To be published by IEEE Computer Society Press in:
Proceedings of the Second International Conference on
Automatic Face and Gesture Recognition '96, October
14-16 1996, Killington, Vermont, USA.

Robust Classification of Hand Postures against Complex Backgrounds

Jochen Triesch and Christoph von der Malsburg[1]
Institut für Neuroinformatik, System-Biophysik
Ruhr-Universität Bochum
D-44780 Bochum, Germany
email: jochen@neuroinformatik.ruhr-uni-bochum.de
[1]also at: University of Southern California
Dept. of Computer Science and Section for Neurobiology
Los Angeles, CA, USA

Abstract

A system for the classification of hand postures against complex backgrounds in grey-level images is presented. The system employs elastic graph matching, which has already been successfully employed for the recognition of faces. Our system reaches 86.2% correct classification on our gallery of 239 images of ten postures against complex backgrounds. The system is robust with respect to certain variations in size of hand and shape of posture.

1. Introduction

There are at least two possible applications for automatic classification of hand postures from video images: Firstly, new man-machine interfaces may be devised, which free the human user from cumbersome input devices such as keyboards, mice, remote controls and so on (for review see [3]). Secondly, the automatic translation of gesture based natural languages (e.g. the American Sign Language) is a long-term goal. Furthermore, a new generation of intelligent robots may learn how to handle objects in its environment by watching human subjects (or other robots) manipulating them. In all of these domains it is important that the system is able to recognize objects despite variations in the image background. A system demanding uniform background is not flexible enough for most real-world applications.

The system presented here employs elastic graph matching for the classification of hand postures in grey-scale images. Graph matching has already been successfully applied to other computer vision tasks, e.g. object recognition and face recognition [7] [11].

2. Object representation with labeled graphs

Hand postures are represented as labeled graphs with an underlying two-dimensional topology. Nodes of the graph are labeled with a local image description called *jet*. Edges are labeled by a distance vector. Figure 1 depicts a graph superimposed on the original image. The jets are based on

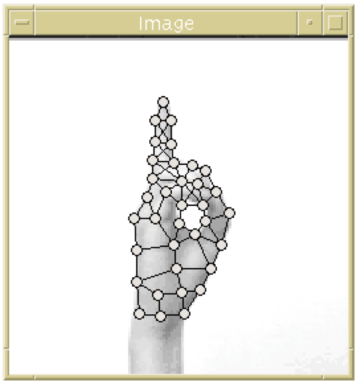

Figure 1. Hand postures are represented by labeled graphs. Attatched to the nodes are so-called jets — local image descriptions based on Gabor-like filters. The edges are labeled with geometrical information.

a wavelet transform with complex Gabor-based kernels:

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2} \exp\left(-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}\right) \left[\exp\left(i\vec{k}\vec{x}\right) - \exp\left(\frac{-\sigma^2}{2}\right)\right] .$$

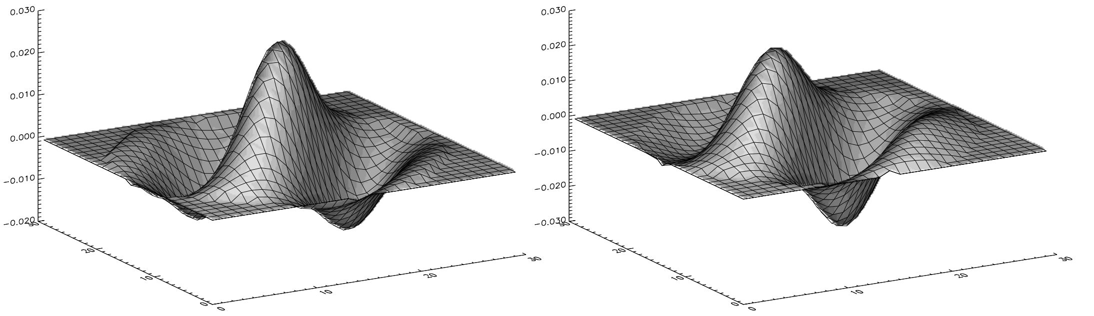

Figure 2. Gabor-based filters are used to extract jet components. The filters have the form of plane waves restricted by Gaussian envelope functions. Left: real part, right: imaginary part. We use filters of three different sizes and eight orientations. The shape of these filters resembles the receptive fields of neurons in the visual cortex of mammals.

Gabor filters are known to resemble the receptive fields of neurons in the primary visual cortex of mammals [5]. Our filters are DC-free and hence their responses are invariant with respect to constant offsets in the grey-level values of an image. They have the form of a plane wave with wave-vector $\vec{k}$ restricted by a Gaussian envelope function (see figure 2). A jet is composed of the results of convolutions with several kernels of different wave-vector $\vec{k}$:

$$\vec{k}_{\nu\mu} = k_\nu e^{i\phi_\mu} \quad \text{with} \quad k_\nu = k_{max}/f^\nu,\ \phi_\mu = \frac{\pi\mu}{8},$$

We use $\nu \in \{0, 1, 2\}$ and $\mu \in \{0, \ldots, 7\}$. The value f is the spacing factor between kernels in the frequency domain. It was chosen to be $1/\sqrt{2}$ with $k_{max} = 1.7$. The width of the Gaussian envelope function is given by $\sigma / \left|\vec{k}\right|$ with $\sigma = 2.5$. A jet is a complex vector composed of the 24 complex filter responses.

Elastic matching of a model graph M to an image means to search for a set $\vec{x}^N$ of node positions simultaneously satisfying two constraints: The local image information attached to each node must match the image region around the position where the node is placed. The distances between the matched node positions must not differ too much from the original distances. We have expressed these demands by the definiton of similarity functions for the nodes and a cost function for the edges of the matched graph.

In order to allow for comparison of a graph's jets to points in an image, we compute jets for every point in the image and compare them to the graph's jets using two similarity functions with different properties [10]:

1. Jet similarity using only magnitudes of the complex filter responses: A jet is a vector of 24 complex numbers $J_j, \quad j \in \{0, \ldots, 24\}$, whose components may be written as $J_j = a_j exp(i\phi_j)$. We define

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

2. Jet similarity using magnitude and phase of the complex filter responses:

$$S_{\text{pha}}(J, J') = \frac{1}{2}\left(1 + \frac{\sum_j a_j {a'}_j \cos(\phi_j - \phi'_j)}{\sqrt{\sum_j a_j^2 \sum_j {a'_j}^2}}\right)$$

Both functions yield similarity values between zero and one. While $S_{\text{abs}}(J, J')$ slowly changes when J' is moved across the image, $S_{\text{pha}}(J, J')$ varies very rapidly because the phases of filter responses change significantly on a spatial scale corresponding to the wave-vector $\vec{k}$ of strongly responding kernels.

3. Bunch-graphs of hand postures

Our aim is the classification of hand postures against complex backgrounds. As the hand may appear in front of lighter background in some parts of the image and against darker background in others, we use the concept of *bunch-graphs* for the representation of hand postures [12] [11] [10]. The idea behind the bunch-graph concept is to express the natural variability in the jets of corresponding points in several images (e.g. several fingertips in front of light or dark background) by labeling each node with a whole collection of jets rather than only a single jet.

For one image of each posture, a graph is created manually. All graphs have 35 nodes and 70 edges. Node positions are chosen to lie on the rim of the hand and on highly textured postions within the hand. An example of a model graph is shown in figure 1. For five other examples of every posture, a graph is constructed in a semi-automatic way. The already existing graph is matched to the image (see below) as a first guess about the correct node positions for the other images. Then, single node postions are corrected by hand if they fall on wrong postions during the matching process. In this way, for each posture two graphs are constructed for three different persons — one in front of a light background, and one in front of a dark background. These six graphs for every posture are then fused into a single bunch-graph (figure 3). The lengths of corresponding edges in the six graphs are averaged, and the node information B^N attatched to each node of the bunch-graph is defined as the set of the six jets of that node N:

$$B^N = \{J^N(1), \ldots, J^N(6)\}$$

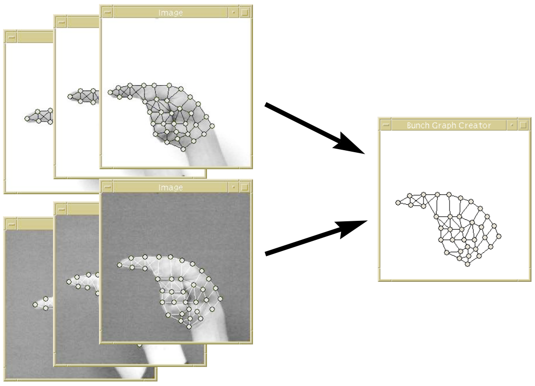

Figure 3. Creation of a bunch-graph from six single graphs. The bunch-graph geometry is averaged from the six graphs, the features attached to each node are sets of six jets composed of the jets of the six graphs.

For the matching process, we need a similarity function comparing the set of jets attatched to each node of the bunch-graph with local image information in a picture. We define the similarity of a set of jets B^N to a single jet $J(\vec{x})$ taken at a point $\vec{x}$ in an image to be the maximum of the similarities of the six single jets.

$$\begin{aligned}
\tilde{S}_{\text{abs}}(B^N, J(\vec{x})) &= \max_j \left\{S_{\text{abs}}(J^N(j), J(\vec{x}))\right\}, \\
\tilde{S}_{\text{pha}}(B^N, J(\vec{x})) &= \max_j \left\{S_{\text{pha}}(J^N(j), J(\vec{x}))\right\} \\
j &= 1, \ldots, 6.
\end{aligned}$$

When a graph G with P nodes is matched to an image at the node positions $\vec{x}^N$, its total similarity is given by the average of all its nodes:

$$\begin{aligned}
\hat{S}_{\text{abs}} &= \frac{1}{P}\sum_N \tilde{S}_{\text{abs}}(B^N, J(\vec{x}^N)) \\
\hat{S}_{\text{pha}} &= \frac{1}{P}\sum_N \tilde{S}_{\text{pha}}(B^N, J(\vec{x}^N))
\end{aligned}$$

During the matching process the model graph may be distorted in order to compensate variations in hand shape. However, big distortions shall be punished by a topological cost term. The cost for a single edge E is defined as:

$$C_{\text{edge}}(E) = \left(\frac{|\text{ original length} - \text{distorted length }|}{\text{original length}}\right)^2$$

We use the square of the relative change in length, because we do not want to punish very small changes but prohibit large distortions. The topological costs of a graph with M edges matched to an image are defined as the average cost for all individual edges:

$$C_{\text{topol}} = \frac{1}{M}\sum_i C_{\text{edge}}(E_i)$$

4. Matching process

As outlined above, elastic matching of a model graph M onto an image means the finding of image node positions $\vec{x}^N$ which yield high image similarities and low topological costs. The matching process operates in three steps.

1. Coarse positioning of the graph: The image is scanned in coarse steps of five pixels in x and y direction without graph distortion. The local image similarities are computed without the phase information of the filter responses, i.e. using $\hat{S}_{\text{abs}}$ as similarity function.

2. Rescaling of the graph: The graph is allowed to grow or shrink by up to 20% as a whole without relative changes of the edge lengths. This step compensates for different sizes of subjects' hands and different distances from the camera. Additionally, the graph is allowed to shift its position by up to twelve pixels in x and y direction.The local image similarities are again computed using $\hat{S}_{\text{abs}}$.

3. Local diffusion of single nodes: All nodes may shift their positions by up to four pixels, using phase-sensitive node similarities and topological costs:

   $$\hat{S}_{\text{total}} = \hat{S}_{\text{pha}} - \lambda\, C_{\text{topol}}.$$

   The coefficient λ controls the rigidity of the image graph, large values penalizing distortions more heavily; after some experiments we settled for a value of

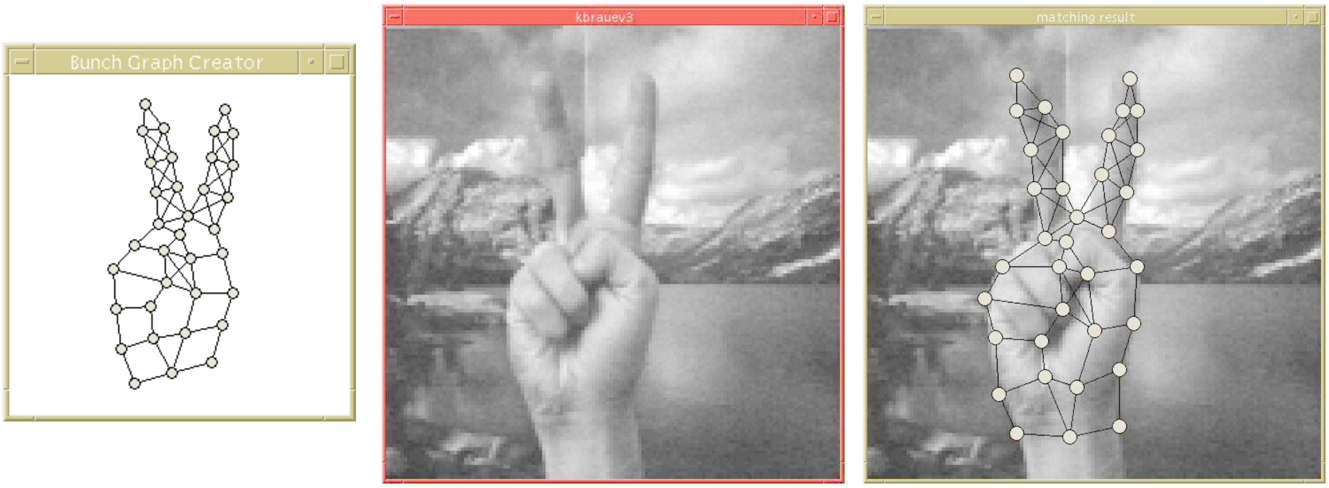

Figure 4. Matching of a model graph to an image. The model graph is allowed to distort during match. Not all nodes are placed perfectly. However, classification is robust with respect to such little imperfections.

$\lambda = 0.3$. The algorithm optimizes the positions of the nodes one after the other. A node is moved exhaustively in a 9 times 9 pixels area around its original postition after the rescaling step. The overall similarity $\hat{S}_{\text{total}}$ for each position is computed and the position with highest similarity is chosen. Then the next node is moved.

A typical result of a match is given in figure 4. Not all nodes are placed perfectly. However, the total similarity $\hat{S}_{\text{total}}$ is usually high enough to allow correct classification.

For an image to be classified the bunch-graphs of all postures are matched to the image, yielding a set of similarity values. The posture of the model graph with the highest similarity is chosen. Our current system sequentially matches bunch graphs for all postures. However, it may be parallelized with adequate hardware [13].

5. Experiments

Our gallery consists of 10 hand signs performed by 24 persons against three backgrounds. The ten hand postures are depicted in figure 5. We recorded 8-bit grey-scale images of 128^2 pixels. For each person the ten postures were recorded in front of uniform light, uniform dark and complex background giving 720 images of which three were lost. Examples of the complex backgrounds are given in figure 6. Note the variabilities in size and shape of the hand posture. The images of three persons against light and dark background (60 images) were the model set and served for the generation of the model graphs as described above. The remaining pictures constituted the test set. Thus model set and test set were disjunct, although the pictures of the three model subjects taken against complex background entered into the test set.

5.1. Classification against uniform background

We tested the system at recognizing the hand postures in front of uniform background. Of the 210 hands against light background not included in the model set 198 were recognized correctly which corresponds to a rate of 94.3%. For the dark background we reached a recognition of 194 out of 208, which is 93.3%. These results are not particularly good considering the simplicity of the task. Note however, that our system is not specialized for these circumstances. The errors are due to big posture variations between subjects.

5.2. Classification against complex background

Recognition against complex background is more difficult for several reasons. In a single image the hand may be seen against lighter background in some places and seen against darker background in other places. Also, the boundary of the hand may be undetectable where hand and background have the same grey-level value. Additionally, parts of the background may often be false targets and can easily be mistaken as parts of the hand. We use the bunch-graph concept to overcome these difficulties. The fused bunch-graphs contain local image descriptions of the hand against light and dark background.

On our test set of 239 images of the ten postures against complex backgrounds our system correctly recognizes 206 images, which corresponds to a recognition rate of 86.2%.

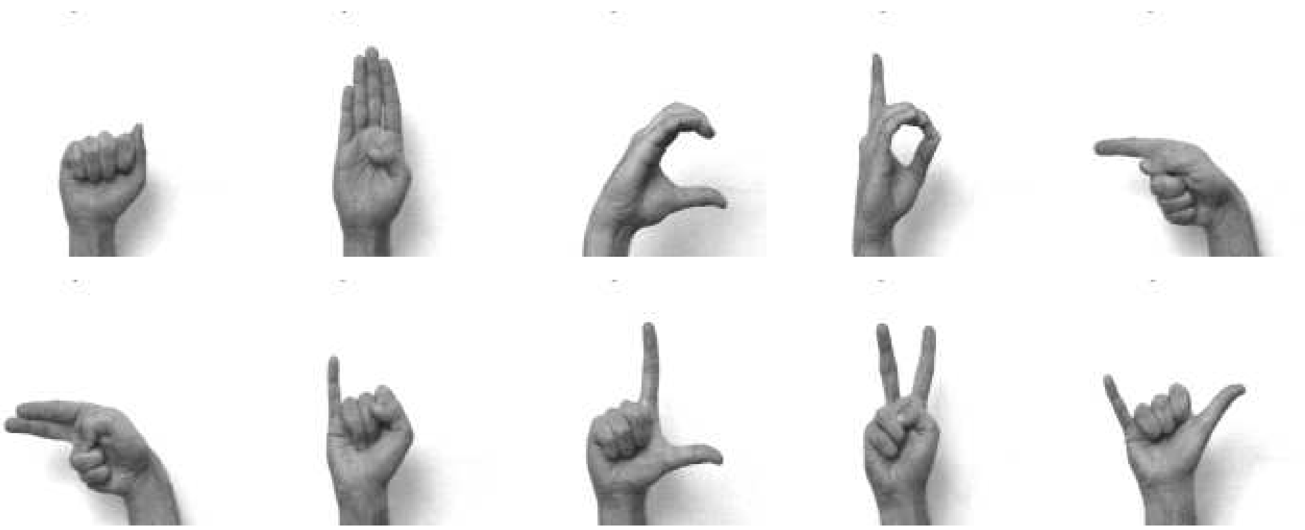

Figure 5. Each subject signed against light, dark and complex background. The image shows the ten signs performed by one subject against uniform light background.

Recognition results

background	number	correct	percentage
complex	239	206	86.2
light	210	198	94.3
dark	208	194	93.3
total	657	598	91.0

Table 1. Recognition results. Correct recognition rate against uniform light and dark background is about 94%. Most errors occur when the system confuses postures five and six.

For the misclassified images, errors occur at all three matching steps:

1. Positioning of the graph may be wrong for the correct hand posture.

2. The graph may be positioned correctly, but the following rescaling brings it to a wrong size.

3. The graph is placed and rescaled correctly, but single nodes are too far away from their correct positions due to big variations in hand shape. Nodes then diffuse to wrong locations during the final matching step.

The results are summarized in table 1.

5.3. Computation times

In its present guise, the system is slow. The Gabor transformation of an image takes 1.85 s on a SUN UltraSPARC-I workstation (167 MHz). The complete matching of a single bunch-modelgraph onto an image takes 1.49 s:

1. Coarse Positioning of the graph: 0.22 s

2. Rescaling of the graph: 0.58 s

3. Local diffusion of single nodes: 0.69 s

These long times are due to the fact that during all matching steps a brute force exhaustive search in a confined region is performed. Application of an appropriate search technique would certainly speed up the system. Note that for classification a model-graph has to be matched on the image for every posture.

6. Discussion

Elastic graph matching has been applied to recognize the posture of a hand against complex background in grey-level images. The bunch-graph concept has been successfully applied to overcome the problem of varying backgrounds. Note that our approach does not employ a seperate segmentation mechanism for the localization of the hand based on grey-levels, stereo-vision, motion or other. Incorporation of these will certainly improve our system's performance. But even without them, our system's recognition is very robust. The main problem we have to cope with are the large variations in the shape of hands referring to the same posture.

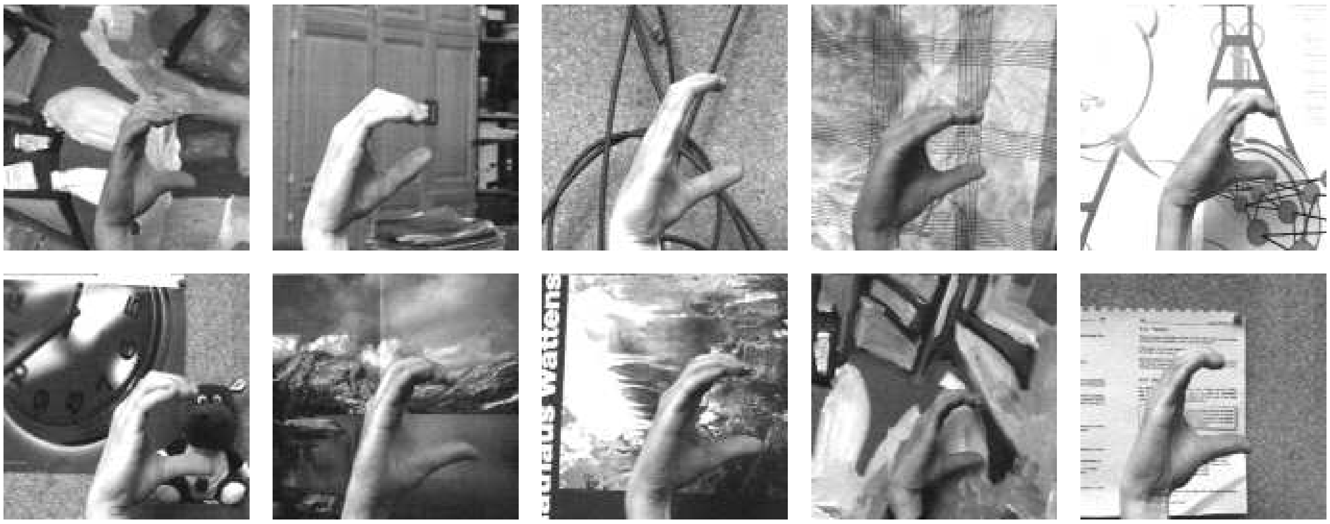

Figure 6. Pose number three performed by 10 different subjects in front of ten of the used complex backgrounds. Complexity of the backgrounds varies, as does the size and shape of the hand. Nevertheless, our system reaches 86.2% correct classification on this set of data.

These variations require the bunch-graph to be greatly distorted. By reducing topological costs big distortions become possible, but we also have more false targets for every node.

In contrast to other systems for hand posture recognition from grey-level images, e.g. [9] [1] [4] [2], we are able to cope with complex backgrounds. We believe that for most real world applications a uniform background is too strong a limitation. However, there exists work about the detection or tracking of hands in front of complex backgrounds [8] [6]. In order to better compare the performance of different systems, it is desirable to use common databases.

Future work will deal with the recognition of gestures, i.e. hand and arm movements in sequences of several images.

References

[1] Y. Cui and J. Weng. Learning-based hand sign recognition. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[2] W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[3] T. S. Huang and V. I. Pavlović. Hand gesture modeling, analysis, and synthesis. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[4] E. Hunter, J. Schlenzig, and R. Jain. Posture estimation in reduced-model guesture input systems. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[5] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[6] C. Kervrann and F. Heitz. Learning structure and deformation modes of nonrigid objects in long image sequences. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[7] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.

[8] B. Moghaddam and A. Pentland. Maximum likelihood detection of faces and hands. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[9] C. Uras and A. Verri. Hand gesture recognition from edge maps. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[10] L. Wiskott. *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*, volume 53 of *Reihe Physik*. Verlag Harri Deutsch, Thun, Frankfurt a. Main, Germany, 1995. PhD thesis.

[11] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition and gender determination. Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, June 26-28, 1995.

[12] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic graph matching. Submitted to IEEE-PAMI, 1996.

[13] R. P. Würtz, J. C. Vorbrüggen, and C. von der Malsburg. A transputer system for the recognition of human faces by labeled graph matching. In R. Eckmiller, G. Hartmann, and G. Hauske, editors, *Parallel Processing in Neural Systems and Computers*, pages 37–41. North Holland, Amsterdam, 1990.