not modeling face deformation per se, but the face recognition problem, it is noteworthy that Maximum Likelihood Estimation is generally not optimal. It may be advantageous to train the HMM parameters under discriminative criteria such as the *Minimum Classification Error* (MCE) or its approximation via the *Maximum Mutual Information Estimation* (MMIE) criterion.

8. REFERENCES

[1] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. on Computers*, vol. 42, no. 3, Mar 1993.

[2] L. Wiskott, J. M. Fellous, N. Krüger and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 775–779, July 1997.

[3] B. Dûc, S. Fischer and J. Bigün, "Face authentication with gabor information on deformable graphs," *IEEE Trans. on Image Processing*, vol. 8, no. 4, Apr 1999.

[4] A. Tefas, C. Kotropoulos and I. Pitas, "Using support vector machines to enhance the performance of elastic graph matching for frontal face recognition," *IEEE Trans. on PAMI*, vol. 23, no. 7, pp. 735–746, Jul 2001.

[5] F. Perronnin, J.-L. Dugelay and K. Rose, "Iterative decoding of two-dimensional hidden markov models," in *ICASSP*, 2003, vol. 3, pp. 329–332.

[6] P. J. Phillips, H. Wechsler, J. Huang and P. Rauss, "The feret database and evaluation procedure for face recognition algorithms," *Image and Vision Computing Journal*, vol. 16, no. 5, pp. 295–306, 1998.

[7] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on PAMI*, vol. 19, pp. 711–720, Jul 1997.

[8] L. R. Rabiner, "A tutorial on hidden markov models and selected applications," *Proc. of the IEEE*, vol. 77, no. 2, Feb 1989.

[9] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. on Image Processing*, vol. 11, no. 4, pp. 467–476, Apr 2002.

[10] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman and T. J. Sejnowski, "Classifying facial expressions," *IEEE Trans. on PAMI*, vol. 21, pp. 974–989, Oct 1999.

[11] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition," in *ICASSP*, 1989, vol. 1, pp. 532–535.