IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):775-779 (1997).

Face Recognition by Elastic Bunch Graph Matching

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

Abstract

We present a system for recognizing human faces from single images out of a large database containing one image per person. Faces are represented by *labeled graphs*, based on a Gabor wavelet transform. Image graphs of new faces are extracted by an elastic graph matching process and can be compared by a simple similarity function. The system differs from the preceding one [1] in three respects. Phase information is used for accurate node positioning. Object-adapted graphs are used to handle large rotations in depth. Image graph extraction is based on a novel data structure, the *bunch graph*, which is constructed from a small set of sample image graphs.

Index Terms: face recognition, different poses, Gabor wavelets, elastic graph matching, bunch graph, ARPA/ARL FERET database, Bochum database.

1 Introduction

The system presented here is based on a face recognition system described in [1]. In this preceding system, individual faces were represented by a rectangular graph, each node labeled with a set of complex Gabor wavelet coefficients, called a *jet*. Only the magnitudes of the coefficients were used for matching and recognition. When recognizing a face of a new image, each graph in the model gallery (database) was matched to the image separately and the best match indicated the recognized person. Rotation in depth was compensated for by elastic deformation of the graphs.

We have made three major extensions to this system in order to handle larger galleries and larger variations in pose, and to increase the matching accuracy, which provides the potential for further techniques to improve recognition rate.

- Firstly, we use the phase of the complex Gabor wavelet coefficients to achieve a more accurate location of the nodes and to disambiguate patterns which would be similar in their coefficient magnitudes.
- Secondly, we employ object adapted graphs, so that nodes refer to specific facial landmarks, called *fiducial points*. The correct correspondences between two faces can then be found across large viewpoint changes.
- Thirdly, we have introduced a new data structure, called the *bunch graph*, which serves as a generalized representation of faces by combining jets of a small set of individual faces.

This allows the system to find the fiducial points in one matching process, which eliminates the need for matching each model graph individually. This reduces computational effort significantly. A more detailed description of this system is given in [2].

2 The System

2.1 Jets

A jet is based on a wavelet transform, defined as a convolution of the image with a family of *Gabor kernels* [3]

$$\psi_j(\vec{x}) = \frac{k_j^2}{\sigma^2} \exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right) \left[\exp(i\vec{k}_j\vec{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (1)$$

in the shape of plane waves with wave vector $\vec{k}_j$, restricted by a Gaussian envelope function with relative width $\sigma = 2\pi$. We employ a discrete set of 5 different spatial frequencies and 8 orientations. For images of size 128×128 pixels, the lowest and highest frequency have a wavelength of 16 and 4 pixels, respectively. The last term in Equation (1) makes the kernels *DC-free*, i.e. the

[0] Laurenz Wiskott was with the Institute for Neural Computation, Ruhr-University Bochum, D-44780 Bochum, Germany (http://www.neuroinformatik.ruhr-uni-bochum.de) when this research was performed. He is now at the Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, San Diego, CA 92186-5800. E-mail: wiskott@cnl.salk.edu.

[0] Jean-Marc Fellous was with the Computer Science Department, University of Southern California, Los Angeles, CA 90089 when this research was performed. He is now at the Volen Center for Complex Systems, Brandeis University, Waltham, MA 02254-9110. E-mail: fellous@cajal.ccs.brandeis.edu.

[0] Norbert Krüger and Christoph von der Malsburg are with the Institute for Neural Computation, Bochum. Christoph von der Malsburg is also with the Computer Science Department, University of Southern California, Los Angeles. E-mail: {nkrueger, malsburg}@neuroinformatik.ruhr-uni-bochum.de.

integral $\int \psi_j(\vec{x}) d^2\vec{x}$ vanishes. This is known as a wavelet transform because the family of kernels is self-similar, all kernels being generated from one *mother wavelet* by dilation and rotation.

A jet $\mathcal{J}$ is defined as the set $\{\mathcal{J}_j\}$ of 40 complex Gabor wavelet coefficients obtained for one image point. It can be written as $\mathcal{J}_j = a_j \exp(i\phi_j)$ with magnitudes $a_j(\vec{x})$, which slowly vary with position, and phases $\phi_j(\vec{x})$, which rotate with a rate set by the spatial frequency or wave vector $\vec{k}_j$ of the kernels. Due to this phase rotation, jets taken from image points only a few pixels apart have very different coefficients, although representing almost the same local feature. This can cause severe problems for matching. We therefore either ignore the phase or compensate for its variation explicitly. The similarity function

$$\mathcal{S}_a(\mathcal{J}, \mathcal{J}') = \frac{\sum_j a_j a'_j}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}} \tag{2}$$

ignores phase [1]. With a jet $\mathcal{J}'$ taken at a fixed image position and jets $\mathcal{J} = \mathcal{J}(\vec{x})$ taken at variable position $\vec{x}$, $\mathcal{S}_a(\mathcal{J}(\vec{x}), \mathcal{J}')$ is a smooth function with local optima forming large attractor basins (see Figure 1), leading to rapid and reliable convergence with simple search methods such as gradient descent or diffusion.

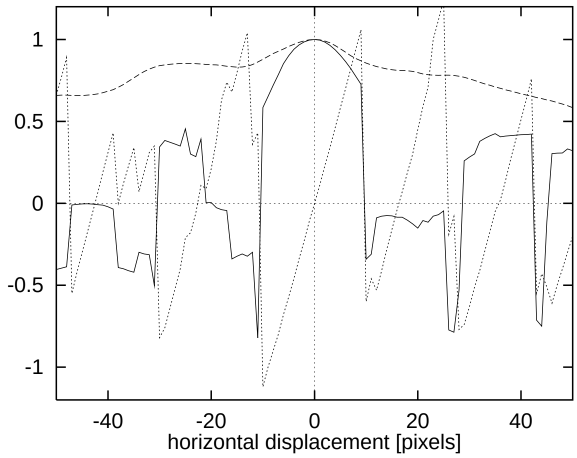

Figure 1: Similarities $\mathcal{S}_a(\mathcal{J}, \mathcal{J}')$ (dashed line) and $\mathcal{S}_\phi(\mathcal{J}, \mathcal{J}')$ (solid line) with jet $\mathcal{J}'$ taken from the left eye of a face, and jet $\mathcal{J}$ taken from pixel positions of the same horizontal line. The dotted line shows the estimated displacement $\vec{d}(\mathcal{J}, \mathcal{J}')$ (divided by 8 to fit the ordinate range). The right eye is 24 pixels away from the left eye, generating a local maximum for both similarity functions and zero displacement close to $d_x = -24$.

Using phase has two advantages. Firstly, phase information is required to discriminate between patterns with similar magnitudes, should they occur. Secondly, since phase varies so quickly with location, it provides a means for accurate jet localization in an image. Assuming that two jets $\mathcal{J}$ and $\mathcal{J}'$ refer to object locations with small relative displacement $\vec{d}$, the phase shifts can be approximately compensated for by the terms $\vec{d}\vec{k}_j$, leading to a phase-sensitive similarity function

$$\mathcal{S}_\phi(\mathcal{J}, \mathcal{J}') = \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \vec{d}\vec{k}_j)}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}}. \tag{3}$$

In order to compute it, the displacement $\vec{d}$ has to be estimated. This can be done by maximizing $\mathcal{S}_\phi$ in its Taylor expansion around $\vec{d} = 0$, which is a constrained fit of the two-dimensional $\vec{d}$ to the 40 phase differences $\phi_j - \phi'_j$ [2, 4]. Large displacements of up to 8 pixels can be estimated if the phases of higher frequency coefficients are corrected by multiples of 2π depending on the disparity estimated from lower frequency coefficients. It is a great advantage of this second similarity function that it yields this displacement information. Profiles of similarities and estimated displacements are shown in Figure 1.

2.2 Graphs

A *labeled graph* $\mathcal{G}$ representing a face consists of N nodes connected by E edges. The nodes are located at facial landmarks $\vec{x}_n, n = 1, ..., N$, called *fiducial points*, e.g. the pupils, the corners of the mouth, the tip of the nose, the top and bottom of the ears, etc. This face graph is *object-adapted* since its geometrical structure is adapted to the structure of the object (see Figure 2). The nodes are labeled with jets $\mathcal{J}_n$. The edges are labeled with two-dimensional distance vectors $\Delta\vec{x}_e = \vec{x}_n - \vec{x}_{n'}, e = 1, ..., E$, where edge e connects node n' with n. (We refer to the geometrical structure of a graph, unlabeled by jets, as a *grid*.) Graphs for different head pose differ in geometry and local features (jets). Although the fiducial points refer to corresponding object locations, some may be occluded, and jets as well as distances vary due to rotation in depth. To be able to compare graphs of different poses, we manually defined pointers to associate corresponding nodes in the different graphs.

In order to extract image graphs automatically for new faces, one needs a general representation rather than models of individual faces. This representation should cover a wide range of possible variations in the appearance of faces, such as differently shaped eyes, mouths, or noses, different types of beards, variations due to sex, age, and race etc. It is obvious that it would be too expensive to cover each feature combination by a separate graph. We instead combine a representative set of M individual model graphs $\mathcal{G}^{\mathcal{B}m}$ $(m = 1, ..., M)$ into a stack-like structure, called a *face bunch graph* (FBG) (see Figure 3).

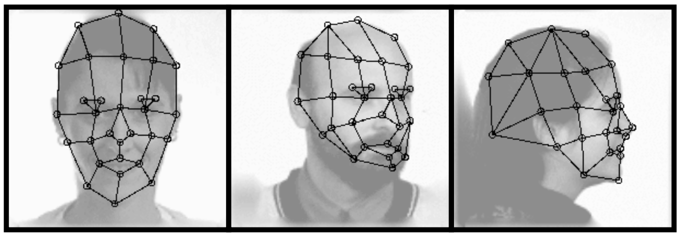

Figure 2: Object-adapted grids for different poses. The nodes are positioned automatically by elastic bunch graph matching. (The grids used in Section 3 for the FERET database had about 14 additional nodes which are not shown here for simplicity). One can see that, in general, the matching finds the fiducial points quite accurately. But mispositioning occurred, for example, for the face in the center. The chin was not found accurately; the leftmost node and the node below should be at the top and the bottom of the ear respectively.

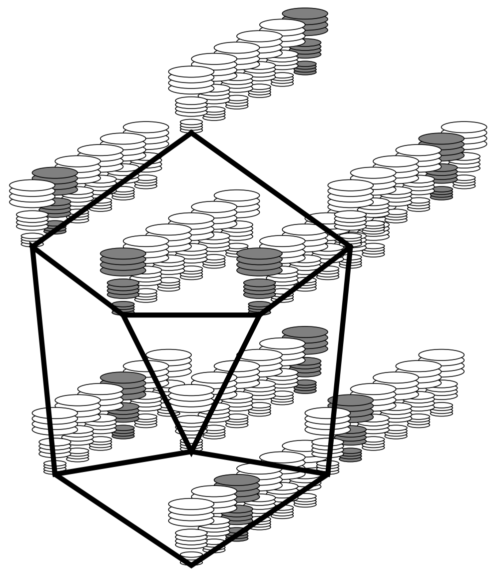

Figure 3: The Face Bunch Graph (FBG) serves as a general representation of faces. Each stack of discs represents a jet. From a bunch of jets attached to a single node only the best fitting one is selected for a match, indicated by grey shading.

Each model graph has the same grid structure and the nodes refer to identical fiducial points. A set of jets referring to one fiducial point is called a *bunch*. An eye bunch, for instance, may include jets from closed, open, female, and male eyes etc. to cover these local variations. The corresponding FBG $\mathcal{B}$ is then given the same grid structure as the individual graphs, its nodes are labeled with the bunches of jets $\mathcal{J}_n^{\mathcal{B}m}$ and its edges are labeled with the averaged distances $\Delta\vec{x}_e^{\mathcal{B}} = \sum_m \Delta\vec{x}_e^{\mathcal{B}m}/M$. During the location of fiducial points in a new image of a face, the procedure described below selects the best fitting jet, called the *local expert*, from the bunch dedicated to each fiducial point. Thus, the full combination of jets in the bunch graph is available, covering a much larger range of facial variation than represented in the constituting model graphs themselves.

2.3 Elastic Bunch Graph Matching

A first set of graphs is generated manually. Nodes are located at fiducial points and edges between the nodes as well as correspondences between nodes of different poses are defined. Once the system has an FBG (possibly consisting of only one manually defined model), graphs for new images can be generated automatically by elastic bunch graph matching. Initially, when the FBG contains only few faces, it is necessary to review and correct the resulting matches, but once the FBG is rich enough (approximately 70 graphs) one can rely on the matching and generate large galleries of model graphs automatically. Matching a FBG on a new image is done by maximizing a *graph similarity* between an image graph and the FBG of identical pose. It depends on the jet similarities and a topography term, which takes into account the distortion of the image grid relative to the FBG grid. For an image graph $\mathcal{G}^{\mathcal{I}}$ with nodes $n = 1, ..., N$ and edges $e = 1, ..., E$ and an FBG $\mathcal{B}$ with model graphs $m = 1, ..., M$ the similarity is defined as

$$\begin{aligned} \mathcal{S}_{\mathcal{B}}(\mathcal{G}^{\mathcal{I}}, \mathcal{B}) &= \frac{1}{N}\sum_n \max_m \left(\mathcal{S}_\phi(\mathcal{J}_n^{\mathcal{I}}, \mathcal{J}_n^{\mathcal{B}m})\right) \\ &\quad -\frac{\lambda}{E}\sum_e \frac{(\Delta\vec{x}_e^{\mathcal{I}} - \Delta\vec{x}_e^{\mathcal{B}})^2}{(\Delta\vec{x}_e^{\mathcal{B}})^2}, \end{aligned} \tag{4}$$

where λ determines the relative importance of jet similarities and the topography term. $\mathcal{J}_n$ are the jets at node n and $\Delta\vec{x}_e$ are the distance vectors used as labels at edges e. Since the FBG provides several jets for each fiducial point, the best one is selected and used for comparison. These best fitting jets serve as *local experts* for the image face. A heuristic algorithm is used to find the image graph which maximizes the graph similarity function. First, the location of the face is found by a sparse scanning of the FBG over the image. Then, the FBG is varied in size and aspect ratio to adapt to the right format of the face. These steps are of no cost in the topography term of the similarity function because the edge labels are transformed accordingly. Finally all nodes are moved locally and relative to each other to optimize the graph similarity further. Only node locations with small estimated disparity are considered. This local distortion is constrained by the topography term.

Since in the FERET database faces vary in size by a factor of three, the matching is done twice. In the first matching step the size and location of the face is determined and the face image normalized in size. The second matching step is used to find the fiducial points for recognition. The two steps use different FBGs with different emphasis and number of nodes. The first step requires several FBGs of different size, the best fitting one of which is used for size estimation. Each image has a label which indicates the pose, so that pose does not need to be determined automatically, though our system is able to determine pose automatically in the same way as size is estimated [5]. The two steps together take less than 30 seconds on a SPARCstation 10-512. Figure 2 shows some automatically positioned grids.

2.4 Recognition

After having extracted model graphs from the gallery images and image graphs from the probe images, recognition is possible with relatively little computational effort by comparing an image graph to all model graphs and selecting the one with the highest similarity value. The similarity function we use here for comparing graphs is an average over the similarities between pairs of corresponding jets. Some jets in one pose may not have a corresponding jet in the other pose because of occlusions. We use the jet similarity function without phase here. It turned out to be more discriminative, possibly because it is more robust with respect to change in facial expression and other variations. Grid distortions are not taken into account. This graph similarity induces a ranking of the model graphs relative to an image graph. A person is recognized correctly if the correct model yields the highest graph similarity, i.e. if it is of rank one. A comparison against a gallery of 250 individuals took slightly less than a second.

3 Experiments

One set of tests was done on the ARPA/ARL FERET database provided by the US Army Research Laboratory. The poses used here are: neutral frontal view (fa), frontal view with different facial expression (fb), half-profile right (hr) or left (hl) (rotated by about 40-70°), and profile right (pr) or left (pl) (see Figure 2 for examples). The size of the faces varies by about a factor of three, which was compensated for by the first matching step. The format of the original images is 256×384 pixels, 256 grey levels. Recognition results are shown in Table 1.

The recognition rate is very high for frontal against frontal images (first row). This is mainly due to the fact that in this database two frontal views show only little variation, and any face recognition system should perform well under these circumstances. See results on the Bochum database for a more challenging test.

Model gallery	Probe images	First rank #	First rank %	First 10 #	First 10 %
250 fa	250 fb	245	98	248	99
250 hr	181 hl	103	57	147	81
250 pr	250 pl	210	84	236	94
249 fa + 1 fb	171 hl + 79 hr	44	18	111	44
171 hl + 79 hr	249 fa + 1 fb	42	17	95	38
170 hl + 80 hr	217 pl + 33 pr	22	9	67	27
217 pl + 33 pr	170 hl + 80 hr	31	12	80	32

Table 1: Recognition results for cross-runs between different galleries. The different compositions in the four bottom rows are due to the fact that not all poses were available for all people. The table shows how often the correct model was identified as rank one and how often it was among the first 10 (4%).

Before comparing left against right poses we flipped all left pose images over. Since human heads are bilaterally symmetric to some degree and since our present system performs poorly on such large rotations in depth (see below), we proceeded under the assumption that it would be easier to deal with differences due to facial asymmetry than with differences caused by substantial head rotation. This assumption is born out at least by the high recognition rate of 84% for right profile against left profile (third row). The sharply reduced recognition rate of 57% (second row) when comparing left and right half-profiles could be due to inherent facial asymmetry, but the more likely reason is the poor control in rotation angle in the database — visual inspection of images shows that right and left rotation angles may differ by up to 30°.

When comparing half profiles with either frontal views or full profiles another reduction in recognition rate is observed (although even a correct recognition rate of 10% out of a gallery of 250 is still high above chance level, which would be 0.4%!). The results are asymmetrical, performance being better when frontal or profile images serve as model gallery rather than if half-profiles are used. This is due to the fact that both frontal and profile poses are much more standardized than half-profiles, for which the angle varies between 40° and 70°. We interpret this as being due to the fact that similarity is more sensitive to depth-rotation than to inter-individual face differences. Thus, when comparing frontal probe images to a half-profile gallery, a 40° half-profile gallery image of a wrong person is often favored over the correct gallery image if, in the latter, the head is rotated by a larger angle. A large number of such false positives considerably degrades the correct-recognition rate. In these experiments we also flipped all left pose images over, so that to a large extent the recognition was not only done across pose but also across mirror reflection.

A second set of tests has been done on the Bochum database [1]. It contains neutral frontal views (fa), frontal views with different facial expression (fb), 11° rotated poses (refered to as 15° in [1] because the gaze is at 15°, but the head rotation is less), 22° rotated poses. For the Bochum database we did not use the normalization stage, because faces varied only little in size.

We used 108 neutral frontal views as a model gallery and the other images as probe galleries. The recognition rates for galleries fb, 11°, and 22° were 91%, 94%, and 88%, respectively. On the same galleries the preceding system [1] achieved 92%, 97%, and 85%. Thus the overall performance is the same. The performance on the fb-gallery is worse than for the corresponding fb-gallery of the FERET database, because the Bochum database shows more variation in facial expression, some faces being even half covered by a hand or hair.

We have introduced phase information in order to improve matching accuracy. We have tested the accuracy on the Bochum database by matching a face bunch graph to images for which all fiducial points were controlled manually. We always left the person on the image out of the face bunch graph, so that no information about that particular person could be used for matching. We ran the same algorithm with phase information and without phase information, i.e. all phases set to zero. Matching accuracy was calculated as the mean Euclidean distance between matching positions and manually controlled reference positions. It was 1.6 and 5.2 pixels with and without phase, and the histograms had their maximum at 1 and 4 pixels distance, respectively. The images had a size of 128×128 pixels.

4 Conclusion

The system presented is general and flexible. It is designed for an *in-class recognition* task, i.e. for recognizing members of a known class of objects. We have applied it to face recognition but the system is in no way specialized to faces and we assume that it can be directly applied to other in-class recognition tasks, such as recognizing individuals of a given animal species, given the same level of standardization of the images. In contrast to many neural network systems, no extensive training for new faces or new object classes is required. Only a moderate number of typical examples have to be inspected to build up a bunch graph, and individuals can then be recognized after storing a single image.

We tested the system with respect to rotation in depth and differences in facial expression. Some experiments included mirror reflection. We did not investigate robustness to other variations, such as illumination changes or structured background. The performance is high on faces of same pose. We also showed robustness against rotation in depth up to about 22°. For large rotation angles the performance degrades significantly. Our system perfoms well compared to other systems. Results of a blind test of different systems on the FERET database were published in [6, 7].

In comparison to the system [1] on the basis of which we have developed the system presented here we have made several major modifications. We now utilize wavelet phase information for accurate node localization. Previously, node localization was rather imprecise. We have introduced the potential to specialize the system to specific object types and to handle different poses with the help of object-adapted grids. The face bunch graph is able to represent a wide variety of faces, which allows matching on face images of unseen individuals. These improvements make it possible to extract an image graph from a new face image in one matching process. Even if the person of the new image is not included in the FBG, the image graph reliably refers to the fiducial points. This considerably accelerates recognition from large databases since for each probe image, correct node positions need to be searched only once instead of in each attempted match to a gallery image, as was previously necessary. We did not expect, and the system does not show, an improvement in terms of recognition rates compared to the preceding system.

The increased matching accuracy, the object adapted graphs, and the face bunch graph provide the basis for further improvements. In an extension of the system presented here, Krüger has developed a method for learning weights emphasizing those nodes which are more discriminative and more robust against noise [8]. On model galleries of size 130–150 and probe images of different pose, an average improvement of the first rank recognition rates of 6.5% has been achieved, from a mean performance of 19.8% without to 26.3% with weights.

Another individual treatment of the nodes has been developed by Maurer & von der Malsburg [9]. They applied linear jet transformations to compensate for the effect of rotation in depth. On a frontal pose gallery of 90 faces and half profile probe images an average improvement of the first rank recognition rate of 15% was achieved, from 36% without rotation to 50% and 53% with rotation, depending on which pose was rotated.

In [10] the bunch graph technique has been used to fairly reliably determine facial attributes from single images, such as sex or the presence of glasses or a beard. If this technique was developed to extract independent and stable personal attributes, such as age, race, or sex, recognition from large databases could be improved and considerably speeded by preselecting corresponding sectors of the database.

Future research on the basic system will have to focus on replacing the manual steps in the initial phase by automatic procedures. The manual selection of fiducial points could be replaced by grouping salient points on the basis of common motion [11]. Monitoring a rotating object by continuously applying elastic bunch graph matching can then reveal which nodes refer to corresponding fiducial points in different poses [12]. See [2] for a more detailed discussion.

Acknowledgements

We wish to thank Irving Biederman, Ladan Shams, Michael Lyons, and Thomas Maurer for very fruitful discussions and their help in the tests on the FERET database. Many thanks go to Thomas Maurer and Jan Vorbrüggen for additional tests on the Bochum database. For the experiments we have used the FERET database of facial images collected under the ARPA/ARL FERET program and the Bochum gallery collected at the Institute for Neural Computation, Ruhr-University Bochum. This work has been supported by grants from the German Federal Ministry for Science and Technology (413-5839-01 IN 101 B9) and from ARPA and the U.S. Army Research Lab (01/93/K-109).

References

[1] 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*, vol. 42, no. 3, pp. 300–311, 1993.

[2] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," Tech. Rep. IR-INI 96-08, Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum, Germany, 1996.

[3] J. G. Daugman, "Complete discrete 2-D Gabor transform by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1169–1179, July 1988.

[4] W. M. Theimer and H. A. Mallot, "Phase-based binocular vergence control and depth reconstruction using active vision," *CVGIP: Image Understanding*, vol. 60, pp. 343–358, Nov. 1994.

[5] N. Krüger, M. Pötzsch, and C. von der Malsburg, "Estimation of face position and pose with labeled graphs," in *Proceedings of the British Machine Vision Conference (BMVC96)*, pp. 735–743, 1996.

[6] P. J. Rauss, J. Phillips, M. K. Hamilton, and A. T. DePersia, "FERET (face-recognition technology) recognition algorithms," in *Proc. of the Fifth Automatic Target Recognizer System and Technology Symposium*, 1996.

[7] P. J. Phillips, P. J. Rauss, and S. Z. Der, "FERET (face recognition technology) recognition algorithm development and test report," Tech. Rep. ARL-TR-995, U. S. Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783-1197, Oct. 1996.

[8] N. Krüger, "An algorithm for the learning of weights in discrimination functions using *a priori* constraints." accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997.

[9] T. Maurer and C. von der Malsburg, "Linear feature transformations to recognize faces rotated in depth," in *Proceedings of the International Conference on Artificial Neural Networks, ICANN'95*, (Paris), pp. 353–358, Oct. 1995.

[10] L. Wiskott, "Phantom faces for face analysis," *Pattern Recognition*, vol. 30, no. 6, pp. 837–846, 1997.

[11] B. S. Manjunath, R. Chellappa, and C. von der Malsburg, "A feature based approach to face recognition," Tech. Rep. CAR-TR-604 or CS-TR-2834, Computer Vision Laboratory, University of Maryland, Colledge Park, MD 20742–3411, 1992.

[12] T. Maurer and C. von der Malsburg, "Tracking and learning graphs on image sequences of faces," in *Proceedings of the ICANN 1996* (C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, eds.), (Bochum), pp. 323–328, Springer Verlag, July 1996.