Embedded Face and Facial Expression Recognition

Antonio Colmenarez
Adaptive Systems Department
Philips Research
Briarcliff Manor, NY 10510
USA

Brendan Frey
Department of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

Thomas S. Huang
Beckman Institute
University of Illinois
Urbana, IL 61801
USA

Abstract

A framework for embedded recognition of faces and facial expressions is described. Faces are modeled based on the appearances and positions of facial features. Hidden states are used to represent discrete facial expressions. A face model is constructed for each person in the database using video segments showing different facial expressions. Face recognition and facial expression recognition are carried out using Bayesian classification. In our current implementation, the face is divided into 9 facial features grouped in 4 regions which are detected and tracked automatically in video segments. We report results on face and facial expression recognition using a video database of 18 people and 6 expressions.

1 Introduction

The goal in most face recognition approaches is to find a similarity measure invariant to illumination changes, head pose and facial expressions so that images of faces can be successfully matched despite of these variations. Several approaches have been studied to deal with these forms of variations.

Elastic graphs (e.g. "Dynamic Link Architecture" [1, 2]) have been used to match images while allowing for non-rigid deformation. Another interesting approach uses a parametric model to deform faces so that the comparison of their appearance is carried out after the pose and facial expression is somewhat compensated [3, 4]. In a face recognition system based on a single frontal view, [5] uses the position of the facial features to estimate the head pose and normalize feature templates accordingly.

It is also worth noting that although many techniques exist for nonrigid facial motion analysis from video, little work have been done to use video for face recognition. Similarly, facial expression pattern have not been used as part of the discrimination criteria on face recognition algorithms. On one approach for person identification based on the analysis of the spatio-temporal patterns of the lips [6], it has been found that combining shape information with intensity information improves recognition accuracy.

Most expression recognition algorithms use models of nonrigid deformation patterns of facial expressions so that expressions can be successfully classified in a person-independent context. These deformation models are based on optical flow [7, 8], 2-D graphical models ("Potential Nets") [9] and local parametric models [10, 11]. Facial appearance variations due to facial expressions that are not well described by these deformation models are simply ignored.

In this paper, we proposed a Bayesian recognition framework in which the faces are modeled considering a discrete set of facial expression states. We use video sequences to train these face models, and use these models in a maximum likelihood setup to carry out face and facial expression recognition. The algorithm finds the face model and the facial expression state that maximize the likelihood of a test image or video sequence.

Additionally, we proposed a model of the global facial appearance based on a combination of individual models of the positions and appearances of facial features. Feature positions are normalized in scale and rotation using the positions of outer corners of the eyes. Individual feature appearances are normalized with the corresponding feature position and the face size. For example, feature appearance is taken as subwindows of size proportional to that of the face and centered at the position of the feature.

In this framework, face recognition is improved by (i) modeling individual feature appearances conditioned to the corresponding feature positions, and (ii) by learning patterns of feature appearances and positions for different facial expression states. Similarly, facial deformation patterns, captured by the positions

of facial features, are used together with changes in individual feature appearances to discriminate among facial expressions. Therefore, face recognition is improved by facial deformation matching, and facial expression recognition is improved by facial appearance matching.

2 Embedded Face and Facial Expression Recognition

Consider the face recognition problem where $\rho \in \{1, 2, \ldots, P\}$ is the index to ρ-th person in a database of P people and $\mathbf{f}$ is the portion of the observed image used for face recognition. Face recognition is carried out using maximum likelihood classification,

$$\rho^* = \arg \max_{\rho=1,\ldots,P} \mathbf{P}(\mathbf{f}|\rho), \tag{1}$$

by selecting the model that maximizes the likelihood probability of the observed image.

We use a hidden, discrete variable $\epsilon = 1, 2, \ldots, N$ to index the facial expressions. The likelihood probability of the image $\mathbf{f}$ given the identity class ρ is computed from

$$\mathbf{P}(\mathbf{f}|\rho) = \sum_{\epsilon=1}^{N} \mathbf{P}(\mathbf{f}|\epsilon, \rho)\mathbf{P}(\epsilon|\rho), \tag{2}$$

where $\mathbf{P}(\mathbf{f}|\epsilon, \rho)$ is the likelihood probability of the portion of the observed image given the facial expression state ϵ and the person identity ρ, and $\mathbf{P}(\epsilon|\rho)$ is the prior probability of the facial expression for that particular person.

Using this framework, the proposed embedded face and facial expression recognition procedure selects the person's model that maximizes the likelihood of the test images. Then, in a person-dependent scheme, the detected facial expression is simply that which maximizes the posteriori probability for that person's model,

$$\epsilon^* = \arg \max_{\epsilon=1,\ldots,N} \mathbf{P}(\mathbf{f}|\epsilon, \rho^*)\mathbf{P}(\epsilon|\rho^*). \tag{3}$$

Although not formally studied in this work, we believe that this architecture can be extended to handle person-independent expression recognition by combining the results obtained with models of different faces in a sufficiently large database of people.

2.1 Modeling Faces with Feature Appearances and Positions

We model faces with a set of regions containing facial features. We assume these facial feature regions $\{\mathbf{r}_i \;\; i = 1, 2, \ldots, R\}$ to be independent fron each other for a given person and facial expression, and compute

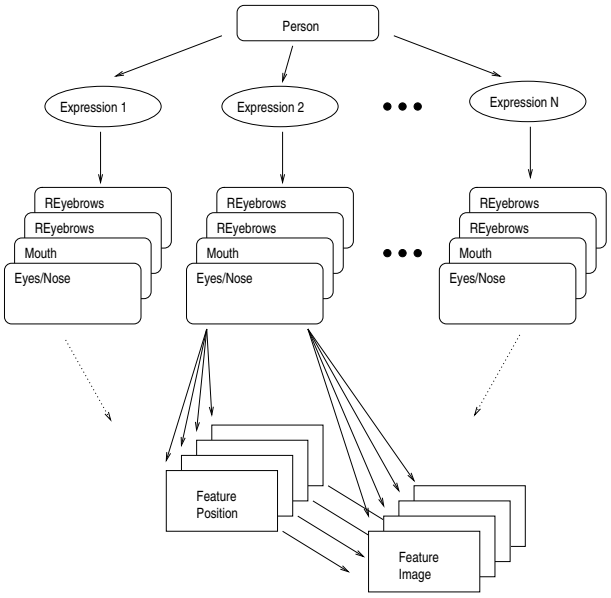

Figure 1: Probability network for embedded face and facial expression recognition

the likelihood probability of the observed image from

$$\mathbf{P}(\mathbf{f}|\epsilon, \rho) = \prod_{k=1}^{M} \mathbf{P}(\mathbf{r}_k|\epsilon, \rho). \tag{4}$$

We compute the likelihood probability $\mathbf{P}(\mathbf{r}_k|\epsilon, \rho)$ of a region based on the position $\mathbf{x}_{ki}$ and appearance $\mathbf{v}_{ki}$ of its F_k facial features as

$$\begin{aligned} &\mathbf{P}(\mathbf{r}_k|\epsilon, \rho) = \\ &\quad \mathbf{P}(\mathbf{v}_{k1}, \ldots, \mathbf{v}_{kF_k}|\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kF_k}, \epsilon, \rho) \\ &\quad \mathbf{P}(\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kF_k}|\epsilon, \rho). \end{aligned} \tag{5}$$

Figure 1 illustrates schematically this probability network. Note that ovals represent hidden states, rounded corner rectangles represent data structures, and sharp-cornered rectangles represent actual observed data.

Figure 2 illustrates the four facial regions and the nine facial features used in our implementation. These face models and recognition algorithms are based on the assumption that the facial features can be accurately detected and tracked in image sequences.

2.2 Modeling Facial Feature Appearances and Positions

Facial features appearances are modeled independently from each other, while positions of facial features are modeled jointly. This follows the idea that changes in local feature appearances are mostly uncorrelated, but facial deformation pattern within regions

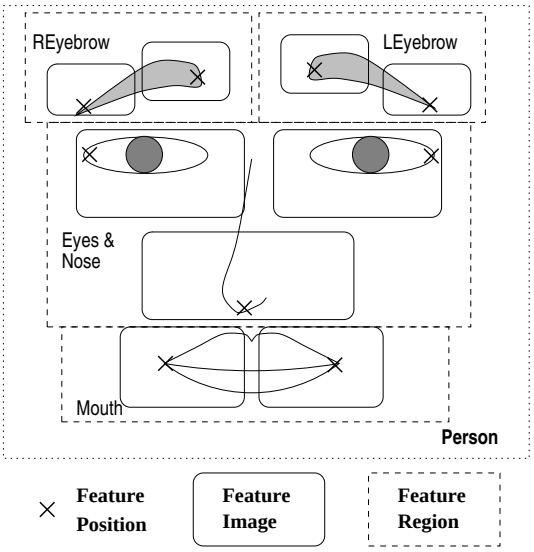

Figure 2: Scheme of the Facial Features and Regions.

are highly correlated as they are driven by higher-level facial expressions. Consequently, (5) turns into

$$\begin{aligned}&\mathbf{P}(\mathbf{r}_k|\epsilon,\rho) = \\ &\quad \mathbf{P}(\mathbf{v}_{k1}|\mathbf{x}_{k1},\epsilon,\rho)\ldots\mathbf{P}(\mathbf{v}_{kF_k}|\mathbf{x}_{kF_k},\epsilon,\rho) \\ &\quad \mathbf{P}(\mathbf{x}_{k1},\ldots,\mathbf{x}_{kF_k}|\epsilon,\rho).\end{aligned} \tag{6}$$

We model the positions of the facial features in each region, $\mathbf{P}(\mathbf{x}_{k1},\ldots,\mathbf{x}_{kF_k}|\epsilon,\rho)$, jointly using q-dimensional Gaussian distributions with full covariance matrices, where $q = 2F_k$.

The appearance of each facial feature is modeled with a p-dimensional Gaussian distribution applied over the principal component subspace. Note that this is different from the eigenface approach in which principal component analysis (PCA) is used to find the sub-space in which all object classes span the most. For simplicity in nomenclature, we use $\mathbf{P}(\mathbf{v})$ instead of $\mathbf{P}(\mathbf{v}_{k1}|\mathbf{x}_{k1},\epsilon,\rho)$ in the remaining of this section.

Let $\mathbf{v} \in \Re^d$ be a d-dimensional random vector with some distribution that we are to model. We use a set of training samples of the class in question to estimate the mean $\bar{\mathbf{v}}$ and the covariance matrix Ω of the class. Using singular value decomposition, we obtain the diagonal matrix Σ corresponding to the p largest eigenvalues of Ω and the transformation matrix $\mathbf{T}$ containing the corresponding eigenvectors.

The conditional probability of $\mathbf{v}$ for a given class is then computed from

$$\begin{aligned}&\mathbf{P}(\mathbf{v}) = \\ &\quad \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp\left[-\tfrac{1}{2}\mathbf{u}'\Sigma^{-1}\mathbf{u}\right] \\ &\quad \frac{1}{\sqrt{2\pi\lambda}} \exp\left[-\frac{\mathbf{r}^2}{2\lambda}\right],\end{aligned} \tag{7}$$

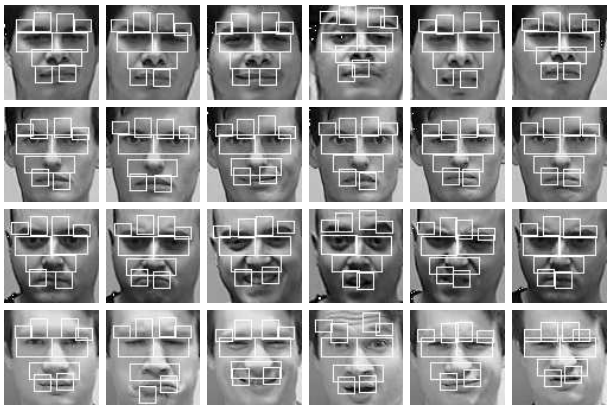

Figure 3: Examples of the facial feature windows and of the six facial expressions.

where $\mathbf{u} = \mathbf{T}(\mathbf{v} - \bar{\mathbf{v}})$ is the projection of $\mathbf{v}$ onto the aforementioned p-dimensional subspace, $r = \sqrt{|\mathbf{v} - \bar{\mathbf{v}}|^2 - |\mathbf{u}|^2}$, is the distance from the subspace, and λ is obtained from the sum of the remaining eigenvalues of Ω.

3 Experiments and Results

We have tested our face and facial expression recognition algorithm using images from a video database of head-and-shoulder scenes. This database consists of 54 video segments containing 18 people with 3 distinct video segment each.

The first segment, of approximately 2600 frames, consists of three repetitions of six facial expressions in the following sequence: neutral, sadness, neutral, happiness, neutral, surprise, neutral, disgust, neutral, anger, neutral. The second set of video segments, of 2020 frames, consist of repetitions of facial gestures while showing three facial expressions: neutral, happiness and anger. The facial gestures include head nodding, head shaking, eye winking, and others. The last set of videos consist of 1350 frames of more relaxed gestures, expressions, and head motion intended for testing face tracking and recognition under extreme facial expressions.

We use our face and facial feature detection and tracking system [12, 13, 14] to locate nine facial features on each frame of these video sequences. The nine facial features detected are two outer eye corners, four eyebrow corners, the center of the nostrils, and two mouth corners.

We trained one model for each person using the first two thirds of the frames of the first set of video segments in our video database and left the rest of the frames for testing our face and facial expression recognition algorithm.

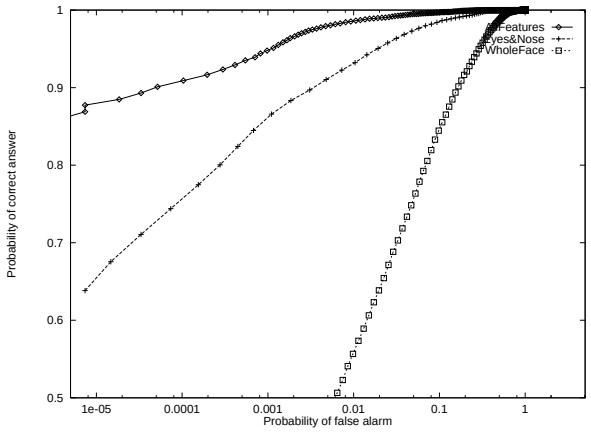

Figure 4: Face recognition performance comparison using facial appearance models of: (i) the whole face, (ii) eye and nose region, (iii) individual facial features

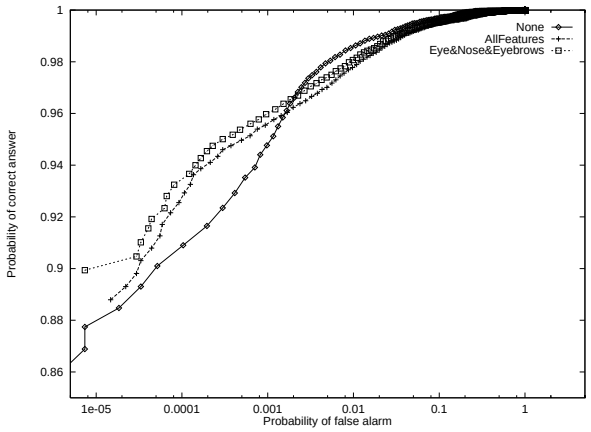

Figure 5: Performance improvement in face recognition by using the facial feature position models in addition to the facial feature appearance models.

Each model, as illustrated in Figure 2, consists of four regions of features and six facial expressions: neutral, sadness, happiness, surprise, disgust and anger. These facial expressions were labeled by hand in the training video segments. Figure 3 shows examples of these six facial expressions and of the image windows of the facial features (e.g. $\mathbf{v}_{ki}$).

For the purpose of comparison, we also trained a simpler version of these face models in which the facial appearance was modeled using images of the whole face. The position of the outer eye corners were used to normalize faces in size and position. These simpler models include hidden states for facial expressions, but exclude any other feature position information.

Face and facial expression recognition was carried out using maximum likelihood decisions. We approximate the likelihood of the face (2) with that of the facial expression that is most likely. So, the face recognition algorithm selects the person's model and the facial expression that maximizes the likelihood of the test image. In a person dependent scheme, the expression recognition algorithm simply selects the facial expression with maximum posterior probability.

Figure 4 shows the ROC's for face recognition systems using (i) the appearance of the whole face, (ii) the appearance of the eye and nose region, and (iii) the combination of the appearance of the facial features: eyes, nose, eyebrows, and mouth. Note how excluding the nonrigid parts of the faces, i.e. the eyebrows and the mouth improve recognition performance. Also note how our face model improve recognition performance even further by using local appearances of all facial features.

Figure 5 shows the improvement in face recognition performance obtained by considering the facial feature position models in addition to the facial feature appearance models. Note that facial deformation patterns do help to discriminate among people.

Table 1 shows the facial expression recognition performance obtained by using the combination of facial feature position and appearance models.

We also measure the expressiveness of the facial feature regions by comparing the recognition performance of each of the facial expressions independently. In the interest of space, detailed results are not provided. Readers are asked to contact the authors.

4 Concluding Remarks

We have presented a Bayesian framework in which modeling facial expression states allow cooperation between face recognition and facial expression recognition improving the overall performance of the system. We have also shown that modeling faces with individual models of the facial feature appearances and facial feature positions improve face and facial expression recognition performance by: (i) increasing robustness in facial appearance matching against facial expression deformation, (ii) considering facial expression deformation patterns for person identification, and (iii) considering both facial feature appearances and positions for expression recognition.

We are currently testing an extension of this framework that accounts for the temporal dependencies between facial expressions. The hidden variable ϵ used to index the facial expression in (2) is replaced by a hidden state in a HMM. Face and facial expression recognition from video segments is then performed using

	Neutral	Sadness	Happiness	Surprise	Disgust	Anger
Neutral	7049	64	51	33	60	28
Sadness	337	633	3	6	44	47
Happiness	292	16	932	9	6	3
Surprise	125	2	6	700	3	0
Disgust	301	21	0	5	661	54
Anger	272	28	8	5	58	664
Error (%)	15.8	17.2	6.8	7.6	20.5	16.5

Table 1: Expression recognition performance using feature positions and images

either the Viterbi algorithm or the forward-backward algorithm.

Acknowledgments

This work was supported by Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0003. Brendan Frey was a Beckman Fellow at the time this research was conducted.

References

[1] L. Wiskott, J. M. Fellous, N. Krger, and C. Malsburg, "Face recognition and gender determination," in *International Conference on Automatic Face and Gesture Recognition*, 1995.

[2] A. Tefas, C. Kotropoulos, and I. Pitas, "Variants of dynamic link architecture based on mathematical morphology for frontal face authentication," in *CVPR*, 1998.

[3] A. Lanitis, C. Taylor, and T. Cootes, "A unified approach to coding and interpreting face images," in *Proc. IEEE Conf. Computer Vision*, 1995.

[4] A. Lanitis, C. Taylor, and T. Cootes, "Automatic identification of human faces using flexible appearance models," in *Procs. Of the 5th Brithish Machine Vision Conference*, 1994.

[5] K.-M. Lam and H. Yam, "An analytic-to-holistic approach for face recognition based on a single frontal view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 673–686, July 1998.

[6] J. Luettin, N. Thacker, and S. W. Beet, "Learning to recognize talking faces," in *Procs. of Int. Conf. On Pattern Recognition*, 1996.

[7] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 636–642, June 1996.

[8] Y. Yacoob and L. Davis, "Computing spatio-temporal representations of human faces," in *IEEE International Conference on Computer Vision*, (Cambridge, MA), IEEE Computer Society Press, June 1995, pp. 70–75.

[9] K. Matsuno, C.-W. Lee, S. Kimura, and S. Tsuji, "Automatic recognition of human facial expressions," in *IEEE International Conference on Computer Vision*, (Cambridge, MA), IEEE Computer Society Press, June 1995, pp. 352–358.

[10] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions," Xerox Palo Alto Research Center, Tech. Rep. CS-TR-3401, Jan. 1995.

[11] M. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Proc. IEEE Int. Conf. Computer Vision*, 1995.

[12] A. Colmenarez, "Facial analysis from continuous video with application to human-computer interface," Ph.D. dissertation, University of Illinois at Urbana-Champaign, March 1999.

[13] A. Colmenarez, B. Frey, and T. Huang, "Detection and tracking of faces and facial features," in *ICIP*, 1999.

[14] A. J. Colmenarez and T. S. Huang, "Pattern detection with information-based maximum discrimination and error bootstrapping," in *Proc. ICPR*, 1998.