Short Papers

Automatic Classification of Single Facial Images

Michael J. Lyons, Julien Budynek, and Shigeru Akamatsu

Abstract—We propose a method for automatically classifying facial images based on labeled elastic graph matching, a 2D Gabor wavelet representation, and linear discriminant analysis. Results of tests with three image sets are presented for the classification of sex, "race," and expression. A visual interpretation of the discriminant vectors is provided.

Index Terms—Computer vision, face recognition, facial expression recognition, Gabor wavelets, principal component analysis, discriminant analysis.

1 Introduction

THE human face plays a central role in social interaction, hence it is not surprising that automatic facial information processing is an important and highly active subfield of pattern recognition research [5]. The face displays a complex range of information about identity, age, sex, "race"[1] well as emotional and attentional state. This paper focuses on the problem of extracting these semantic-level attributes of an individual face from single digital images. The examples chosen to demonstrate our method are facial expression, sex, and "race," however, the technique may extend to other facial attributes.

The method proposed in this paper synthesizes aspects of two major approaches to facial image processing: Gabor-wavelet-labeled elastic graph matching [9], [12] and "eigenface" or "Fisherface" algorithms [11], [1], [8] based on statistical representation of face space. Both the eigenface and Fisherface techniques require precise normalization and registration of facial internal features. Moreover, performance of the eigenface algorithm is improved by morphing the face to a standard shape [3]. By contrast, with this algorithm, a graph structure is registered approximately with the head. The features used, based on the 2D Gabor wavelet transform, are a compromise in the trade-off between spatial and spatial frequency domain accuracy [4] and are robust to small changes in the grid node positions. A pixel-based input representation, such as is used in previous work on eigenfaces, is not as robust to errors in registration. This is the major novel feature of the algorithm we describe: It combines the advantages of the Gabor wavelet representation with the ability to train the system simply and quickly from examples in a manner similar to the Fisherface algorithm.

1. The term "race" is used in quotation marks in this article to indicate a restricted usage which is specified in Section 2.2. More generally, the definition of the term race is complex and controversial.

- *M.J. Lyons and S. Akamatsu are with ATR Human Information Processing Research Labs, 2-2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan. E-mail: {mlyons, akamatsu}@hip.atr.co.jp.*
- *J. Budynek is with EuroBios, Immeuble Jean Monnet, 11 place des Vosges-La Défense 5, 92061 La Défense Cedex, France. E-mail: julien.budynek@polytechnique.org.*

Manuscript received 12 May 1999; revised 15 Nov. 1999. Recommended for acceptance by D. Kriegman. For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 109795.

2 Face Classification Algorithm

The algorithm can be divided into two broad steps: registration of a grid with the face and face classification based on feature values extracted at grid points. In this paper, facial grids are registered either automatically, using labeled elastic graph matching [9], [12] (as in Section 3, which describes a live demo system), or by manually clicking on points of the face (as in Sections 4 and 5, which describes basic research on facial expression recognition). This paper is concerned with face classification after the grid has been registered and the algorithm may be adapted for use with other grid registration schemes. Labeled elastic graph matching has been described in detail in the papers cited and will not be discussed in depth here.

Images are first transformed using a multiscale, multiorientation set of Gabor filters (Fig. 1). The grid is then registered with the face. Two types of grid are considered in this paper: a rectangular grid (Section 3) and a fiducial grid with nodes located at easily identifiable landmarks of the face (Sections 4 and 5). The amplitude of the complex valued Gabor transform coefficients are sampled on the grid and combined into a single vector, the labeled graph vector (or LG vector in Fig. 1). The ensemble of LG vectors from a training set of images are subjected to principal components analysis (PCA) to reduce the dimensionality of the input space. LG vectors project into the lower dimensional PCA space (LG-PCA vectors). The ensemble of LG-PCA vectors from the training set are then analyzed using linear discriminant analysis (LDA) in order to separate vectors into clusters having different facial attributes. Input vectors in the original LG space may then be analyzed using the same LDA to determine their attributes.

2.1 Two-Dimensional Gabor Wavelet Representation

Use of the 2D Gabor wavelet representation in computer vision was pioneered by Daugman in the 1980s [4]. More recently, von der Malsburg's group has developed a face recognition system making use of this representation [9], [12].

A complex-valued 2D Gabor function is a plane wave restricted by a Gaussian envelope:

$$\Psi(\mathbf{k},\mathbf{x}) = \frac{\mathbf{k}^2}{\sigma^2}\exp\left(-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right)\left[\exp(i\mathbf{k}.\mathbf{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right]. \quad (1)$$

The multiplicative factor $\mathbf{k}^2$ ensures that filters tuned to different spatial frequency bands have approximately equal energies. The term $\exp(-\sigma^2/2)$ is subtracted to render the filters insensitive to the overall level of illumination. The Gabor wavelet representation allows description of spatial frequency structure in the image while preserving information about spatial relations. The complex amplitude of the transforms is used as features to test for the presence of spatial structure, restricted to a band of orientations and spatial frequencies, within the Gaussian envelope. The amplitude information degrades gracefully with shifts in the image location at which it is sampled, over the spatial scale of the envelope.

For the 256×256 images used in the experiments below, five spatial frequencies were used, with $k_i = \pi/2^i$ and i from 1 to 5. Six angular orientations (from 0 to 150 degrees in 30 degree steps) were used. For all experiments, $\sigma = \pi$, setting the bandwidth of the filters to roughly one octave in spatial frequency. Input images are convolved with the Gabor filters and the magnitudes of the complex-valued filter responses are sampled at points on the facial grid and combined into a single LG vector. For the experiments in Section 3 (grid of 49 nodes), the LG vector is of dimension 1,470, whereas, for Sections 4

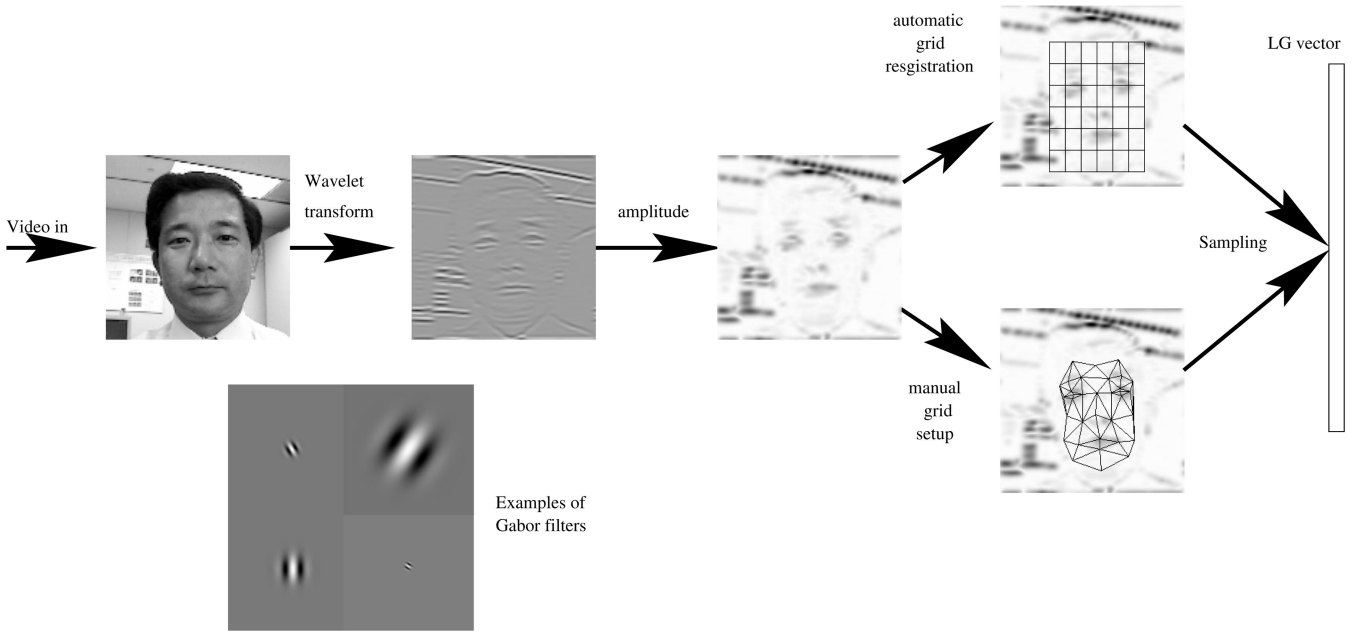

Fig. 1. The Gabor-labeled elastic graph representation of a facial image.

and 5 (grid of 34 nodes), it is of dimension 1,020.

2.2 Discriminant Analysis

The examples we consider can be treated using two class discriminant analysis, e.g., male or female. For facial expression, the presence or absence of each expression is tested and the outcomes used to classify the expression.[2] Application of a binary classifier to "race" is possible only because our training and test sets consist of faces which are clearly identifiable as either "East Asian" or "non-East Asian."

Two-class linear discriminant analysis seeks a single projection optimally separating the two labeled clusters in the training set, while minimizing variance within each projected cluster. A complete description of LDA may be found in Duda and Hart [6], whose notation we preserve here. Consider a set of n d-dimensional vectors $\mathbf{x_1}, \ldots \mathbf{x_n}$, with n_1 vectors in the set $\mathcal{X}_1$ and n_2 in the set $\mathcal{X}_2$. The projection of the sample $\mathbf{x}$ onto direction defined by vector $\mathbf{w}$ is $y = \mathbf{w^t x}$. The *scatter* of the projected vectors is defined as:

$$\tilde{s}_i^2 = \sum_{\mathbf{y} \in Y_i} (\mathbf{y} - \tilde{\mathbf{m}}_\mathbf{i})^2,$$

where $\tilde{\mathbf{m}}_\mathbf{i}$ is the mean of the projected samples of set i. Scatter *within-cluster* and *between-cluster*, are defined, respectively, as:

$$S_W = \sum_{i=0}^{1} \sum_{\mathbf{x} \in X_i} (\mathbf{x} - \mathbf{m_i})(\mathbf{x} - \mathbf{m_i})^\mathbf{t}, \quad S_B = (\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^\mathbf{t}$$

S_B, as the outer product of two vectors, has rank of at most one and, for any $\mathbf{w}$, $S_B\mathbf{w}$ is in the direction of $\mathbf{m_1} - \mathbf{m_2}$.

We seek a projection direction, defined by vector $\mathbf{w}$, along which the ratio of the between class scatter to the within class scatter, $J(\mathbf{w})$ is maximized,

$$J(\mathbf{w}) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_W \mathbf{w}}. \tag{2}$$

2. Our preliminary experiments on facial expression classification using multiclass LDA show significantly lower classification rates.

A vector maximizing J over $\mathbf{w}$ must satisfy the generalized eigenvalue problem, $S_B\mathbf{w} = \lambda S_W \mathbf{w}$. Since $S_B\mathbf{w}$ is in the direction of $\mathbf{m_1} - \mathbf{m_2}$ then $\mathbf{w} = S_W^{-1}(\mathbf{m_1} - \mathbf{m_2})$. Hence, for the two class problem, one need not solve the full generalized eigenvalue system.

The number of training images, typically of order 10^2, is smaller than the input dimensionality of the LG vectors, which is roughly 10^3. Therefore S_W is singular. In analogy with the Fisherface method, the data set is first projected into a lower dimensional space found using principal components analysis (PCA), then LDA is applied. Input LG vectors are first transformed by subtracting the mean: $\mathbf{\Phi_i} = \mathbf{x_i} - \mathbf{m}$. The principal components of the training data set are given by the eigenvectors of its covariance matrix, $C = \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Phi_i}\mathbf{\Phi_i}^\mathbf{t}$. Because of the high dimensionality of the LG vectors, C is very large; however, there are only $n-1$ nonzero eigenvalues and only the corresponding eigenvectors are relevant for describing the distribution of the training set. In practice, only N eigenvectors having the largest eigenvalues (and, hence, the largest variance in the data set) are kept and the discriminant analysis can be performed in a space having smaller dimension N, in which the within-class scatter matrix is nonsingular. If W_{pca} is the matrix of eigenvectors having the N largest eigenvalues (W_{pca} is of dimension $d \times N$), (2) becomes

$$J(\mathbf{w}) = \frac{\mathbf{w}^t W_{pca}^t S_B W_{pca} \mathbf{w}}{\mathbf{w}^t W_{pca}^t S_W W_{pca} \mathbf{w}}.$$

If every eigenvector with nonzero eigenvalue is included in W_{pca}, then the within-class scatter of projected training samples can be reduced to zero. Including too many of the $n-1$ eigenvectors in the LDA analysis results in overfitting to the training set and no improvement to or, in some cases, worsening of the generalization rate. The number of retained eigenvectors was chosen empirically to optimize generalization performance.

2.3 Image Classification

To classify an input LG vector, it is projected along the corresponding discriminant vector calculated from training examples. The distance to each cluster center is calculated, normalized by the standard deviation, $\tilde{\sigma}_j$, of the projected cluster

Fig. 2. Examples of facial images acquired during a live demo of the system at the annual ATR Open House.

$$d_j = \frac{(\mathbf{x} - \mathbf{m}) \cdot \mathbf{w} - \tilde{\mathbf{m}}_j}{\tilde{\sigma}_j}, \tag{3}$$

where $j \in \{0, 1\}$ for the two clusters. The input sample is classified as a member of the nearest cluster.

3 Experiments: Sex, "Race," and Expression

A classifier was trained to categorize face images according to sex: male/female, "race": east-Asian/other, expression: smile/other. The image set for these experiments was acquired during a live demo held at ATR on 5 and 6 November 1998. The faces are, in almost all cases, easily recognizable as either east-Asian or not. The demo is fully automatic, positioning a 7×7 rectangular grid on the face using our local implementation of the elastic graph matching algorithm [9], [12]. The facial registration grid has only four parameters, the x and y coordinates of the center-of-mass and the horizontal and vertical grid line spacing. The image set includes a total of 182 images, consisting of 106 male faces (76 females), 135 East Asian faces (47 non-East Asian), and 84 smiles (98 nonsmiles). Sample images are shown in Fig. 2, which displays a typical range of variation in the image conditions and grid position. The system was periodically retrained as more images were acquired. The entire procedure was carried out independently for each of the three LDA projections. As the number of images in the training set increased, the correct classification rate also increased. Fig. 3 shows the performance of the three classifiers for various training set sizes. In these experiments, the system was trained on all samples of the training set but one identity and then tested on that person's images. This "leave-one-out" procedure was repeated for each identity and the results averaged. Generalization performance was 91 percent for expression, 95 percent for "race," and 92 percent for sex recognition. The slight decrease in performance for the largest image sets may be due to a statistical fluctuation or a change in image acquisition conditions during the course of the live demonstration, which took place over the space of two consecutive working days, as no special care was taken to keep the conditions constant.

4 Experiments: Facial Expression

Six binary classifiers, one for each of the six fundamental facial expressions (happy, sad, angry, fearful, surprised, disgusted), were trained independently and combined to build a facial expression categorizer. For an input image that is positively classified for two or more expressions, the normalized distance, (3), to the cluster centers is used as a deciding factor. An input image that is not positively classified for any category is categorized as neutral. These procedures are appropriate for the expression

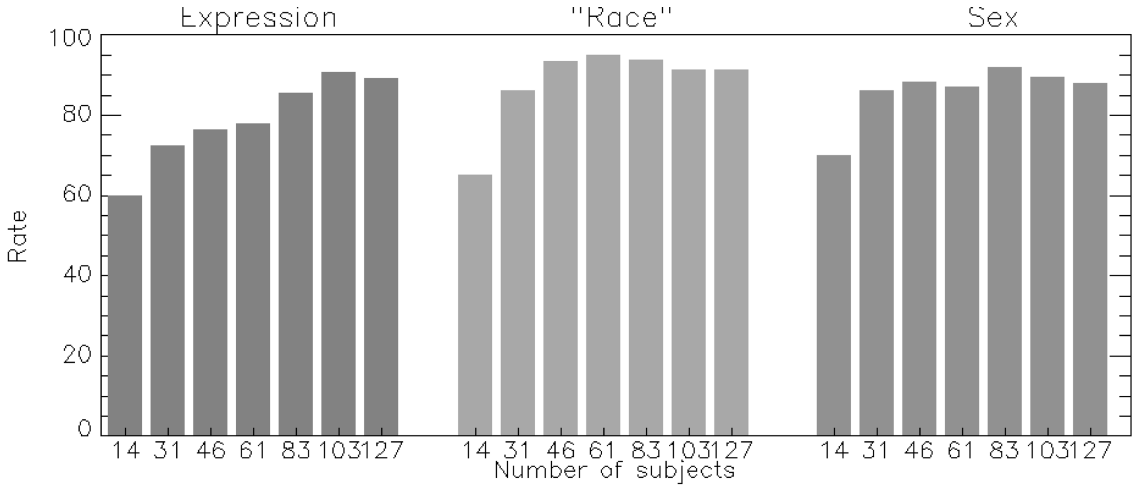

Fig. 3. Generalization rates of the algorithm on three classification tasks as the number of training images is increased.

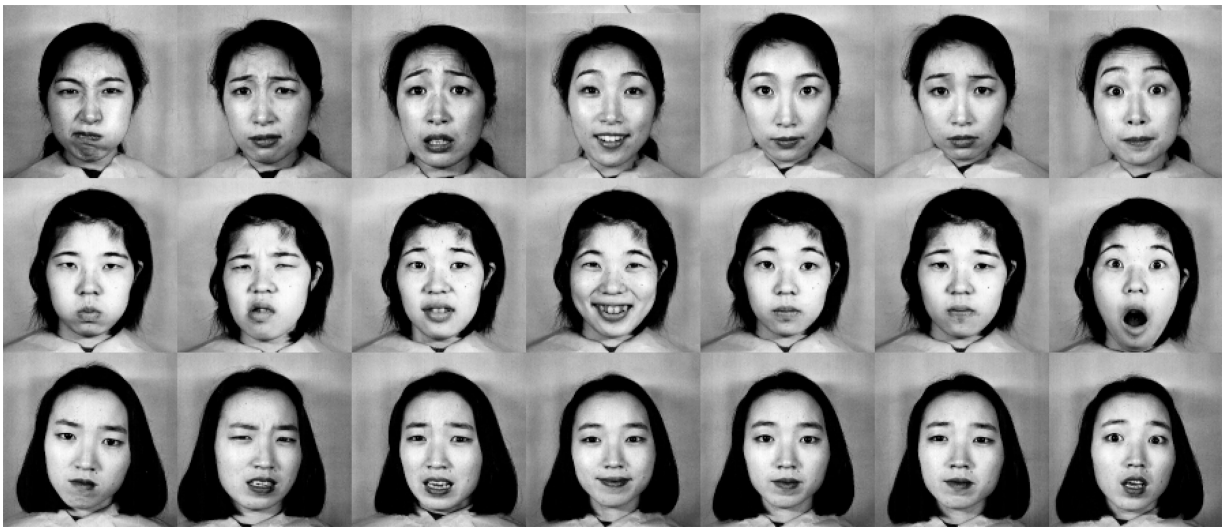

Fig. 4. Samples from the Japanese females facial expression image set.

databases here. However, for mixed-expression data, an alternate approach may be more suitable.

Because some of the facial expressions are distinguished by subtle changes in facial configuration (e.g., disgust and anger), we reasoned that more spatial accuracy in grid point registration might be necessary for this task. A fiducial grid (shown in Fig. 1) was positioned by manually clicking on 34 easily identifiable points of each facial image.

The expression classifier was first tested using a set of 193 images of expressions posed by nine Japanese females, which has been used in two previous studies of facial expression recognition [10], [13]. Each expresser posed three or four examples of each of the six fundamental facial expressions and a neutral face. Sample images from the set are shown in Fig 4. In the first study, we used the same testing paradigm as Zhang et al. [13]. The entire set of images was divided into 10 segments; the discriminant vectors were calculated using nine of these segments and the generalization performance tested on the remaining segment, with the results averaged over all 10 distinct partitions. Fig. 5 shows the results plotted as a function of the number of eigenvectors retained before LDA. The generalization rate for this system is 92 percent. To measure generalization over identity, the the image set was partitioned into nine segments, each corresponding to one expresser. The system was trained on eight of the segments and then tested on the ninth. This was repeated for all nine possible partitions of training and testing data and the results were averaged. The average generalization rate for recognition of expression for a novel expresser was 75 percent.

The system was also tested using the facial expression image set of Ekman and Friesen [7], consisting of 110 images, of which 51 are male and 59 are female. The system has a peak generalization rate of 82 percent tested on all expression categories. Not all expressions were equally well recognized by the system. Table 1 shows a confusion matrix showing misclassification rates for expressions.

5 Saliency Maps

The algorithm requires no explicit specification of which parts of the facial image are pertinent to the classification process. It is interesting to ask which aspects of the input data are most useful for characterizing faces. The magnitude of each component of the discriminant vector determines its influence on the classification decision and is, therefore, a measure of the saliency of the corresponding feature. The left side of Fig. 6 displays discriminant vector component magnitude averaged over all frequencies and orientations and facial expressions at each fiducial point on the face. The size of each filled circle is proportional to the discriminant vector component magnitude. The figure shows that

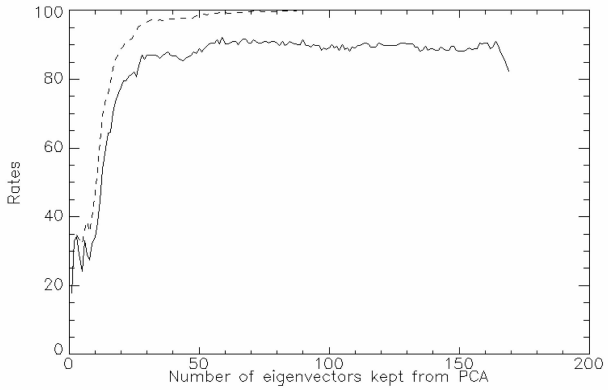

Fig. 5. Generalization (solid) and training set (dashed) performance rate for the facial expression classifier, as tested using the Japanese female facial expression image set.

TABLE 1
Confusion Matrix for the Facial Expression Classifier as Measured on Ekman Image Set

ang	dis	fea	hap	sad	sur	I/O
88	13	13	6	6	0	ang
0	73	0	6	6	0	dis
0	0	73	6	0	8	fea
0	0	0	78	0	0	hap
13	13	0	0	81	0	sad
0	0	0	0	0	85	sur
0	0	13	6	6	8	neu

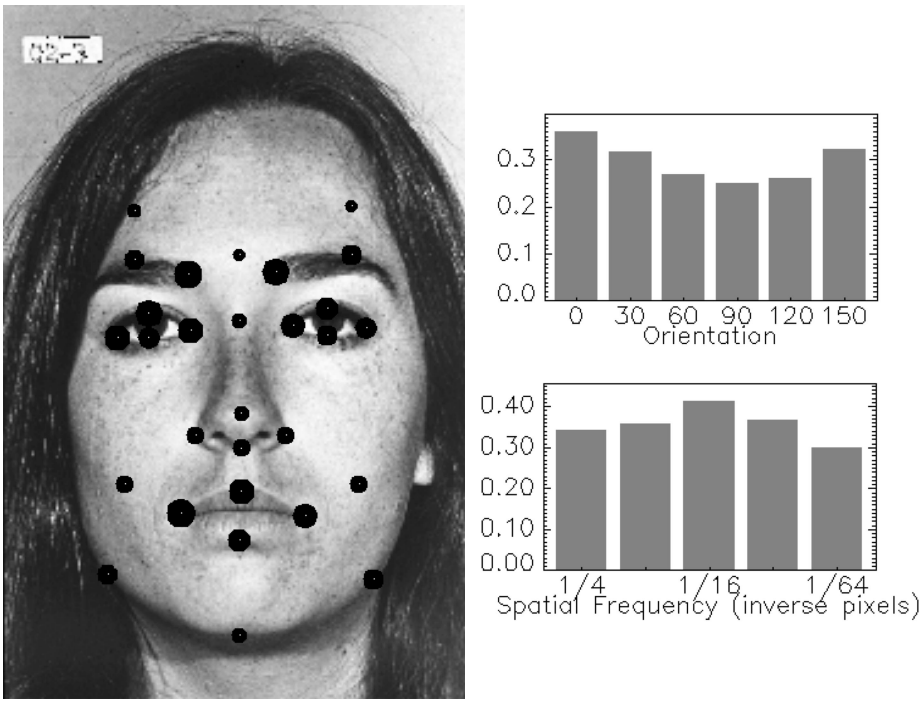

Fig. 6. Node, orientation, and frequency saliencies.

the eyes and mouth are the most critical areas of the face for determining facial expressions. The upper right of Fig. 6 plots discriminant vector component magnitude averaged over all frequencies, locations on the fiducial graph, and expressions as a function of spatial orientation. The graph shows that horizontally oriented filters are the most useful for recognizing facial expressions and vertical orientations are the least useful. The lower right of Fig. 6 plots discriminant vector component magnitude averaged over orientation, location, and expression as a function of spatial frequency.

6 Discussion

This paper presented a novel algorithm for automatically extracting semantic-level information about faces from digital images. The algorithm synthesizes aspects of two major streams of research on face processing: the labeled elastic graph matching approach of the von der Malsburg group [9], [12] and Eigenface/Fisherface algorithms [11], [1], [8].

Experiments with an automatically positioned rectangular grid on images taken under live conditions showed that the system was quite robust to shifts in node position, maintaining generalization rates that exceed 90 percent for sex, "race," and discrimination of two expressions. This compares favorably with previous results using other single-image methods on similar binary classification tasks, summarized in Wiskott et al. [12].

A further advantage of the simple LDA-based classification scheme is that it can be trained quickly. During a live demonstration, the system was retrained nearly each time a new image was acquired. Completely retraining the system from scratch is also straightforward so that it can be rapidly adapted to other binary classification tasks or the local requirements of the implementation.

Sections 4 and 5 tested the algorithm on a finer discrimination task: recognition of the 6 basic facial expressions. The generalization rate of the expression recognizer for the Japanese female image set was 92 percent, essentially the same as the 90 percent obtained with a multilayer perceptron in the study of Zhang et al., suggesting that the linear LDA algorithm is sufficiently powerful for this classification task. However, fewer hidden units were used by the nonlinear perceptron to attain this generalization rate. The average generalization rate over expressor identity was 75 percent for the Japanese female image set and 82 percent for the Ekman pictures. This is still remarkably high considering the classifier has only about 10 individuals in the training set to learn which featural changes are due to identity and which are due to expression. The rate is not significantly different from the 86 percent reported by Padgett and Cottrell [2] on the same set of images, using principle components analysis and a multilayer perceptron classifier. In their work, however, input images were manually cropped and registered before analysis. Higher generalization rates for expression recognition might be obtained if the number of individuals in the training set was increased, as in Fig. 3 for the smile/nosmile classifier.

The current algorithm is limited in that it may only be used to extract categorical information about faces and neglects any information that cannot be treated by multiple binary classification. Our algorithm is also insensitive to color, which is often present in single images, though often unreliable because of the difficulty of accurate camera calibration. It would be interesting to explore the utility of color information for face classification. A further deficiency of the algorithm is that not all expressions are recognized equally well (see Table 1). However, this may be intrinsic to the facial expression recognition problem itself.

The saliency information displayed in Fig. 6 shows that the regions around the eyes and mouth are more important than other areas of the face for classifying the facial expressions. Filters of intermediate spatial frequency were found to be slightly more informative for expression classification. Notably, filters having horizontal orientation were more heavily weighted in the discriminant vector than other orientations. This seems intuitively correct since the most noticeable expressive motions of the face are the opening and closing of the mouth and eyes and raising and lowering of the eyebrows. Displacement of roughly horizontal edges forms the largest component of these motions.

ACKNOWLEDGMENTS

The authors thank Miyuki Kamachi (ATR) and Jiro Gyoba (Tohoku University) for permission to use the facial expression database and Sebastien Jehan, a student intern from the Institut National des Télécommunications, France, for contributions to the implementation of the live demo.

REFERENCES

[1] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherface: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[2] C. Padgett and G. Cottrell, "Identifying Emotion in Static Face Images," *Proc. Second Joint Symp. Neural Computation,* vol. 5, pp. 91-101, La Jolla, Calif., 1995.

[3] I. Craw, N. Costen, T. Kato, G. Robertson, and S. Akamatsu, "Automatic Face Recognition: Combining Configuration and Texture," *Proc. Int'l Workshop Automatic Face and Gesture Recognition,* M. Bichsel, ed., pp. 53-58, 1995.

[4] J.G. Daugman, "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *J. Optical Soc. Am. A,* vol. 2, pp. 1,160-1,169, 1985.

[5] J. Daugman, "Face and Gesture Recognition: An Overview," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 675-676, July 1997

[6] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.

[7] P. Ekman and W.V. Friesen, "Pictures of Facial Affect," Human Interaction Laboratory, Univ. of California Medical Center, San Francisco, 1976.

[8] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *J. Optical Soc. Am. A,* vol. 14, no. 8, pp. 1,724-1,733, 1997.

[9] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Wurtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers,* vol. 42, no. 3, pp. 300-311, Mar. 1993.

[10] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," *Proc. Third IEEE Conf. Face and Gesture Recognition,* pp. 200-205, Nara, Japan, Apr. 1998.

[11] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[12] L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg, "Face Recognition and Gender Determination," *Proc. Int'l Workshop Automatic Face and Gesture Recognition,* M. Bichsel, ed., pp. 92-97, 1995.

[13] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron," *Proc. Third IEEE Conf. Face and Gesture Recognition,* pp. 454-459, Nara, Japan, Apr. 1998.