

Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron

Zhengyou Zhang^{1,3} Michael Lyons¹ Michael Schuster² Shigeru Akamatsu¹

¹ ATR Human Information Processing Research Laboratories

² ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

³ INRIA, 2004 route des Lucioles, BP 93, F-06902 Sophia-Antipolis Cedex, France
e-mail: zzhang@sophia.inria.fr, zzhang@hip.atr.co.jp

Abstract

In this paper, we investigate the use of two types of features extracted from face images for recognizing facial expressions. The first type is the geometric positions of a set of fiducial points on a face. The second type is a set of multi-scale and multi-orientation Gabor wavelet coefficients extracted from the face image at the fiducial points. They can be used either independently or jointly. The architecture we developed is based on a two-layer perceptron. The recognition performance with different types of features has been compared, which shows that Gabor wavelet coefficients are much more powerful than geometric positions. Furthermore, since the first layer of the perceptron actually performs a nonlinear reduction of the dimensionality of the feature space, we have also studied the desired number of hidden units, i.e., the appropriate dimension to represent a facial expression in order to achieve a good recognition rate. It turns out that five to seven hidden units are probably enough to represent the space of feature expressions.

1. Introduction

There are a number of difficulties in facial expression recognition (FER) due to the variation of facial expression across the human population and to the context-dependent variation even for the same individual. Even we human beings may make mistakes [7]. On the other hand, FER by computer is very useful in many applications such as human behavior interpretation and human-computer interface.

An automatic FER system needs to solve the following problems: detection and location of faces in a cluttered scene, facial feature extraction, and facial expression classification.

Face detection has been studied by many researchers, and it seems that most successful systems are based neural networks [19, 17]. Once a face is detected in the image, the corresponding region is extracted, and is usually normalized to have the same size (for example, the same distance between two eyes) and the same gray level. In this paper, we do not address the face detection problem.

Facial feature extraction attempts to find the most appropriate representation of the face images for recognition. There are mainly two approaches: holistic template-matching systems and geometric feature-based systems [4]. In holistic systems, a template can be a pixel image or a feature vector obtained after processing the face image as a whole. In the latter, principal component analysis and multilayer neural networks are extensively used to obtain a low-dimensional representation. In geometric feature-based systems, major face components and/or feature points are detected in the images. The distances between feature points and the relative sizes of the major face components are computed to form a feature vector. The feature points can also form a geometric graph representation of the faces. Feature-based techniques are usually computationally more expensive than template-based techniques, but are more robust to variation in scale, size, head orientation, and location of the face in an image. The work to be described in this paper is, to some extent, an hybrid approach. We first locate a set of feature points, and then extract a set of Gabor wavelet coefficients at each point through image convolution.

Compared with face recognition, there is relatively a small amount of work on facial expression recognition. The first category of previous work uses image sequences. Suwa et al. [18] did a preliminary analysis of facial expressions by tracking the motion of twenty identified spots. Mase [13] uses the means and variances of optical flow data at evenly divided small blocks. Yacoob and Davis [22] use the inter-frame motion of edges extracted in the area of the mouth,

nose, eyes, and eyebrows. Bartlett et al. [2] use the combination of optical flow and principal components obtained from image differences. Essa and Pentland [8] builds a dynamic parametric model by tracking facial motion over time, which can then be used for analyzing facial expressions. The second category of previous work tries to classify facial expressions from static images. Turk and Pentland [20] represent face images by eigenfaces through linear principal component analysis. Padgett and Cottrell [14] use an approach similar to eigenfaces but with seven pixel blocks from feature regions (both eyes and mouth). Cottrell and Metcalfe [5] use holistic representations based on principal components, extracted by feed forward networks. Rahardja et al. [15] also use holistic representations with neural networks, but the images are represented in a pyramid structure. Lanitis et al. [10] use parameterized deformable templates (flexible models) which take into account both variations in shape and grey-level appearance.

In this paper, we extract two types of features from face images in order to recognize facial expressions. The first type is the geometric positions of a set of fiducial points on a face. The second type is a set of multi-scale and multi-orientation Gabor wavelet coefficients extracted from the face image at the fiducial points. They can be used either independently or jointly. The architecture we developed is based on a two-layer perceptron. The recognition performance with different types of features will be compared. Since the first layer of the perceptron actually performs a nonlinear reduction of the dimensionality of the feature space, we will also study the desired number of hidden units, i.e., the appropriate dimension to represent a facial expression in order to achieve a good recognition rate. Finally, we note that a similar representation of faces has been developed in Wiskott et al. [21] for face recognition, where they use a labeled graphs, based on a Gabor wavelet transform, to represent faces, and face recognition is done through elastic graph matching.

2. Data Set and Representation

The database we use for our experiments contains 213 images of female facial expressions. The head is almost in frontal pose. Original images have been rescaled and cropped such that the eyes are roughly at the same position with a distance of 60 pixels in the final images (resolution: 256 pixels \times 256 pixels). The number of images corresponding to each of the 7 categories of expression (neutral, happiness, sadness, surprise, anger, disgust and fear) is roughly the same. A few of them are shown in Fig. 1. For details on the collection of these images, the reader is referred to [12].

Each image is represented in two ways. The first uses 34 fiducial points as shown in Fig. 2. They have been se-

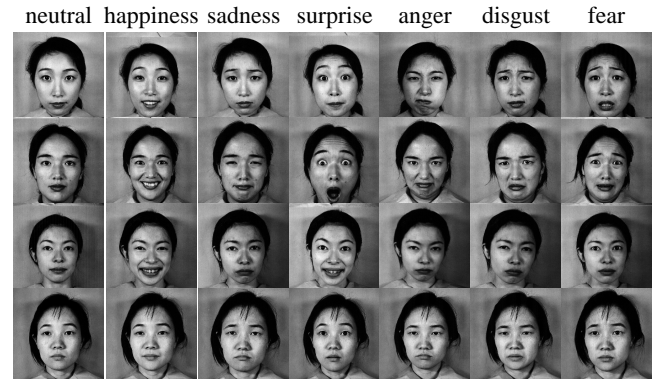


Figure 1. Facial expression database: Examples

lected manually. Development of a technique for automatically extracting these points is under way. (An automatic technique for building a similar representation has already been reported in the literature [9, 21].) The image coordinates of these points (geometric positions) will be used as features in our study. Therefore, each image is represented by a vector of 68 elements.

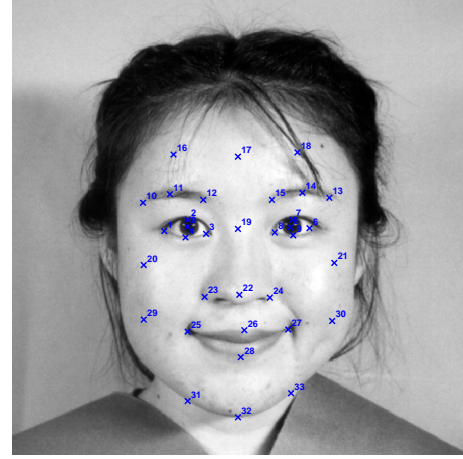


Figure 2. Geometric representation: 34 fiducial points to represent the facial geometry

The second way is use features extracted by using 2-D Gabor transforms [6, 11]. A 2-D Gabor function is a plane wave with wavevector \mathbf{k} , restricted by a Gaussian envelope function with relative width σ :

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

We set $\sigma = \pi$ for our 256 \times 256 images. We use a discrete set of Gabor kernels which comprise 3 spatial frequencies, i.e., scales, (with wavenumber $k = \|\mathbf{k}\| = (\pi/4, \pi/8, \pi/16)$ in inverse pixels) and 6 distinct orientations from 0° to 180°, differing in 30° steps. Two examples with three of the total 18 even Gabor kernels are shown in Fig. 3. Each image is

convolved with both even and odd Gabor kernels at the location of the fiducial points as shown in Fig. 2. We have therefore 18 complex Gabor wavelet coefficients at each fiducial point. In our study, only the magnitudes are used, because they vary slowly with the position while the phases are very sensitive. In summary, with Gabor wavelet coefficients, each image is represented by a vector of 612 (18×34) elements.

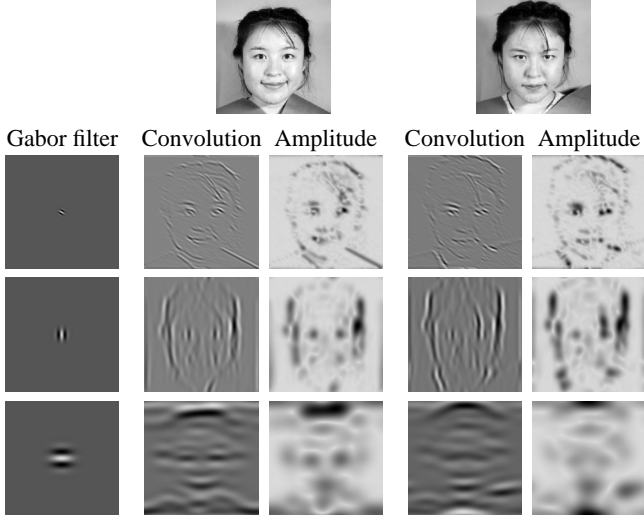


Figure 3. Gabor wavelet representation: Examples of three kernels

3. The Architecture and Training

The architecture of our FER system is based on a two-layer perceptron (see Fig. 4). As described in Sect. 2, an image is first preprocessed, and two sets of features (geometric positions and Gabor wavelet coefficients) are extracted. These features are fed in the input units of the two-layer perceptron. The objective of the first layer is to perform a non-linear reduction of the dimensionality of feature space, depending on the number of hidden units. Note that there are no interconnections in the first layer between geometric and Gabor-wavelet parts, because they are two pieces of information very different in nature. The second layer makes a statistical decision based on the reduced set of features in the hidden units. An output unit is associated with a particular facial expression, so our system contains 7 output units. Each output unit gives an estimate of the probability of the input image belonging to the associated facial expression.

The FER problem is considered as a statistical classification problem. The training is done by minimizing the cross-entropy for multiple classes [3]:

$$E = - \sum_n \sum_{k=1}^c t_k^n \ln \left(\frac{y_k^n}{t_k^n} \right)$$

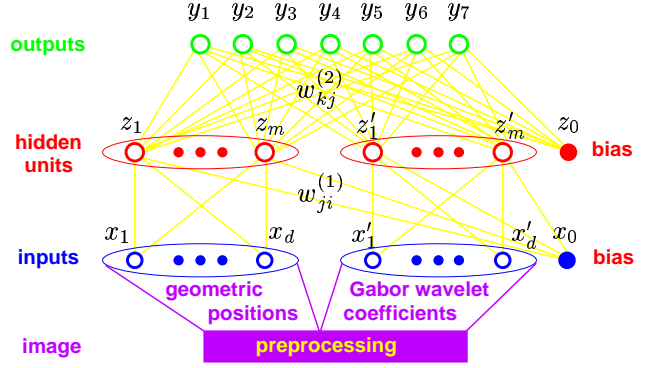


Figure 4. The architecture: There are no interconnections in the first layer between geometric and Gabor-wavelet parts

where t_k^n and y_k^n are respectively the pattern target value and network output value, representing the probability that input \mathbf{x}^n belongs to class \mathcal{C}_k . The activation function of the output units is the *softmax* function:

$$y_k = \frac{\exp(a_k)}{\sum_{k'=1}^c \exp(a_{k'})}$$

where $a_k = \sum_j w_{kj} z_j$ and z_j is the output of hidden unit j . The activation function of the hidden units is 'tanh':

$$g(a_j) = \tanh(a_j) \equiv \frac{e^{a_j} - e^{-a_j}}{e^{a_j} + e^{-a_j}}$$

where $a_j = \sum_i w_{ji} x_i$ and x_i is the value of input unit i .

The two-layer perceptron is trained through Rprop (Resilient propagation) [16], which is a local adaptive learning scheme, performing supervised batch learning. The idea is to eliminate the harmful influence of the size of the partial derivative on the weight step. In consequence, the weight update depends only the sign of the derivative, and is exclusively determined by a weight-specific, so-called "update-value" $\Delta_{ij}^{(t)}$:

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t)} > 0 \\ +\Delta_{ij}^{(t)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t)} < 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\frac{\partial E}{\partial w_{ij}}^{(t)}$ denotes the summed gradient information over all patterns of the pattern set. The update-value $\Delta_{ij}^{(t)}$ itself is adapted based on a sign-dependent learning process:

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^+ * \Delta_{ij}^{(t-1)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t-1)} * \frac{\partial E}{\partial w_{ij}}^{(t)} > 0 \\ \eta^- * \Delta_{ij}^{(t-1)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t-1)} * \frac{\partial E}{\partial w_{ij}}^{(t)} < 0 \\ \Delta_{ij}^{(t-1)} & \text{otherwise} \end{cases}$$

where $0 < \eta^- < 1 < \eta^+$. (we use $\eta^- = 0.5$, $\eta^+ = 1.2$).

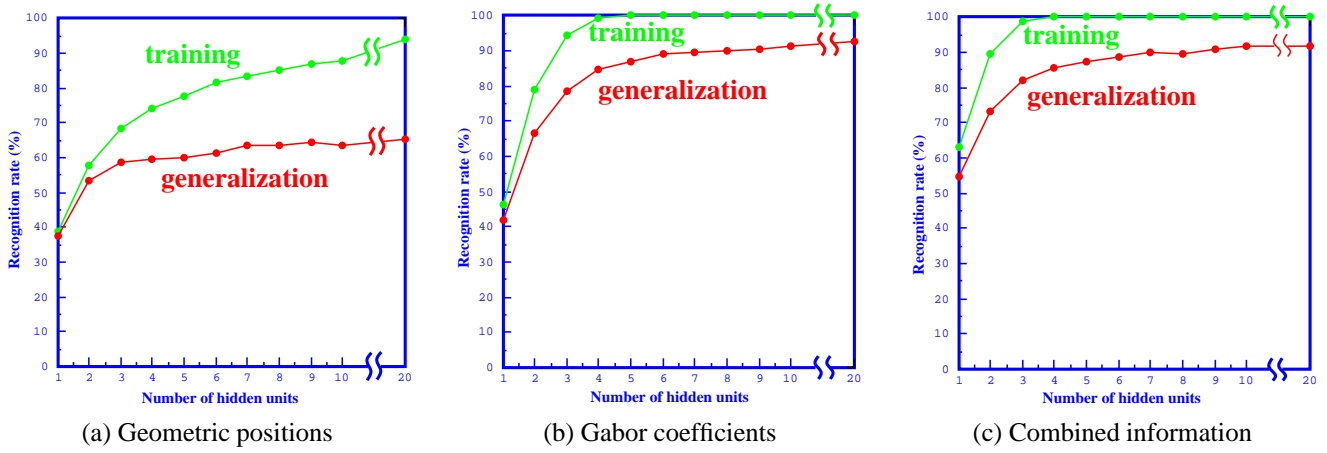


Figure 5. Recognition rate with respect to the number of hidden units

4. Experiments

4.1. Computer Recognition Results

Our goal is to develop a recognition system which not only works well on the training data but also gives good predictions for new data. Since the size of our database is limited (213 images), we use the cross-validation technique [3] to test different configurations of our FER architecture. More precisely,

- We partition the data set at random into S distinct segments (we set $S = 10$).
- We then train a two-layer perceptron using data from $S - 1$ of the segments and test its performance, by evaluating the error function (recognition rate), using the remaining segment.
- The above process is repeated for each of the S possible choices of the segment to be omitted from the training process.
- Finally, we average the results over all S trained two-layer perceptrons.

Since the training is a nonlinear optimization problem, the final result depends on the initial guess of the weights of the perceptrons. So, each perceptron is furthermore trained ten times with randomly initialized weights. Thus, the result for each configuration shown below is the average of the results produced by 100 trained two-layer perceptrons.

We have carried out experiments on the FER using the developed architecture by using geometric positions alone, using Gabor wavelet coefficients alone, and by using the combination of the two pieces of information. In order to investigate the appropriate dimension to code the facial expression, we vary the number of hidden units from 1 to 20. The perceptrons with geometric positions alone were trained by running 250 cycles through all the training data, while other perceptrons were trained by running only 100 cycles.

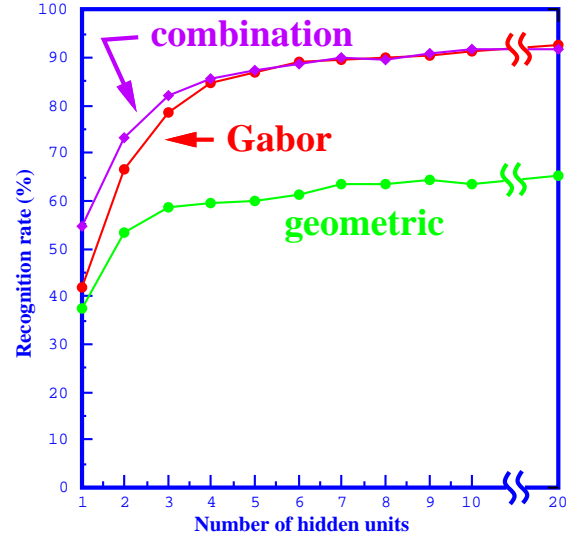


Figure 6. Comparison of the generalized recognition rates

The recognition rates on the training data and on the test data (generalization) with respect to the number of hidden units are shown in Fig. 5. In particular, the generalized recognition rates are compared in Fig. 6.

From the experimental results, we have the following observations:

- Gabor coefficients are much more powerful than geometric positions;
- At least two hidden units are necessary to code reasonably facial expressions;
- Probably from 5 to 7 hidden units are sufficient to code precisely facial expressions;
- Adding geometric positions improves the recognition rate only for low dimensional coding (with less than 5 hidden units). No improvement is observed when 5 or more hidden units are used.



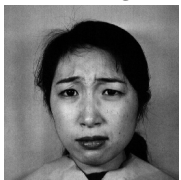

Label: Surprise	NN outputs	Label: Happiness	NN outputs
	Neu. 0.000		Neu. 0.122
	Hap. 0.000		Hap. 0.720
	Sad. 0.000		Sad. 0.000
	Sur. 1.000		Sur. 0.000
	Ang. 0.000		Ang. 0.000
	Dis. 0.000		Dis. 0.000
	Fear 0.000		Fear 0.158
Label: Disgust	NN outputs	Label: Fear	NN outputs
	Neu. 0.001		Neu. 0.002
	Hap. 0.000		Hap. 0.000
	Sad. 0.428		Sad. 0.022
	Sur. 0.000		Sur. 0.000
	Ang. 0.016		Ang. 0.001
	Dis. 0.555		Dis. 0.005
	Fear 0.000		Fear 0.970

Figure 7. Examples of correct labeling




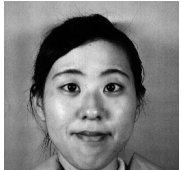
Label: Anger	NN outputs	Label: Fear	NN outputs
	Neu. 0.000		Neu. 0.010
	Hap. 0.000		Hap. 0.000
	Sad. 0.091		Sad. 0.502
	Sur. 0.000		Sur. 0.000
	Ang. 0.224		Ang. 0.001
	Dis. 0.685		Dis. 0.007
	Fear 0.000		Fear 0.479
Label: Sadness	NN outputs	Label: Happiness	NN outputs
	Neu. 0.099		Neu. 0.822
	Hap. 0.846		Hap. 0.017
	Sad. 0.052		Sad. 0.150
	Sur. 0.000		Sur. 0.011
	Ang. 0.000		Ang. 0.000
	Dis. 0.002		Dis. 0.000
	Fear 0.002		Fear 0.000

Figure 8. Examples of disagreement

The recognition rate (i.e., the agreement with the labeling provided by the expressors) achieved by our system is 90.1% with 7 hidden units. This should be compared with the agreement between human subjects and expressors' labeling. In the study of Lyons et al. [12], 60 human non-expert subjects were asked to rate each facial image for content of the six basic facial expressions. In 20.5% highest rating (averaged over all subjects) disagreed with the expression label of the image. This is similar to the results reported in the literature but with different image database [2, 10]. Several sources of this disagreement may be identified. The expressor may have posed the expression inaccurately or even incorrectly in some cases. The experimental subjects may have confused one expression with another when performing the rating task (for example, fear may be confused with surprise and anger with disgust). Finally, in a small percentage of cases there is also a possibility that the images were mislabelled by the experimenter.

In order to give the reader a concrete feeling of the FER results, we show a few examples in Fig. 7 and Fig. 8. The original labeling in the database and our system outputs are both shown. Note that, our system provides the probabil-

ity it believes that an image belongs to each of the facial expression classes. The examples shown in Fig. 7 have obtained a consistent labeling from our system, while for those in Fig. 8, our system does not agree with the labeling given in the database. Note that even in the latter case, our system usually gives a reasonable result, because the expressor may have posed an incorrect expression.

4.2. Experiments After Excluding Fear Images

The expressors found it most difficult to pose fear expressions accurately. In addition, human has more difficulty in recognizing fear. There is some evidence supporting the hypothesis that fear expressions are processed differently from the other basic facial expressions [1]. If we exclude the fear images from the database, an experiment with 30 human non-experts shows that in 85.6% of all cases, human subjects agree with the expressors' labeling, about 6% higher than when fear images are included. Hence, we have repeated exactly the same analysis as in the last subsection but with a dataset in which all fear images were excluded. The results are shown in Fig. 9. The same general observations can be made. When 7 hidden units are used, our system achieves a generalized recognition rate of 73.3% with geometric positions alone, 92.2% with Gabor wavelet coefficients alone, and 92.3% with combined information.

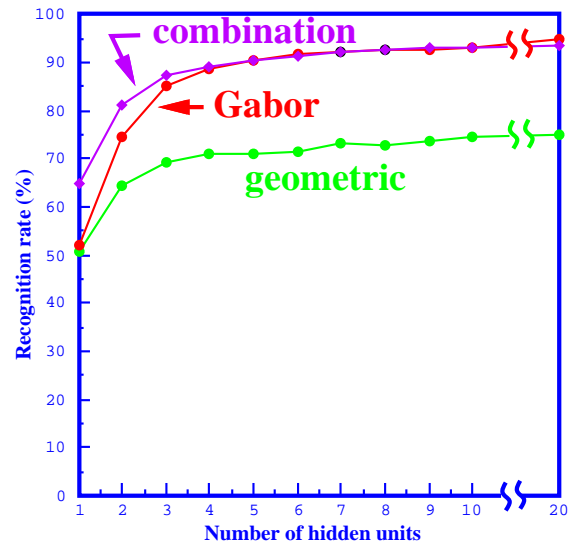


Figure 9. Comparison of the generalized recognition rates when the fear images have been excluded from the database

5. Conclusion

In this paper, we have compared the use of two types of features extracted from face images for recognizing facial expressions. The first type is the geometric positions of a set of fiducial points on a face. The second type is a set of multi-scale and multi-orientation Gabor wavelet coefficients extracted from the face image at the fiducial points. They can be used either independently or jointly. We have developed an architecture based on a two-layer perceptron. Comparison of the recognition performance with different types of features shows that Gabor wavelet coefficients are much more powerful than geometric positions and that the agreement between computer and the expressors' labeling is higher than that between human subjects and the expressors' labeling.

Furthermore, since the first layer of the perceptron actually performs a nonlinear reduction of the dimensionality of the feature space, we have also studied the desired number of hidden units, i.e., the appropriate dimension to represent a facial expression in order to achieve a good recognition rate. It turns out that at least two hidden units are necessary to code reasonably facial expressions and that five to seven hidden units are probably enough to give a precise representation.

Acknowledgment

We thank Miyuki Kamachi and Jiro Gyoba (Kyushu University, Japan) for providing the facial expression database.

References

- [1] R. Adolphs, D. Tranel, H. Damasio, and A. Damasio. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, 372:669–672, 1994.
- [2] M. Bartlett, P. Viola, T. Sejnowski, L. Larsen, J. Hager, and P. Ekman. Classifying facial action. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*. MIT Press, Cambridge, MA, 1996.
- [3] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.
- [4] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, May 1995.
- [5] G. Cottrell and J. Metcalfe. Face, gender and emotion recognition using holons. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 3*, pages 564–571. Morgan and Kaufman, San Mateo, 1991.
- [6] J. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 36(7):1169–1179, July 1988.
- [7] P. Ekman and W. Friesen. *Unmasking the Face: A guide to recognizing emotions from facial expressions*. Consulting Psychologists Press, Palo Alto, CA, 1975.
- [8] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.
- [9] M. Lades, J. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.
- [10] A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, July 1997.
- [11] T. Lee. Image representation using 2d Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, Oct. 1996.
- [12] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. 1997. submitted for publication.
- [13] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions E*, 74(10):3473–3483, 1991.
- [14] C. Padgett and G. Cottrell. Identifying emotion in static images. In *Proceedings of the 2nd Joint Symposium on Neural Computation*, volume 5, pages 91–101, La Jolla, CA, 1997.
- [15] A. Rahardja, A. Sowmya, and W. Wilson. A neural network approach to component versus holistic recognition of facial expressions in images. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, volume 1607 of *SPIE Proc.*, pages 62–70, 1991.
- [16] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In H. Ruspini, editor, *Proceedings of the International Conference on Neural Networks*, pages 586 – 591, San Francisco, CA, Mar. 1993.
- [17] H. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical Report CMU-CS-95-158R, School of Computer Science, Carnegie Mellon University, Nov. 1995.
- [18] M. Suma, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *Proceedings of the 4th International Joint Conference on Pattern Recognition*, pages 408–410, 1978.
- [19] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. Technical Report A.I. Memo 1521, CBCL Paper 112, MIT, Dec. 1994.
- [20] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–86, Mar. 1991.
- [21] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.
- [22] Y. Yacoob and L. Davis. Recognizing facial expressions by spatio-temporal analysis. In *Proceedings of the International Conference on Pattern Recognition*, volume 1, pages 747–749, Jerusalem, Israel, Oct. 1994. Computer Society Press.