Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition ,28-30 March, 2000, Grenoble France, IEEE Computer Society, pp. 202-207. See also: http://www.mic.atr.co.jp/~mlyons

Classifying Facial Attributes using a 2-D Gabor Wavelet Representation and Discriminant Analysis

Michael J. Lyons, Julien Budynek, Andre Plante†, Shigeru Akamatsu
ATR Human Information Processing Research Labs
†ATR Media Integration & Communication Labs
2-2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
mlyons@hip.atr.co.jp

Abstract

A method for automatically classifying facial images is proposed. Faces are represented using elastic graphs labelled with with 2-D Gabor wavelet features. The system is trained from examples to classify faces on the basis of high-level attributes, such as sex, ”race”, and expression, using linear discriminant analysis (LDA). Use of the Gabor representation relaxes the requirement for precise normalization of the face: approximate registration of a facial graph is sufficient. LDA allows simple and rapid training from examples, as well as a straightforward interpretation of the role of the input features for classification. The algorithm is tested on three different facial image datasets, one of which was acquired under relatively uncontrolled conditions, on tasks of sex, ”race” and expression classification. Results of these tests are presented. The discriminant vectors may be interpreted in terms of the saliency of the input features for the different classification tasks, which we portray visually with feature saliency maps for node position as well as filter spatial frequency and orientation.

1 Introduction

The human face displays a range of information relevant to social interactions with other human beings. The ability to process this information automatically can aid technology-mediated human-human communication. The largest number of studies of facial information processing by machines, concern the recognition or verification of facial identity. However, other types of information displayed by the face such as age, sex, emotional and attentional state, and other characteristics are also important. This paper concerns the extraction such symbolic-level attributes of an individual face from 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 we propose here synthesizes aspects of two major approaches to facial image processing: Gabor-wavelet-labelled elastic graph matching [8, 12], and “Fisherface” algorithms [1, 7] based on statistical representation of face space. Both the eigenface and the more recent Fisherface techniques require precise normalization and registration of facial internal features. Performance of the eigenface algorithm is improved by morphing the face to a standard shape [3]. By contrast, with the algorithm proposed in this paper a graph structure is registered approximately with the head. The input features, the complex amplitude of the 2-D Gabor wavelet transform, are not as sensitive to position as individual pixel grayvalues hence are can tolerate more error in registration of the grid. The novel feature of the algorithm proposed in this paper is that it combines the stability 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.

2 The Algorithm

The following is a summary of the steps in the algorithm. Individual steps are detailed in the subsections below and shown schematically in figures 1 and 2. Images are transformed using a multi-scale, multi-orientation set of Gabor filters. A grid is then automatically registered with the face using a variant of the elastic graph matching method [8]. 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). 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 projected into the lower dimensional PCA space (LG-PCA vectors). The ensemble of LG-PCA vectors from the training set are then

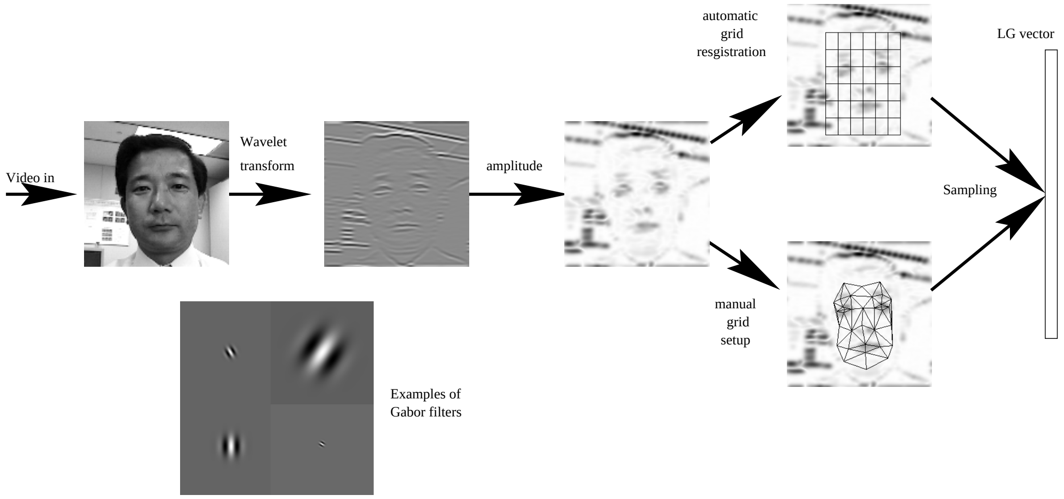

Figure 1. 2-D Gabor wavelet representation of a facial image.

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 1980's [4]. More recently von der Malsburg's group has developed a face recognition system making use of this representation [8, 12]. The Gabor wavelet representation allows description of spatial frequency structure in the image while preserving information about spatial relations.

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

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

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.

For the 256×256 images used in the experiments below, 5 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.

2.2 Discriminant Analysis

The examples considered can be treated using two class discriminant analysis, *eg.* male or female. For facial expression, the presence or absence of each expression is tested and the outcomes used to classify the expression. 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 [5], whose notation we preserve here. Consider a set of n d-dimensional vectors $\mathbf{x_1}, ...\mathbf{x_n}$, 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:

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

$$S_B = (\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^\mathbf{t}$$

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 $\mathbf{m_1} - \mathbf{m_2}$.

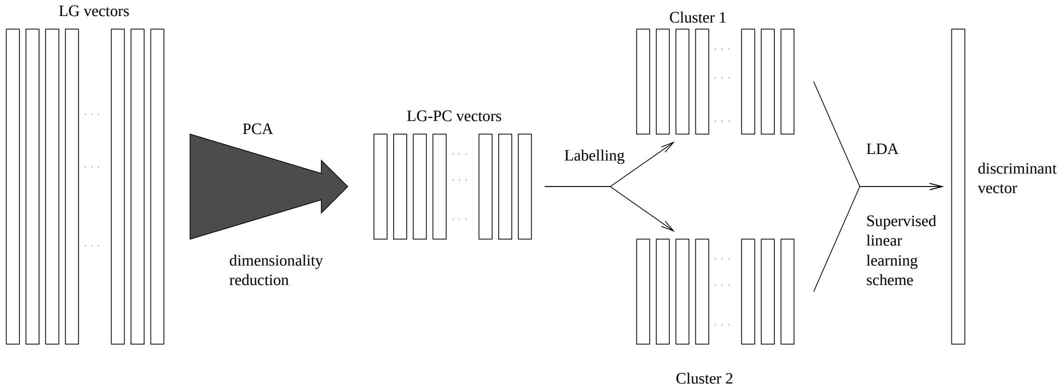

Figure 2. Dimensionality reduction and discriminant analysis of the LG vectors.

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,

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

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 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,$C = \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Phi_i}\mathbf{\Phi_i^t}$. Because of the high dimensionality of the LG vectors, C is very large, however there are only $n-1$ non-zero 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 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 non-singular. If W_{pca} is the matrix of eigenvectors having the N largest eigenvalues (W_{pca} is of dimension $d \times N$) then,

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

If every eigenvector with non-zero 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 over-fitting 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.

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

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

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

3 Experiments: Facial Expression Classifier

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. Fiducial grids, with nodes on featural landmarks of the face, were positioned by hand. A fully automatic implementation would be possible using labelled elastic graph matching [8, 12]. For an input image that is positively classified for two expressions, the normalized distance 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 databases here. However for mixed-expression data, an alternate approach may be more suitable.

The expression classifier was first tested using a set of 193 images of expressions posed by 9 Japanese females, which has been used in previous studies of facial expression recognition [9]. Each expresser posed 3 or 4 examples of each of the six fundamental facial expressions and a neutral face. Samples from the database may be seen in [9]. The entire set of images was divided into 10 segments; the discriminant vectors were calculated using 9 of these segments and the generalization performance tested on the remaining segment, with the results averaged over all 10 distinct partitions. The generalization rate for this system is 92%. To measure generalization over identity, the the image set was partitioned into 9 segments each corresponding to one expresser. The system was trained on 8 of the segments and then tested on the 9th. This was repeated for all 9 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%.

The system was also tested using the facial expression image set of Ekman and Friesen [6], consisting of 110 images, of which 51 are male and 59 are female. The system has a peak generalization rate of 82% tested on all expression categories.

4 Experiments: Live Demo

A classifier was trained to categorize face images according to sex: male/female, "race": east-Asian/other, expression: smile/other. The system was created for a demo held annually at ATR and all images in the dataset were acquired under live conditions, with some variation lighting, background and pose. 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 [8, 12]. The facial registration grid has 4 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 non-smiles). Sample images are shown in figure 3, which display a typical range of variation in the image conditions and grid position. The system was periodically re-trained 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. 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% for expression, 95% for "race", and 92% for sex recognition.

5 Saliency maps

The magnitude of each component of the discriminant vector determines its its influence on the classification decision, and is therefore a measure of the saliency of the corresponding feature. Figure 4 display the discriminant vector magnitude averaged over all frequencies and orientations at each position on the rectangular grid for sex, "race" and expression classification tasks for the images in the ATR Open House database. The size of each filled circle is proportional to the discriminant vector component magnitude. The figures show that different regions of the face are salient for different classification tasks. The top panel of figure 5 plots discriminant vector component magnitude averaged over frequency, position, 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 panel of figure 5 plots discriminant vector component magnitude averaged over orientation, node, and expression as a function of spatial frequency, and shows that intermediate frequencies are the most salient.

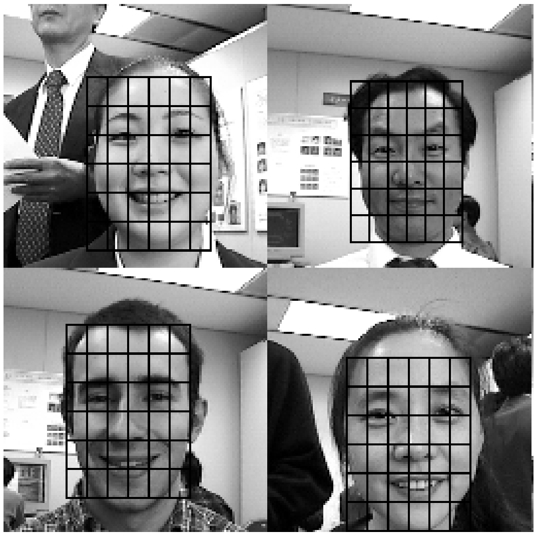

Figure 3. Sample images from the ATR Openhouse database.

6 Discussion

This paper presented a new algorithm for automatically extracting high-level information from facial images. The algorithm is a hybrid system using labelled elastic graph matching [8, 12] to register a grid with the face and extract Gabor wavelet features and a classifier similar to that used by the Fisherface method [1, 7].

Use of a representation based on the amplitude of 2-D Gabor wavelet transform relaxes the requirement for exact registration of the internal features of the face. Substantial

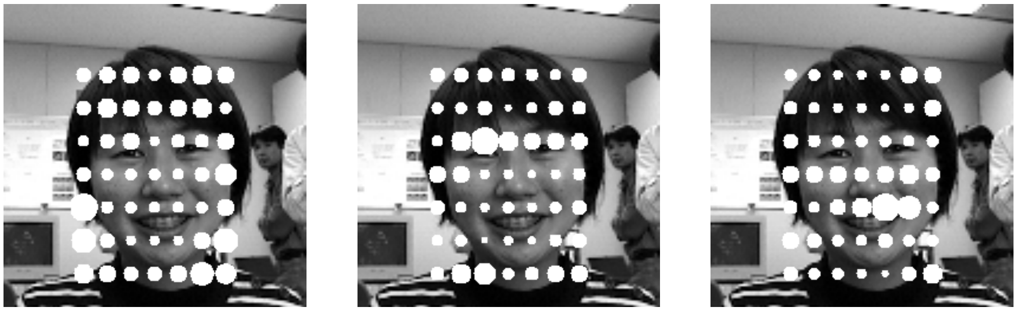

Figure 4. Node saliencies. Left: Male/Female. Middle: East Asian/Not East Asian. Right: Smile/No-smile.

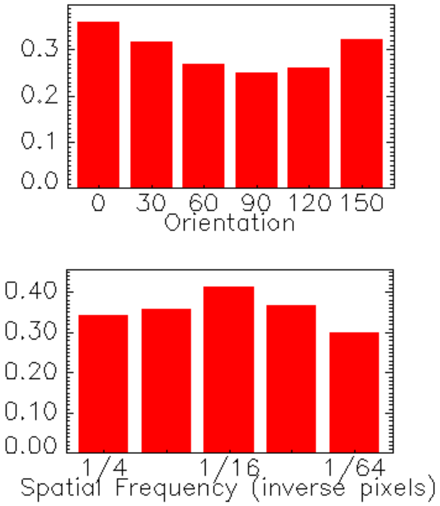

Figure 5. Orientation and spatial frequency saliencies for the Ekman and Friesen database.

shifts of the grid position (see fig. 3) were tolerated, and generalization exceeding 90% for sex, “race”, and expression maintained. These rates compare favorably with previous results using other single-image methods on binary classification tasks [12]. The LDA algorithm can be trained in about a minute: during the live demo the system was retrained nearly each time a new image was acquired. Hence, the system can be quickly adapted to changes in local conditions of use (lighting, camera or background) by retraining the algorithm in the new environment.

Results for recognizing expression from single images were not significantly different from those obtained using a multi-layered perception by Zhang *et al.* on the same (Japanese female) database, suggesting that the linear LDA algorithm is powerful enough for this classification task (92% for the LDA vs. 90% with the non-linear perceptron). Fewer hidden units were needed by the non-linear perceptron to attain a similar generalization rate than number of PC's retained with the linear classifier. The average generalization rate over expressor identity was 75% for the Japanese female image set and 82% for the Ekman pictures. This is still remarkably high considering the classifier has only about 10 individuals in the training set to learn which variations are due to identity and which are due to expression. The rate is not significantly different from the 86% 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 over identity for expression recognition are to be expected if the number of individuals in the training set is increased.

The saliency information displayed in figure 4 shows that different areas of the face are useful for different classification tasks. Furthermore, as seen in figure 5, saliency varies with spatial frequency and filter orientation. Filters of intermediate spatial frequency were found to be the slightly

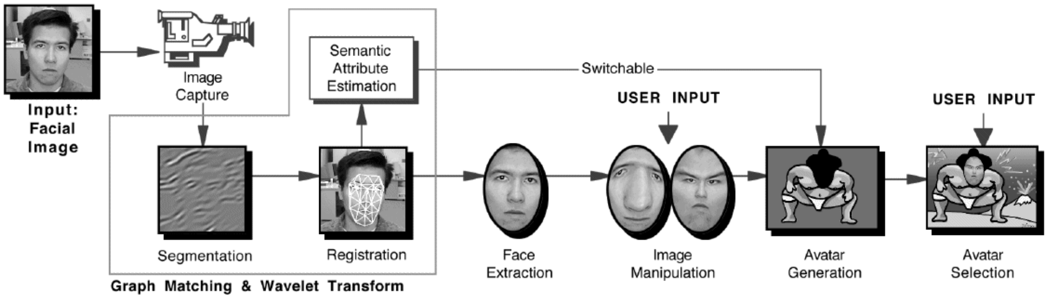

Figure 6. Automatic facial attribute estimation demo system.

more informative for expression classification. Filters having horizontal orientation were more heavily weighted in the discriminant vector than other orientations. This may reflect the fact that 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. Vertical displacement of roughly horizontal edges form the largest component of these motions.

A somewhat humorous but commercially viable application of this algorithm in the realm of virtual reality and/or entertainment technology has been developed. Figure 6 shows a schematic for a proposed automatic personalized avatar generation system. An avatar is a users abstract representation in a virtual reality environment. Using elastic graph matching to find and normalize a face in an input image, and the classification algorithm described in the present paper to characterize it, the face is cropped from the image and applied to a body selected automatically using the automatically determined facial attributes. Distortion of the facial image or morphing using graph coordinates may be used to comic effect. The live demo described in section 4 is in fact a prototype of such a system.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherface: Recognition Using Class Specific Linear Projection,” *IEEE PAMI*, vol .19, no. 7, pp. 711-720, 1997

[2] C. Padgett and G. Cottrell, “Identifying emotion in static face images,” *Proceedings of the 2nd Joint Symposium on Neural Computation*, vol. 5, pp. 91-101, La Jolla, CA, 1995.

[3] I. Craw, N. Costen, T. Kato, G. Robertson, and S. Akamatsu, “Automatic Face Recognition: Combining Configuration and Texture,” *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, Bichsel, M., (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,” *Journal of the Optical Society of America A*, vol. 2, pp. 1160-1169, 1985.

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

[6] P. Ekman, W. V. Friesen, “Pictures of Facial Affect,” Human Interaction Laboratory, University of California Medical Center, San Francisco, 1976

[7] K. Etemad and R. Chellappa, “Discriminant Analysis for Recognition of Human Face Images,” *Journal the of Optical Society of America A*, vol. 14, no. 8, pp. 1724-1733, 1997.

[8] 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 Transactions on Computers*, vol. 42, no. 3, pp. 300-311, 1993.

[9] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba “Coding Facial Expressions with Gabor Wavelets,” *Proceedings of the Third IEEE Conference on Face and Gesture Recognition*, Nara, Japan, April 1998, pp. 200-205.

[10] M. Lyons, J. Budynek, and S. Akamatsu “Automatic Classification of Single Facial Images,” *IEEE PAMI*, vol. 21, no. 12, 1999, in press.

[11] M. Lyons, K. Morikawa, and S. Akamatsu “A linked aggregate code for processing faces,” *Pragmatics and Cognition*, vol. 8, no. 1, pp. 63-81, 2000.

[12] L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg, “Face Recognition and Gender Determination,” *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, Bichsel, M., (Ed), pp. 92-97, 1995.