Proceedings of the Intern. Workshop on Automatic Face- and Gesture-Recognition,
pages 92-97 (1995), Zürich

Face Recognition and Gender Determination*

Laurenz Wiskott †, Jean-Marc Fellous ‡, Norbert Krüger †, Christoph von der Malsburg †‡

†Ruhr Universität Bochum
Institut für Neuroinformatik
D-44780 Bochum, Germany
WWW: http://www.neuroinformatik.ruhr-uni-bochum.de
E-mail: laurenz@neuroinformatik.ruhr-uni-bochum.de

‡University of Southern California
Dept. of Computer Science and Section for Neurobiology
Los Angeles, CA 90089, USA

Abstract

The system presented here is a specialized version of a general object recognition system. Images of faces are represented as graphs, labeled with topographical information and local templates. Different poses are represented by different graphs. New graphs of faces are generated by an elastic graph matching procedure comparing the new face with a set of precomputed graphs: the "general face knowledge". The final phase of the matching process can be used to generate composite images of faces and to determine certain features represented in the general face knowledge, such as gender or the presence of glasses or a beard. The graphs can be compared by a similarity function which makes the system efficient in recognizing faces.

1 Introduction

Face recognition systems can be subdivided into two main categories [1] depending on the nature of the coding of an input picture and its processing. Schemes that use pixels (grey-level values) as the basis for their coding and various forms of statistical analysis for the processing are often referred to as template-based approaches [2, 3]. Procedures that utilize various properties of a face (such as facial topology, hair color, ...) as their code for the face are, on the other hand, referred to as feature-based [4]. In general, each of these two classes of systems bear advantages and drawbacks regarding database size, the uncontrolled nature of the input stimuli (head orientation, illumination differences or partial occlusions for example), variable picture quality (signal/noise), to cite only a few. It is likely that a robust and efficient system achieving face recognition will require a hybrid approach.

The work presented below uses labeled graphs of two-dimensional views as a hybrid representation of faces. The nodes are labeled with jets, a special class of local templates built on the basis of wavelet transforms. The edges are labeled with distance vectors similar to the geometric features in [1]. More abstract features like gender are determined from the local templates.

*Supported by grants from the German Federal Ministry for Science and Technology (413-5839-01 IN 101 B9) and from the US Army Research Lab (01/93/K-0109).

A small set of manually controlled model graphs serve as a "general face knowledge". It represents the face space and is used to generate graphs of new faces by elastic graph matching. By this means large sets of model graphs (called galleries) can be generated automatically. The gallery is distinct from the general face knowledge since the former represents a set of individual persons to be recognized while the latter represents the face space in general and might contain much fewer samples than the gallery.

Recognizing a new face is done in three stages. In a preprocessing stage the location and size of the face is estimated and the image is rescaled accordingly. In a second stage the general face knowledge is matched to the image by maximizing a similarity function. By this means facial landmarks (termed hereafter fiducial points) are located and an image graph is generated. In the last stage the generated graph is compared to all individual model graphs of the gallery. The most similar model is taken to be the correct one.

Matching of the general face knowledge also provides information on the basis of which a composite or phantom face can be generated and high level features such as gender can be determined. Once the general face knowledge is generated under manual control, no further user intervention is needed for storing and recognizing new individuals. New image graphs are generated by the matching process and compared with a simple similarity function.

The system is an extended algorithmic version of a fully neural system for robust object recognition [5, 6, 7]. The advantage of the approach is its simplicity and flexibility. Only few modifications are required to apply the system to different tasks, such as object recognition in cluttered scenes with significant mutual occlusion [8], face recognition [3], in cases where input images are scaled and rotated in the image plane [9], or the determination of face features such as gender. The main goal of this work is not to build a specialized, highly optimized system for face recognition, but to contribute to the larger effort of creating a robust and flexible general purpose system that can be used to solve different visual tasks.

2 Face Representation

We use graphs G with an underlying two-dimensional topography. The nodes are labeled with jets J_n and the edges are labeled with distance vectors $\Delta\vec{x}_e$. In the simplest case the graph has the form of a rectangular grid with constant spacing between nodes.

The jets are based on a wavelet transform, which is defined as a convolution with a family of complex Gabor kernels

$$\psi_j(\vec{x}) = \frac{k_j^2}{\sigma^2} \exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right) \left[\exp(i\vec{k}_j\vec{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right],$$

providing at each location $\vec{x}$ the coefficients

$$J_j(\vec{x}) = \int I(\vec{x}')\psi_j(\vec{x}-\vec{x}')d^2\vec{x}'$$

given the image grey level distribution $I(\vec{x})$.

This preprocessing was chosen for its theoretical properties and its biological relevance, since the receptive fields of simple cells in the primary visual cortex are of similar shape as the Gabor kernels [10, 11]. They are localized in both space and frequency domains and have the shape of plane waves of a wave vector $\vec{k}_j$ restricted by a Gaussian envelope function of width σ/k with $\sigma = 2\pi$. In addition the kernels are corrected for their DC value, i.e., the integral $\int \psi_j(\vec{x})d^2\vec{x}$ vanishes. All kernels are similar in the sense that they can be generated from one kernel simply by dilation and rotation. We use kernels of five different sizes, index $\nu \in \{0,\ldots,4\}$, and eight orientations, index $\mu \in \{0,\ldots,7\}$. Each kernel responds best at the frequency given by the characteristic wave vector

$$\vec{k}_j = \begin{pmatrix} k_\nu \cos\phi_\mu \\ k_\nu \sin\phi_\mu \end{pmatrix}, \quad k_\nu = 2^{-\frac{\nu+2}{2}}\pi,\ \phi_\mu = \mu\frac{\pi}{8},$$

with index $j = \mu + 8\nu$.

The full wavelet transform provides 40 complex coefficients at each pixel (5 frequencies, 8 orientations). We will refer to this array of coefficients at one pixel as the jet $J(\vec{x})$, see figure 1.

The complex jet coefficients J_j can be written as $J_j(\vec{x}) = a_j(\vec{x})\exp(i\phi_j(\vec{x}))$ with a smoothly changing magnitude $a_j(\vec{x})$ and a phase $\phi_j(\vec{x})$ spatially varying with approximately the characteristic frequency of the respective Gabor kernel. Due to this variation one cannot compare the jets directly, because small spatial displacements change the individual coefficients drastically. One can therefore use either the magnitudes only or one has to compensate explicitly for the phase shifts due to a possible displacement. The two corresponding similarity functions are

$$S_a(J,J') = \frac{\sum_j a_j a'_j}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}}$$

and

$$S_\phi(J,J') = \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \vec{d}\vec{k}_j)}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}}.$$

where $\vec{k}_j$ is the characteristic wave vector of the respective Gabor kernel and $\vec{d}$ is an estimated displacement vector which compensates for the rapid phase shifts. $\vec{d}$ is determined by maximizing S_ϕ in its Taylor expansion around $\vec{d} = 0$ [12], which is a constraint fit of the two-dimensional $\vec{d}$ to the 40 phase differences $\phi_j - \phi'_j$.

The jets and the similarity functions are robust against changes in lighting conditions in two respects. Firstly, since the kernels are DC free, the jets are invariant with respect to general offsets in the image grey values. Secondly, since the similarity functions S are normalized, they are invariant with respect to contrast variations.

3 Elastic Graph Matching

In order to generate a new image graph G^I of a face, a procedure is applied which matches a stack of existing model graphs (the general face knowledge) with the image. In this section we consider only one model graph G^M. The extension to a stack of model graphs is described in the next section.

Let us assume that the model graph has N nodes labeled with jets J_n and E edges labeled with distance vectors $\Delta\vec{x}_e$. A generated image graph must have the same structure, i.e., the same number of nodes and the same pairs of nodes connected by an edge. The nodes should also be located at corresponding "fiducial points" in the faces, e.g., the left eye or the tip of the nose. (In case of the rectangular graphs, only few nodes are located on specific points, the two eyes and the line between the two lips. The others are located according to the regular structure of the grids, but they also refer to roughly the same location for all faces.) The labels J and $\Delta\vec{x}$ may differ, depending on the individual face. The latter are the distance vectors between the pixel coordinates from which the image jets were taken.

The similarity between the image graph and a model graph will depend on the jet similarities and the geometrical distortion between image and model graph. As the overall graph similarity we define

$$S_m(G^M,G^I) = \frac{1}{N}\sum_n S_\phi(J_n^M, J_n^I) - \frac{\lambda}{E}\sum_e (\Delta\vec{x}_e^M - \Delta\vec{x}_e^I)^2$$

where λ is a parameter controlling the relative importance of template and topographical similarities.

In the matching process a sequence of modifications to the image graph is chosen under the constraint that a change is accepted only if the graph similarity increases relative to the previous one. Doing this in a hierarchical, coarse-to-fine manner leads to a good approximation of the optimal image graph in a reasonable amount of time, see figure 2.

In order to find the correct pixel positions as precisely as possible we use the jet similarity function with phase and we only allow deformations of the graph towards pixels in the image that give an estimated displacement $\vec{d}$ in S_ϕ of less than one pixel, i.e., $|d| < 1$.

4 General Face Knowledge

Jets extracted from different faces can vary significantly. Hence one cannot expect to reliably find the fiducial points by matching one model to the image

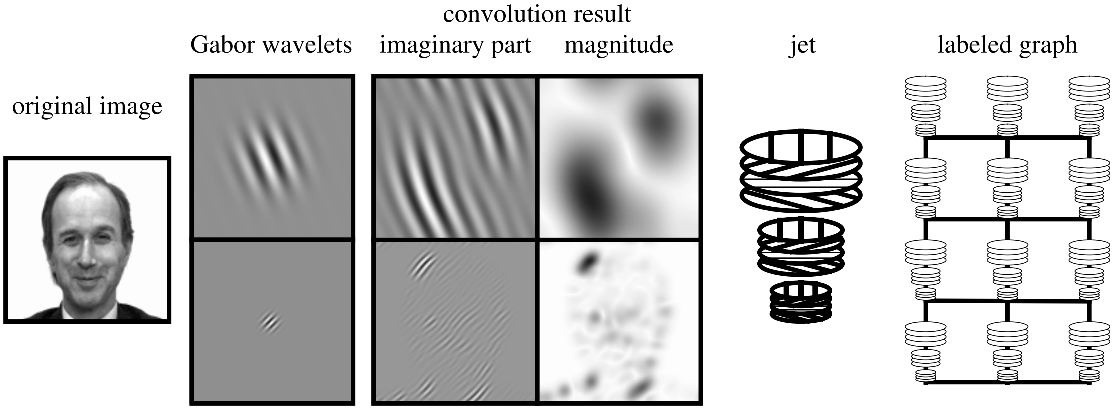

Figure 1: The graph representation of a face is based on a wavelet transform, a convolution with Gabor kernels of different size and orientation. The phase varies according to the main frequency of the kernels (see imaginary part) while the magnitude varies smoothly. The set of coefficients of the transform at one picture location is referred to as a jet and is computed on the basis of a small patch of grey values. A sparse set of such jets together with some topographic information constitutes an image graph representing an object, such as a face.

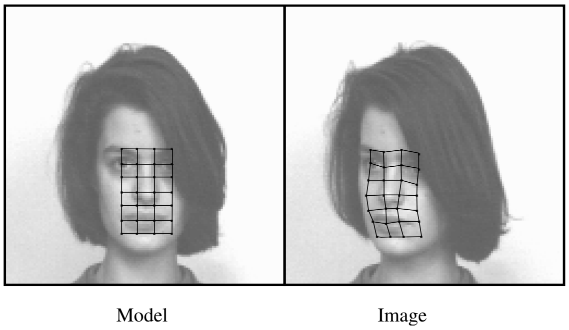

Figure 2: A model graph matched with a new image of the same person. The matching process attempts to find the image graph that is most similar to the model graph, i.e., the one with the most similar local templates and the minimal graph distortions.

of a different person. We solve this problem by using a set of different model graphs, the "general face knowledge", which covers the face space.

In the general face knowledge all model graphs have the same structure, with nodes referring to the same fiducial points. All the nodes referring to the same fiducial point are bound together and represent various instances of this local face region. The edge labels are averaged over the whole general face knowledge, thus leading to an average geometry.

The cost function defined above changes as each node of the image graph can be compared with the corresponding node of any of the models in the stack. When matching, we check all of them and use the one fitting best, see figure 3. We assume that for each new face and for each fiducial point we have an 'expert' jet in the general face knowledge, sufficiently similar to the jet of the new face at that location, to determine the precise position of the fiducial point. Beside yielding an image graph, the matching process also provides information about which model is most similar to the new face at any fiducial point. Such information is important for generating phantom faces and for determining face features.

5 Phantom Faces and Determining Face Features

What can we say about the new face if we discard all of its template information, i.e., the original image jets, just keeping the geometry of the matched image graph and the identity of the expert jet for each node?

First we are going to reconstruct the face image on the basis of the match results; we build a composite or phantom face resembling the original. For each node of the graph we copy the local grey level distribution of the respective expert model and apply a smooth transient to the patches of the neighboring nodes. This very simple method gives a good reconstruction of the original, see figure 4. Such a phantom face is typically composed of patches from about ten to twenty different models.

The very simple and general idea to determine face features now is the following: If the expert jets are taken mostly from female models, one can expect that the phantom face will look female and consequently that the original face was probably a female as well. This also holds for other features, such as facial hair or glasses. If the expert models for the lower half of the image graph are mostly bearded, then the original face was probably bearded as well, and similarly for glasses. One only has to label all models in the model stack with their respective features, decide which region of the face is relevant for a certain feature, and then compare which feature was most often provided by the expert models in that region.

In our test runs we used a gallery of 112 neutral frontal views, 65% of which were male, 19% were bearded, and 28% had glasses. Each of the 112 faces was analyzed while the remaining 111 models served

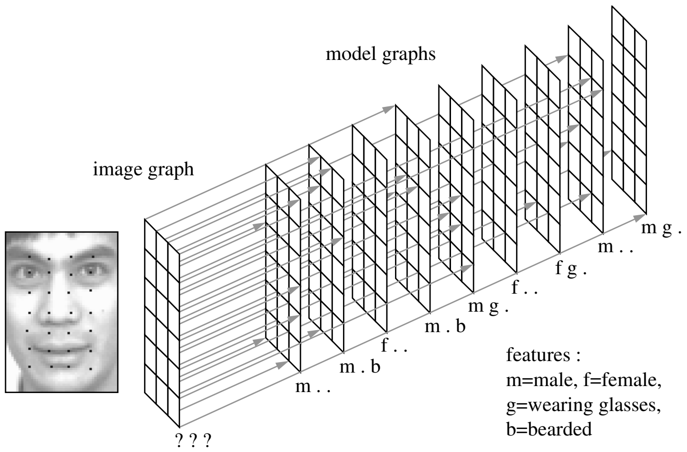

Figure 3: The stack structure of the general face knowledge. Here, we show how the individual nodes of an image graph will fit best to different model graphs. Each model graph is labeled with known features, on the basis of which the features of the new face can be determined.

as the general face knowledge. The 112 model graphs of 7×4 rectangularly arrayed nodes were positioned by hand; the image graphs were generated automatically. The relevant regions were chosen by hand for all three features: All nodes were considered to be relevant for gender determination, while we used only the lower three rows for the beard feature and the upper four rows for the glasses feature. If the number of relevant nodes labeled with a certain feature is above chance level, the system decided on this feature for the image face. For example, if more than 65% nodes were labeled male, the face was determined to be male. Results of this procedure were 90.2% correct gender classification, 92.9% correct beard detection, and 90.2% correct glasses detection. In order to show that the performance on gender determination is not due to facial hair, we tested it on a reduced general face knowledge and test set of 91 beardless faces with 57% males. The result was 91.3% correct gender classification. The difference is not significant; in general performance increases with the size of the general face knowledge.

The evaluation of the node labels can be considered naive and requires the choice of the relevant nodes by hand, but it shows the principle. We have tested a Bayesian approach as well and got an improvement of 1–3%. The Bayes approach also determines the relative reliability of the nodes. For beard and glasses, the lower and the upper rows, respectively, were more reliable, as expected. For gender determination there was a slight emphasis on the lower rows, even if only beardless faces were considered. The classification performance relies on what is represented in the general face knowledge. One cannot expect that with a Caucasian general face knowledge the system performs very well on Asian people, for example. We assume, however, that with an appropriate general face knowledge, other features like age, ethnic group, or facial expression could be detected.

The performance of the system is comparable to others. Brunelli and Poggio [13] trained a hyper basis function network on automatically extracted geometrical features. They achieved a correct gender classification rate of 87.5%. Golomb et al. [14] used a template-based approach. They trained a back-propagation network on a compressed representation (40 units) of low resolution face images of 30×30 pixels and achieved a performance of 91.9%. They used limited hair information and aligned the faces under manual control.

6 Rotation in Depth, Object Adapted Graphs, and Face Recognition

The system as described so far relies on one two-dimensional view only. The elastic graph matching provides robustness against rotation in depth up to about 20 degrees. More drastic rotations have to be handled by a new two-dimensional view of that different pose. For a reasonable comparison of jets one has to define grids of fiducial points adapted to the specific object. The frontal view graph and half profile graph consequently have different structure and geometry, but for most of the nodes in one pose there is a corresponding node in the other pose, referring to the same fiducial point. The structure of these graphs and the links between the nodes belonging to the same fiducial point are defined by hand. Once a minimal general face knowledge for both poses is established, the very same matching process as described above is applied and further model graphs can be generated automatically. We used a basic general face knowledge of 70 manually checked models per pose to build larger galleries automatically.

The linear scale of the faces in the original images varied by about a factor three. A preprocessing phase was necessary to rescale the faces to a normalized size. Different general face knowledges with a few models of small, middle, or large faces were matched to the original images. The match with the highest similarity value was evaluated. The distance between top and bottom node leads to an appropriate scaling

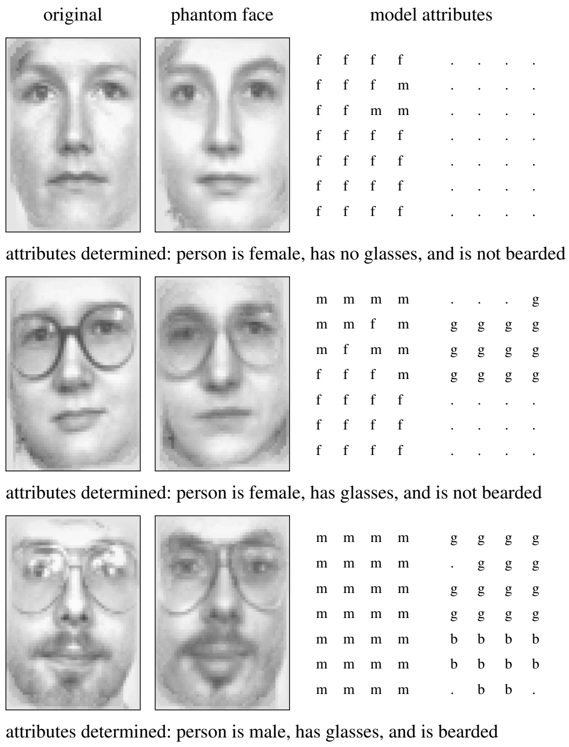

Figure 4: Shown is the original and the phantom face for three different persons. Notice that the phantom image was generated only on the basis of information provided by the match with the general face knowledge; no information from the original image was used. That is the reason why certain details, such as the reflections on the glasses or the precise shape of the lips of the top image are not reproduced accurately. The fields of labels on the right side indicate the features of the models that were used as experts for the individual nodes; m: male, f: female, b: bearded, g: glasses.

factor and the center of the graph serves as center for the rescaled image. For this preprocessing, graphs with a different grid structure were used. Nodes were positioned at points easy to find but not necessarily reliabe for recognition, e.g., the outline of the head, see figure 5. The pose of the faces was known *a priori* and needed not to be determined automatically.

Once an image graph is generated by graph matching with the general face knowledge, it can be compared to individual model graphs of a gallery without further distortion, just by pointwise comparison of jets. The topographical information is not used. Hence for the recognition task the similarity of two graphs is simply defined as the average similarity between their jets:

$$S_r(G^M, G^I) = \frac{1}{N} \sum_n S_a(J_n^M, J_n^I)$$

Here jet similarities S_a based only on the magnitudes turned out to be more discriminative than the similarities S_ϕ, which include phase.

We tested the system on the ARPA/ARL FERET

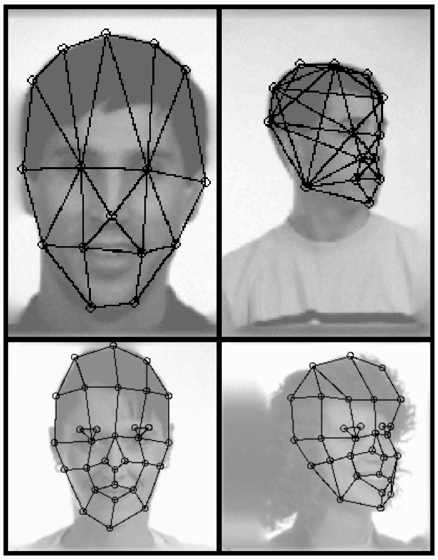

Figure 5: Object adapted graphs for frontal and half profile view. The nodes are positioned automatically by elastic graph matching. The two top images show two original images with large size variation and grids for preprocessing with many nodes on the outline. The two bottom images are already rescaled to normal size. Those grids have more nodes on the face, which is more appropriate for recognition. For the recognition results given below grids with 48 and 46 nodes were used.

database by comparing frontal against frontal views and half profile against frontal views. The two frontal views differed in facial expression and the half profile pose was rotated by about 40 to 70 degrees, in some cases turning almost to profile view. In the first test we compared 300 frontal views against 300 different frontal views of the same persons and achieved a recognition rate of 97.3%. 99.0% were among the first 15 best matches. In a second test we compared 300 half profiles against 300 frontal views of the same persons with a recognition rate of 13.3%. 44.0% were among the first 15 best matches.

The performance is high on frontal views and it was shown that the system is robust with respect to rotations in depth up to 20 degrees [3]. The results are poor for faces of very different pose, which is known to be a much more difficult task for human subjects as well [15]. Nevertheless, using different two-dimensional views for different poses plus the information which nodes in the different views belong to the same fiducial point makes it possible to apply more sophisticated methods to deal with the rotation transformation, as shown in another contribution [16].

7 Conclusions

Based on the system described in [3] we have made three major modifications of which only the last one is restricted to the in-class recognition task, i.e., a task in which objects belonging to one known class have to be recognized.

Phase information was used for a more accurate positioning of the nodes at the fiducial points.

Object adapted graphs were introduced to deal with different views. The nodes then are related to fiducial object points and the graph geometry changes depending on the 3D structure of the object.

The general face knowledge is the only new concept tailored to face recognition or rather in-class recognition. By combining jets of a relatively small set of model graphs, a large face space can be covered.

The modified system has several advantages. Firstly the previous system [3] matched each model of the gallery separately to the face image. By introducing the general face knowledge and by using phase information, image graphs can be generated with no model knowledge about the individual persons. This allows separating the graph generation phase from the recognition phase, which makes the system much faster by generating an image graph only once and not for each model repeatedly.

Secondly the object adapted graphs provide means to deal with a set of different poses. Nodes can refer to the same fiducial points regardless of viewing direction. It also becomes possible to focus on points of special interest or reliability.

Thirdly the use of phase information provides relatively precise node locations that can potentially be used as an additional recognition or feature determination cue. So far only the jets are evaluated. Previously the localization of the nodes was very rough and of little use for the recognition.

The system requires some manual control when generating a general face knowledge. Apart from this, no training is required to build a gallery of new faces to recognize. The models are generated automatically, stored, and compared by a simple similarity function. Only one model per person is required. Nevertheless, different kinds of learning can be introduced. Experiments have been made with jet transformations to account for rotation in depth [16] and with local weights to emphasize reliable nodes [17].

Acknowledgments

We wish to thank Irving Biederman, Ladan Shams, Michael Lyons, and Thomas Maurer for very fruitful discussions and their help in evaluating the performance of the system. Portions of the research in this paper use the FERET database of facial images collected under the ARPA/ARL FERET program.

References

[1] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

[2] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[3] M. Lades et al. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput.*, 42(3):300–311, 1993.

[4] I. Craw, H. Ellis, and J.R. Lishman. Automatic extraction of face features. *Pattern Recognition Letters*, 5:183–187, 1987.

[5] C. von der Malsburg. The correlation theory of brain function. Internal report, 81-2, Max-Planck-Institut für Biophysikalische Chemie, Postfach 2841, 3400 Göttingen, FRG, 1981.

[6] C. von der Malsburg. Nervous structures with dynamical links. *Ber. Bunsenges. Phys. Chem.*, 89:703–710, 1985.

[7] E. Bienenstock and C. von der Malsburg. A neural network for invariant pattern recognition. *Europhysics Letters*, 4:121–126, 1987.

[8] L. Wiskott and C. von der Malsburg. A neural system for the recognition of partially occluded objects in cluttered scenes. *Int. J. of Pattern Recognition and Artificial Intelligence*, 7(4):935–948, 1993.

[9] J. Buhmann, M. Lades, and C. von der Malsburg. Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings of the IJCNN International Joint Conference on Neural Networks*, pages II 411–416, San Diego, June 1990. IEEE.

[10] J.P. Jones and L.A. Palmer. An evaluation of the two dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, 58:1233–1258, 1987.

[11] R.L. DeValois and K.K. DeValois. *Spatial Vision.* Oxford Press, 1988.

[12] W. M. Theimer and H. A. Mallot. Phase-based binocular vergence control and depth reconstruction using active vision. *CVGIP: Image Understanding*, 60(3):343–358, November 1994.

[13] R. Brunelli and T. Poggio. Caricatural effects in automated face perception. *Biological Cybernetics*, 69:235–241, 1993.

[14] B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski. Sexnet: a neural network identifies sex from human faces. In D.S. Touretzky and R. Lippman, editors, *Advances in Neural Information Processing Systems 3*. Morgan Kaufmann, SanMateo,CA, 1991.

[15] P. Kalocsai, I. Biederman, and E.E. Cooper. To what extent can the recognition of unfamiliar faces be accounted for by a representation of the direct output of simple cells. In *Proceedings of ARVO*, May 1994.

[16] T. Maurer and C. von der Malsburg. Single-view based recognition of faces rotated in depth. In *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition*, Zürich, 1995.

[17] N. Krüger. Learning Weights in Discrimination Functions using a priori Constraints. Submitted to 17. Symposium der deutschen Arbeitsgemeinschaft für Mustererkennung (DAGM), Bielefeld, Germany, September 1995.