The Bochum/USC Face Recognition System and How it Fared in the FERET Phase III Test

Kazunori Okada[1], Johannes Steffens[1,2,3], Thomas Maurer[2], Hai Hong[1], Egor Elagin[1,3], Hartmut Neven[1,3], and Christoph von der Malsburg[1,2]

[1] Computer Science Department *and* Center for Neural Engineering
University of Southern California
Los Angeles, CA 90089-2520, USA
kazunori@selforg.usc.edu
[2] Institut für Neuroinformatik
Ruhr-Universität Bochum
D–44780 Bochum, Germany
[3] Now also at Eyematic Interfaces, Inc
827 20th Street, Santa Monica, CA 90403, USA

Summary. This paper summarizes the Bochum/USC face recognition system, our preparations for the FERET Phase III test, and test results as far as they have been made known to us. Our technology is based on Gabor wavelets and elastic bunch graph matching. We briefly discuss our technology in relation to biological and PCA based systems and indicate current activities in the lab and potential future applications.

1. Introduction

Vision is the most important of our senses by which we establish continuity between past and present. Vision is difficult for the simple fact that present scenes never repeat past examples in detail. Bridging that difference is the challenge. Vision has many aspects among which object recognition is but one. Object recognition requires the detection of similarity in spite of image variation in terms of translation, rotation scaling, pose (rotation in depth), deformation, illumination, occlusion, noise and background. Moreover, depending on the specific task, an object may have changing attributes, e.g., surface markings.

In principle, there are three types of information as a basis for generalization from past samples to present instances. One is the information in those samples themselves. It is of extreme biological importance to generalize from minimal sample bases. A second is the structural commonality of an individual object with others. This is a prominent aspect of face recognition, but plays an important role also for more variegated objects as far as they are composed of common shape primitives [Biederman, 1987]. A third type of information is based on first principles which can be built into a system and need not be derived from experience at all. An example of first principles are the transformation laws within the image plane — translation, rotation and scaling. In general, a vision system will exploit a mixture of all three information sources.

Face recognition is a rather particular example of object recognition in that all faces are qualitatively similar to each other and the distinctions to be made are of a gradual nature. For a discussion of face recognition in distinction to other object

recognition tasks see [Biederman and Kalocsai, 1997]. There is the common expectation that technical systems are potentially superior to human face recognition in being able to make precise metric measurements. Unfortunately, this is vitiated by even small variations in pose and facial expression.

We are dealing here with the problem of recognizing a person from a single photograph against a gallery of hundreds of persons, each represented again by a single photograph — the task set by the FERET program. The task as such is virtually impossible for humans to perform, due to the practical impossibility of memorizing (or repeatedly looking through) data bases of thousands of images as was required in the program's test. But even deciding whether two images presented in direct sequence do or do not refer to the same person is made difficult by variation in pose (or expression) [Kalocsai et al., 1994], [Biederman and Kalocsai, 1997]. The difficulty arises from the fact that a single photo doesn't contain enough information about a face's depth profile to predict images of different pose. The face recognition system we have developed is distinguished from others by a larger extent to which its generalization capabilities are based on general principles instead of on statistical learning. We will come back to this point at the end of the paper.

This report succinctly describes the basic system as developed previously [Lades et al., 1993], [Wiskott et al., 1997] and the particular improvements in preparation for the latest FERET test, as well as some details of our system implementation. We then discuss performance of our system resulting from in-house preparation tests and the FERET phase III test, which we have taken in March of 1997 and which has been partially reported [Phillips and Rauss, 1997]. We conclude by mentioning current activities in the lab and potential future applications of our technology, and by discussing our technology in relation to biological and PCA-based systems.

2. The System as Previously Developed

2.1 The Wavelet Transform

Previous versions of our system are described in [Lades et al., 1993],[Wiskott et al., 1997]. The basic data format of our system is the Gabor-based wavelet

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{k^2}{\sigma^2} e^{-\frac{k^2}{2\sigma^2}x^2} \left\{ e^{i\,\mathbf{k}\,\mathbf{x}} - e^{-\frac{\sigma^2}{2}} \right\}. \tag{2.1}$$

The wavelet is a plane wave with wave vector $\mathbf{k}$, restricted by a Gaussian window, the size of which relative to the wavelength is parameterized by σ. The second term in the brace removes the DC component. A wavelet, centered at image position $\mathbf{x}$, is used to extract the wavelet component $J_{\mathbf{k}}$ from the image with gray level distribution $I(\mathbf{x})$,

$$J_{\mathbf{k}}(\mathbf{x}) = \int d\mathbf{x}'\; I(\mathbf{x}')\; \psi_{\mathbf{k}}(\mathbf{x} - \mathbf{x}'). \tag{2.2}$$

We typically sample the space of wave vectors $\mathbf{k}$ in a discrete hierarchy of 5 resolution levels (differing by half-octaves) and 8 orientations at each resolution level, thus giving 40 complex values for each sampled image point (the real and imaginary components referring to the cosine and sine phases of the plane wave). We designate the samples in $\mathbf{k}$-space by the index $j = 1, \ldots, 40$ and consider all wavelet components centered in a single image point as a vector which we call a *jet*. A jet describes the local features of the area surrounding $\mathbf{x}$. If sampled with sufficient density, the image can be reconstructed from jets within the bandpass covered by the sampled frequencies.

2.2 Graphs and Their Similarity

To describe the aspect of an object (in this context, a face) we use a labeled graph, the nodes of which refer to points on the object's aspect and are labeled by jets. Edges of the graph are labeled with distance vectors between the nodes. To compare jets and graphs, similarity functions are defined. If two graphs are of equal geometry, their similarity is the simple sum of pair-wise jet similarities. If the graphs have relative distortion, a second term can be introduced [Lades et al., 1993] to take this into account. An important feature of our system is that we use two different jet similarity functions for two different and even complementary tasks. If the components of a jet J are written in the form $J_j = a_j\, e^{i\phi_j}$, with amplitude a_j and phase ϕ_j, one form for the similarity of two jets J and J' is the normalized scalar product of the amplitude vector

$$S(J, J') = \frac{\sum_j a_j a'_j}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}}. \tag{2.3}$$

The other similarity function has the form

$$S(J, J') = \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \mathbf{d}\mathbf{k}_j)}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}}. \tag{2.4}$$

This function contains the relative displacement vector $\mathbf{d}$ between the image points to which the two jets refer. When comparing two jets during graph matching, the similarity between them is maximized with respect to $\mathbf{d}$, leading to an accurate determination of jet position. This idea goes back to [Fleet and Jepson, 1990], [Theimer and Mallot, 1994]. We use it in the form developed by [Wiskott, 1995]. We are using both similarity functions, preferring the phase-insensitive version, eq. (2.3), which varies smoothly with relative position, when first matching a graph, and using the phase-sensitive version, eq. (2.4), when being interested in accurate positioning.

2.3 Elastic Graph Matching

The fundamental process with our system is elastic graph matching. In it, a *model graph* — a graph derived from a facial image with appropriate node positions — is compared to a test image. In the process, the nodes of the model graph are tentatively positioned over the image, jets are extracted from those image points and the similarity of the the thus-defined image graph to the model graph is determined. This similarity is optimized by varying node positions in the image. In an initial phase, this variation takes the form of a global move of a rigid copy of the model graph's node positions. In a second phase, image nodes are allowed to move individually, introducing elastic graph distortions. In order to *find* a decent match we use the phase-insensitive similarity function, eq. (2.3). With this similarity function, graphs and jets are attracted to their match points in the image over large distances by a smoothly ascending similarity gradient. When trying to *locate* a jet with great accuracy we use the phase sensitive similarity function, eq. (2.4), which by utilizing the phase is very sensitive to small jet displacements.

2.4 Elastic Bunch Graph Matching

When attempting to find an as yet unknown face in an image and to define a graph to represent it, we make use of a data structure called a *bunch graph* [Wiskott et al., 1995]. It is similar to the graph as described above, but instead of attaching only a single jet to each node, we attach a whole bunch of jets, each derived from a different facial image. To form a bunch graph, a collection of facial images (the *bunch graph gallery*) is marked with node locations at defined positions of the head. We call these positions *landmarks.* They are found by a semi-automatic process [Wiskott et al., 1997]. When matching a bunch graph to an image, the jet extracted from the image is compared to all jets in the corresponding bunch attached to the bunch graph and the best-matching one is selected. This process is called *elastic bunch graph matching.* Constructed with a judiciously selected gallery, a bunch graph covers a great variety of faces with different local properties.

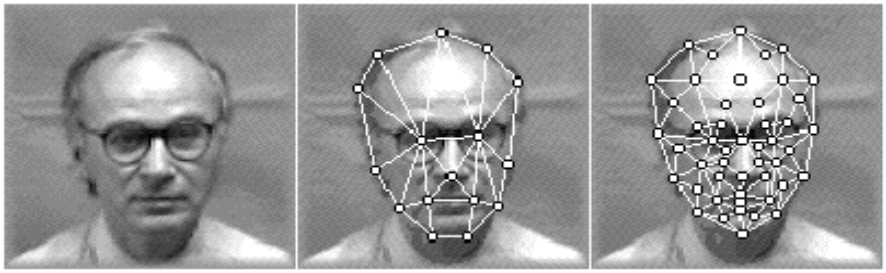
We accomplish recognition of the input face in three stages — face finding, landmark finding, and recognition by comparison. The first two stages serve to create a scale invariant model of the face in an input image. Both stages are based on elastic bunch graph matching, although with different parameter settings corresponding to different level of detail. Faces of different pose (in the FERET test the pose was identified in the file name) are processed in the same manner but with different bunch graphs customized for the relevant poses. In the last stage, face models are compared to achieve recognition. Each stage is discussed in detail below, in the order in which they are actually performed.

2.5 Face Finding

This first stage serves to find a face in an image and determine its size. This is accomplished by a set of matches to bunch graphs of appropriate pose and of three different sizes. The detailed schedule of this match is described in [Wiskott et al., 1997]. The best matching bunch graph determines the size and position of the face. We next place a square frame around the face so that the face occupies about a quarter of the area of the frame. The resulting image is warped to a standard size (currently 128×128 pixels), and a new wavelet transform is computed, thus defining the *image frame.* The image frame is passed to the next module, the Landmark Finder. See Fig. 2.1 B for the graph placed over the facial image during face finding. The reliability of this step in letting the face fall entirely inside the image frame is crucial to the success of the system.

2.6 Landmark Finding

Although in the face finding step a set of nodes was placed over the face, the basic procedure is now repeated with a bunch graph containing more nodes and a larger bunch graph gallery. The purpose of this step is to find facial landmarks with high positional accuracy and reliability and to encode the information contained in the image as accurately as possible. This step is equally crucial since a node not correctly placed over its landmark will lead to distorted similarity values during the comparison stage. Fig. 2.1 C shows a typical result of this stage. This model graph represents all the information extracted from an image. For a face in frontal pose it contains 48 nodes, compared to the 16 nodes used during face finding.



A B C

Fig. 2.1. Graph representation of a facial image. A: Input image. B: Face-finding graph. C: Model graph as defined for landmark-finding. The image frame determined by face finding is used in each case.

2.7 Graph Comparison

The model graphs produced as the result of the landmark finding step are compared pairwise to compute a similarity value. This value is computed as the sum of jet similarities between pairs of corresponding nodes divided by the number of pairs, using the phase-insensitive similarity function, eq. (2.3). Since model graphs for different poses differ in structure, a little conversion table was used to identify correspondence between nodes referring to the same landmark. The result of the graph comparison step is a complete comparison score, containing for each of the face entries in the gallery provided by ARL an ordered list of all other entries in descending order of similarity.

3. Algorithmic Improvements for FERET Phase III Test

A large part of our effort in preparation of the FERET Phase III test consisted in re-implementation of the previous system in the object-oriented program library FLAVOR (see section 3.4), as well as testing, debugging and parameter optimization.

It is a characteristic of the FERET data base that although most pairs of pictures of one person were shot in the same photo session (*same session* images), some were taken in different sessions (*duplicates*), sometimes more than a year apart. The same-session images contain a number of false cues, such as identical lighting and background, similar geometry (e.g., distance to camera) and camera settings, as well as identical clothing and hair style. We have made no effort whatsoever to exploit any of these cues. As only duplicates are relevant for practical applications of face recognition technology, we concentrated our system optimization effort on those. We also made efforts to achieve robustness with respect to at least small pose variations. With this motivation we have added three methods to our system. They are applied after the landmark finding process, section 2.6, in the following order.

3.1 Facial Histogram Equalization

In order to adjust for differences in lighting and in camera setting (which may lead to partial film saturation), we apply a technique called *histogram equalization.* In this technique, a gray value histogram is computed for an image, and depending on its shape a non-linear gray scale transfer function is computed and applied, to spread out intensity levels near histogram peaks and compress them near troughs. The particular version we apply is an adaptation of [Bates and McDonnell, 1986]. We apply histogram equalization after landmark finding. At that stage we define the smallest rectangular image segment containing the whole face as defined by the graph of landmarks. We compute the gray value histogram and the non-linear transfer function from this rectangle only (thus ignoring histogram distortions in the background) but apply the resulting equalization to the whole image frame. We then perform another Gabor wavelet transform to compute corrected jets for the model graph (actually this is the same transformation mentioned at the end of the next section).

3.2 Rescaling Gabor Filters

A coarse face size adjustment is implicit in our face finding procedure, section 2.5. The accuracy of this size determination is sufficient to define the image frame used for the landmark finding stage, but we observed that due to occasional misplacement of nodes the facial size in the image frame may still vary. This residual size variation is small enough not to compromise the reliability and accuracy of landmark finding, but it leads to distortion of the wavelet components extracted in the wavelet transform: a linear size scaling of the face translates directly into an inverse linear frequency scaling of the wavelets (only the product $\mathbf{kx}$ appears in the definition of the wavelet, eq. (2.1)).

We measure the exact facial size in the image frame by computing the mean Euclidean distance of all landmarks (nodes of the model graph) from their center of gravity, and comparing this number to the one derived from the standard graph used in the definition of the landmark finding bunch graph. The ratio of these two numbers, the *size adjustment factor,* is used to recompute wavelet components, with wavenumber $\mathbf{k}$ adjusted accordingly in eq. (2.1). In principle, scale-adjusted wavelet components can be computed by interpolation between neighboring frequency levels [Lades, 1995]. However, for the sake of higher accuracy we recomputed a wavelet transform from the image frame with the adjusted frequencies.

3.3 Jet Transformation for Face Rotation in Depth

The FERET test was to contain sub-tests with depth-rotated probe images. As all practical applications of face recognition technology will have to deal at least to some extent with pose variation, we made efforts to cope with this problem.

The FERET program insisted that a person be recognized on the basis of a single gallery image. There is no reliable method to compute the depth-profile of a face from a single image. Without exact knowledge of the three-dimensional shape of an object, the correspondence between images of different pose cannot be established accurately. This general impasse is mitigated by the fact that human faces share the general shape of their depth profile. Using an average facial depth profile it is therefore possible to predict a rotated pose to some degree of accuracy, which is perhaps a basis for improved recognition of the depth-rotated faces.

As far as a sparse set of point correspondences is concerned, this strategy is already implicit in our basic technology as described above, specifically in the average pose-specific graphs used in our bunch graphs, section 2.4, and the correspondences between nodes relating to the same landmark in different graphs. A more complicated story, however, is the adjustment of wavelet components, which contain the bulk of the information about facial identity. To some extent, Gabor-based wavelets are robust to the distortions implicit in small depth rotations of objects with generally smooth surfaces. This robustness has been the basis of our performance on depth rotation in previous FERET tests.

Here is the simple idea on which we base our approach for jet transformation [Maurer and von der Malsburg, 1995]. Assume we were dealing with a totally flat surface in three-dimensional space with some gray level distribution painted onto it, and a jet encoding this distribution around a given point. It would then be possible to accurately predict the transformation of the jet components due to depth rotation of the surface (assuming an isotropic radiance profile). Assuming the jet to be taken at the origin $\mathbf{x} = \mathbf{0}$ of the coordinate system and assuming the rotation to be about this point, the projection to the image plane of other points $\mathbf{x} = (x_1, x_2)$ is transformed according to $\mathbf{x}' = \mathbf{A}\ \mathbf{x}$, with $\mathbf{A}$ determined by the rotation angles in the image plane and in depth. This translates into the jet transformation

$$J'_{\mathbf{k}} = \int d\mathbf{x}\ I(\mathbf{A}^{-1}\mathbf{x})\ \psi_{\mathbf{k}}(\mathbf{x})\ = \int d\mathbf{x}\ I(\mathbf{x})\ \psi_{\mathbf{k}}(\mathbf{A}\,\mathbf{x})\ \det(\mathbf{A}). \tag{3.1}$$

As we want to stick to our sampling grid in $\mathbf{k}$ space, we make the Ansatz

$$\psi_{\mathbf{k}}(\mathbf{A}\,\mathbf{x})\ \det(\mathbf{A}) \approx \sum_{\mathbf{k}'} c_{\mathbf{k}\mathbf{k}'}(\mathbf{A})\ \psi_{\mathbf{k}'}(\mathbf{x}), \tag{3.2}$$

although this can only be an approximation, with an accuracy that increases with sampling density in $\mathbf{k}$-space. Multiplication of eq. (3.2) with $\psi_{\mathbf{k}''}(\mathbf{x})$ and integration leads to a system of linear equations to determine the $c_{\mathbf{k}\mathbf{k}'}(\mathbf{A})$. All integrals can be solved analytically; details can be found in [Maurer and von der Malsburg, 1995]. Once the $c_{\mathbf{k}\mathbf{k}'}(\mathbf{A})$ are determined, the jet can be transformed according to

$$\mathbf{J}' = \mathbf{C}(\mathbf{A})\ \mathbf{J}. \tag{3.3}$$

We can thus compute the transformation matrix $\mathbf{C}$ to transform the jet — our local visual feature vector — from one perspective into the other, given only the normal vectors of the surface before and after the rotation relative to the camera.

To apply this bit of theory to face recognition we have to assume that the surface around node points is flat to some degree of accuracy, and we have to determine the orientation of this area, i.e., its normal vector, for the two poses being compared. Once this normal vector and the geometric transformation of the face (the rotation angle) are known, the jet at this node can be transformed analytically.

As facial normal vectors are not available to us directly we have adaptively determined estimates from training galleries of 50–80 persons for each pair of poses to be compared. On these faces the flexible grids are placed automatically as described in section 2.6. For a given pair of corresponding nodes we create trial values for the normal angles, transform the jets for all persons from the rotated pose to the frontal pose and compare the two sets in terms of the recognition performance measure

$$E = \frac{1}{N^2} \sum_i \sum_j (s_{ii} - s_{ij}), \tag{3.4}$$

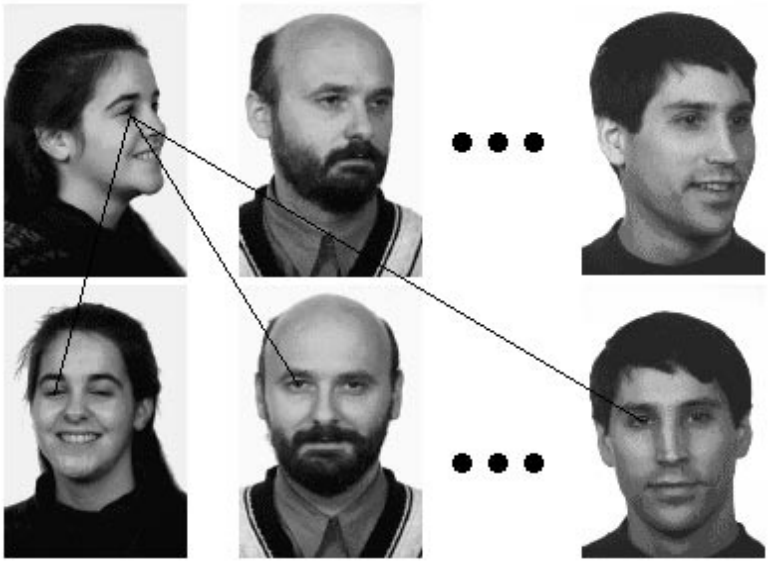


Fig. 3.1. Typical half profile and frontal faces. To learn the effective normal angles for the left eye, the left eyes of all half poses are compared with the left eyes of all frontal poses. Note the low degree of uniformity of poses labeled "half profile" in the FERET data base.

where the s_{ii} are jet similarities for the same person and s_{ij} for different persons. Stepping the trial values for normal angles through all possible orientations in steps of 5° horizontally and vertically, we optimize E. The procedure is repeated for all nodes of the flexible grids visible in both views, nodes having no correspondent in the other view being ignored (see Fig. 3.1). By this procedure we get an average set of effective normal vectors which determine — together with the head rotation angle — the transformation of the jets between the two poses. Let us remark here that we have only two free parameters per node (the two normal angles) for 50–80 data points (the jets of all persons at this node). As a consequence, there will be no generalization problem, which is confirmed by our tests.

After learning the transformation on a training gallery of quarter rotated faces, we obtained the following error rates on test images from the FERET training data base against a frontal gallery of 596 different persons, with a recognition attempt being counted as an error if it failed to identify the correct person as the best match:

Averaged error rates in %

Jet transformation	No	Yes
Quarter–Frontal (138 probes)	22.6	13.8
R–Frontal (237 probes)	13.5	9.2

Jet transformation thus almost halves the error rates on the quarter rotated faces (22.5° rotation angle). With the same set of normal angles it reduced error rates by a third on the so-called R faces (about half-way between quarter and frontal), proving the method to be robust against pose measurement inaccuracies. Further results are reported in section 4.1.

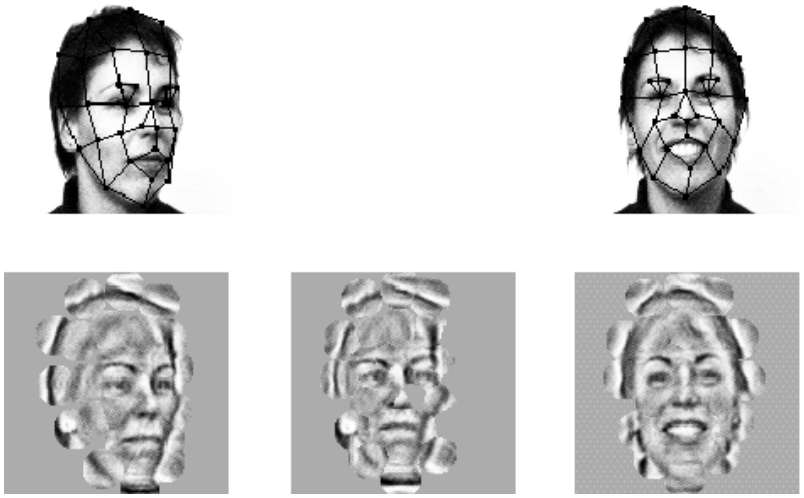


Fig. 3.2. Visualization of the pose transformation on one example. The original images together with the automatically placed grids are shown in the top row, and in the bottom row left and right the respective reconstructions. In the middle of the bottom row there is the reconstruction of the transformed face graph, which has to be compared with the right one for recognition. Only jets at nodes visible in both views are compared.

By reconstructing images from jets (see, e.g., [Pötzsch et al., 1996]), this pose transformation can be visualized, see Fig. 3.2.

One could argue that the assumption of area flatness is grossly violated in certain facial locations, such as near the eyes or the tip of the nose. However, even if that is the case the system tries, during the optimization of E, to find effective normal vectors to describe the transformation as accurately as possible.

3.4 Implementation Issues

We describe here some of the technical issues we faced for the latest FERET test, which we took in March 1997. For this test, we re-implemented the application on the basis of *FLAVOR,*[1] an extensive C++ class library of image processing algorithms designed and written by Christoph von der Malsburg's groups at the Institut für Neuroinformatik, Ruhr-Universität Bochum and at the Laboratory of Computational and Biological Vision, University of Southern California [Rinne et al., 1997]. FLAVOR provided us with all the core algorithms and a lot of support functions, e.g., for the display of results, and a user interface. We integrated the new features described above into FLAVOR and built an application program suited to the needs of the FERET test. FLAVOR thus gave us a stable starting point and a good environment for rapid prototyping and development.

We then optimized the data structures and memory allocation of our system's code in order to reduce the application's computation time. After these optimizations, the computation time to create a gallery entry was approximately one minute

[1] *Flexible Library for Active Vision and Object Recognition*

on a workstation with a 60 MHz SuperSPARC processor; the computation time for a recognition run was approximately ten seconds when the probe image's pose was frontal and twenty seconds when rotation in depth had to be compensated for, using the same processor with a gallery of 3816 entries.

The computation time is crucial, both for comparing our system to human performance and for fielded applications, where recognition times on the order of a few seconds are required. The computation time of our system was somewhat slow for this, though we have since improved on this: We have implemented an on-line system based on the same technology which tracks faces in real time, using a more powerful platform. This system can process 8 persons per minute without compromising much of the recognition performance [Steffens et al., 1997].

Memory requirements are another practical issue when trying to apply a system to real-world conditions. The size of a test image (256×384 pixels) in the FERET database is 96 KBytes; the size of a model graph containing 48 nodes and jets with 40 components is approximately 16 KBytes, when stored in binary. This could be considered as more than 80 % of data compression for a face model, although there are other methods which can achieve better compression factors [Turk and Pentland, 1991]. One possible way for a representation based on jets to achieve higher compression factors is to cluster jets or jet coefficients using their regularities. Krüger *et al.* proposed a clustering algorithm for reducing the number of jets in a bunch graph [Krüger et al., 1997]. Lately, Kalocsai *et al.* showed that it is possible to eliminate certain filters without sacrificing discriminative power of the representation [Kalocsai et al., 1997]. The jet and filter clustering is important not only for data compression, but also for reducing computation time.

3.5 Test Procedure

The test was administered such that, first, all 3816 images supplied by the test conductor were used to generate model graphs for later recognition purposes. This process of gallery creation consumed most of the processing time (see below). Then, the same model graphs were used as a probe set, i.e., run through the recognition process and compared against the whole gallery. The resulting matrix of similarity values is the major test result. It was then analyzed by the test conductor by restricting both gallery and probe sets to appropriate subsets, which generated the performance measurements reported in the following section (see [Phillips and Rauss, 1997] for details).

Our system was tested in two conditions — with and without coordinates of the eyes in the images supplied as additional information. Thus, for this test we actually performed two sets of gallery creation (one using the eye coordinate information, the other ignoring it) and two sets of 3816 recognition runs against the full gallery. The processing time for the entire test procedure was approximately 26 hours (22 hours for gallery creation and 4 hours for the recognition runs), for which we employed six workstations (two with 150 MHz microSPARC II and four with 60 MHz SuperSPARC processors) running in parallel.

4. Test Results

The results of the latest FERET test are described in this section. The results of in-house pre-tests for confirming the system improvement are described, followed by an evaluation of the FERET test results reported by Phillips and Rauss [Phillips and Rauss, 1997].

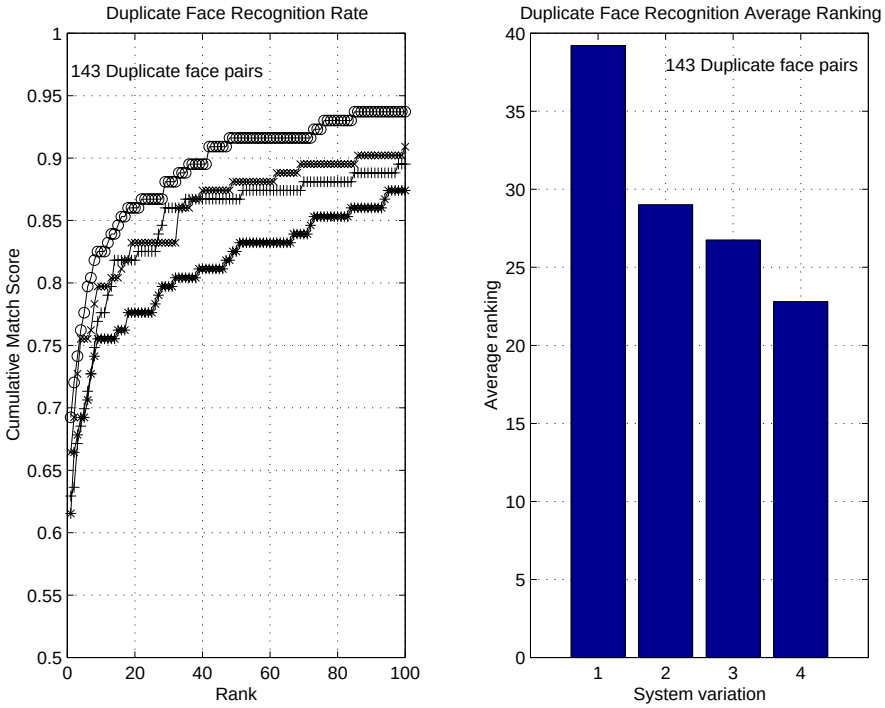


Fig. 4.1. Recognition results of FA vs FB tests in the duplicate face recognition task for various system settings. *Duplicate Face Recognition Rate* (left) shows percentage of successful duplicate face recognition. *Duplicate Face Average Ranking* (right) is an average over recognition ranks for each probe. Rank 1 corresponds to a correct recognition and a lower value means better performance. Legend: $*$ and 1: without new functions; $+$ and 2: histogram equalization only; $\times$ and 3: Gabor filter rescaling only; $\circ$ and 4: both new functions.

4.1 Pre-test Results

To evaluate our system improvements described in section 3., we performed some in-house pre-tests with the training data consisting of 526 sets of images provided by ARL [Phillips et al., 1996]. Each set comprised pictures of the same person taken in different photo sessions. Each set from one session contained two frontal face images marked FA and FB and an optional number of the rotated face images. For some persons, images from multiple photo sessions exist which were, in some cases, taken more than a year apart (duplicate sets). A pair of images from the same set contained either slight variations in facial expression (between FA and FB) or variations in depth rotation (between FA and rotated faces). On the other hand, a pair of images from different sets but belonging to the same person (duplicates) contained greater variations of facial appearance due to changes of illumination, background, facial expression, hair style and face size on top of the variations between FA and FB or FA and rotated faces. See [Phillips et al., 1996] for a detailed description of the data set.

Our system performed very well on the same session recognition task: the recognition rate (RR) of FB probe set (526 entries) was 96.6% and RR of QR probe set

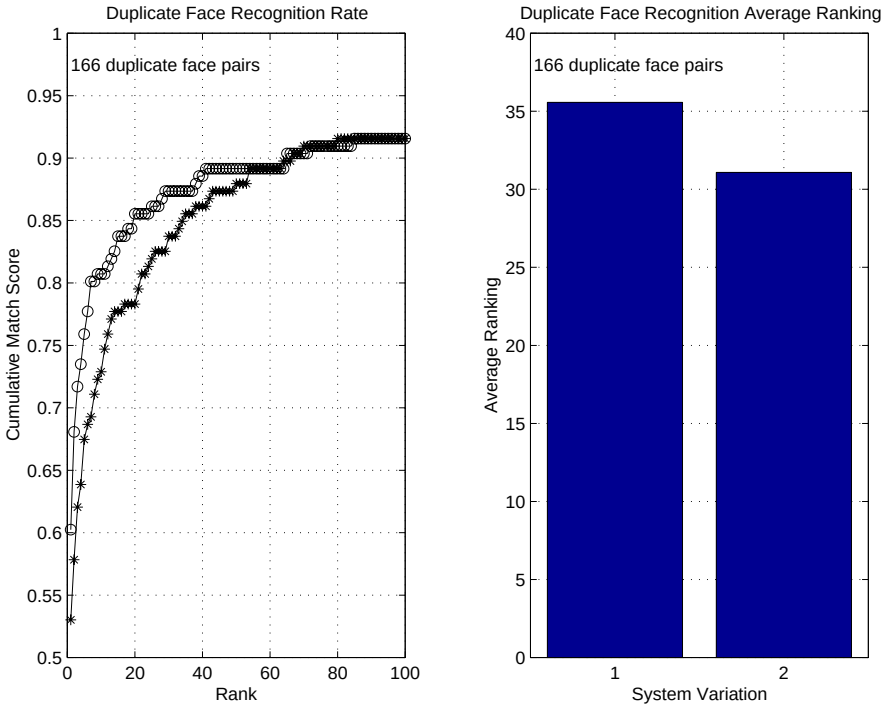


Fig. 4.2. Recognition results of the FA vs QR test in the duplicate face recognition task. Legend: ∘ and 2: with jet transformation; ∗ and 1: without jet transformation. The jet transformation pushes the recognition rates on the rotated duplicates close to those of the frontal ones (see Fig. 4.1); thus, rotation in depth does not seem to be the major problem on these images.

(374 entries) was 87.2%. It turned out to be difficult to evaluate improvement of our system by the same session recognition task due to the already high recognition performance. We therefore concentrated our efforts on the duplicate face recognition task. There were 143 duplicates in the FA vs FB test and 166 duplicates in the FA vs QR test.

We evaluated our system in two conditions: 1) recognition tests of a probe set with an FA gallery, in which a probe was recognized correctly if the best match in the gallery was from the same session of the probe and 2) recognition tests of the same probe set with the same gallery but correct recognition of a probe was granted if the second best match was a duplicate of the probe. Note that the latter condition is harder than the former because of the greater variations of duplicates. We will refer to the former as the same session recognition task and to the latter as the duplicate face recognition task. For both conditions, we performed two tests, one with the FB probe set and the other with a quarter rotated face (QR) probe set. The size of probe sets varied depending on the type of test and condition but the size of the gallery was always fixed at 526 entries. Eye coordinate information was not used for any of the tests.

We evaluated the performance of histogram equalization and Gabor filter rescaling by the FA vs FB test in the duplicate face recognition task. Recognition results

with and without the new functions are shown in Fig. 4.1. For testing the performance of the jet transformation to compensate for rotation in depth, the FA vs QR test in the duplicate face recognition task was used. Recognition results with and without jet transformation are compared in Fig. 4.2. Both facial histogram equalization and Gabor filter rescaling were used as preprocesses for the FA vs QR tests.

Fig. 4.1 and Fig. 4.2 indicate a significant drop in recognition performance for the duplicate face recognition task when compared to the same session recognition task. This clearly shows that the great amount of face and image variations in duplicates indicates a need for improving the system further [Phillips et al., 1997]. By applying histogram equalization and Gabor filter rescaling, the recognition performance of the FA vs FB test was improved as shown by a 12% increase in recognition rate. The result also shows that improvement included not only the recognition rate but also ranks of failure cases so that average rank of all probes was significantly improved (42% decrease in average ranking). For the FA vs QR test, similar improvements were observed by correcting for face rotation in depth through application of the jet transformation described in section 3.3 (14% increase in recognition rate and 13% decrease in average ranking; see Fig. 4.2). These results suggest that our system has considerably improved, especially in the case when facial and image properties vary greatly, such as is the case with the duplicate images.

4.2 The FERET Test Result

Phillips and Rauss [Phillips and Rauss, 1997] reported results of the FERET phase III test. This test was administered in September 1996 and March 1997. There were ten groups of participants. Their report included a performance analysis of ten systems for the same session recognition and the duplicate face recognition tasks, which were explained in the previous section; however, only tests which made use of the eye coordinate information is reported there.

Our pre-test results, presented in section 4.1 and produced with a much smaller data set, were well replicated qualitatively in their report [Phillips and Rauss, 1997]. In the FA vs FB test of the same session recognition task, three systems, including ours, showed similarly high performance (approximately 95% RR). In the FA vs FB test of the duplicate face recognition task, two systems, including ours, outperformed others (approximately 60% RR), when 1196 entries in the gallery and 722 entries in the probe set were used. It is notable that on the subset of duplicates whose images were taken more than one year apart, our system was significantly better than all others: With 864 entries in the gallery and 234 entries in the probe set, our system achieved a recognition rate of approximately 52%, while the second best system scored approximately 35%.

Note that results without the use of eye coordinates were not included in Phillips and Rauss' report. This type of test was taken by only two participants, including our group. When eye coordinates are not given, a system has to find facial landmarks automatically and this adds another factor that potentially degrades recognition performance. Finding facial landmarks is a sub-category of the face recognition problem that is known to be difficult in general. Table 4.1 summarizes our results of the FERET phase III test in detail. Data in the table referring to the test with eye coordinate information correspond to the data reported by Phillips and Rauss.

In Table 4.1, we can see similar performance with or without the use of eye coordinate information; the degradation of recognition rate when eye coordinates were unavailable was at most 3%. These results can be explained by our reliable process of facial landmark finding. The FC images were introduced to the FERET tests for phase III. They are frontal images using modified illumination (taken with

only natural light in a photo studio by turning off the studio lighting). Our system performed very well on this task for both categories. Our system's low recognition performance on the Quarter 2 task can be explained by the additive effects from two types of variations, depth rotation and duplicates. Some inaccuracies of pose labels which we found in our training data set (see Fig. 3.1) might also have contributed to the decrease of recognition performance for rotated face images.

The low recognition rate in the report of Phillips and Rauss, produced with the most realistic duplicate data set (Dup 3), showed that current technology is not close to solving the face recognition problem completely. Although the performance of any system using a limited data set cannot be fully translated to more realistic situations, their results highlight the quality of our technology. Our face recognition system seems to be closest to reaching the level of robust recognition of faces with realistic variations. One of our future work directions is to enhance the performance of our system in this domain.

5. Technical Applications

The technology we describe here has a wide range of applications, some of which have already been realized. In the field of security, face recognition technology can be used for access control to high-security areas, as realized in the commercial system ZN-Face [Konen and Schulze-Krüger, 1995], and the application can be easily extended to picture I.D. verification. In the area of criminal investigation, big galleries of potential suspects can be reduced, with the help of a sample image or of a composite [Konen, 1996], to a sample of 50 or so images. A witness can be expected to examine these preselected samples before losing concentration or recollection.

Table 4.1. Results for different recognition tasks. For a number of different recognition tasks, the recognition ratio (in %) for three different ranks is shown, for both the case with and without eye coordinates being supplied to the system. The sizes of the gallery and probe sets are also given. Definition of tasks following the FERET program's reports [Phillips et al., 1996], [Phillips and Rauss, 1997]: FA vs FB: same session recognition task; FA vs FC: illumination variations (see text) in FC set; Dup 1: probe set contains all duplicate images available; Dup 2: probe set contains duplicate images where difference is whether eye glasses are worn; Dup 3: probe set contains duplicate images taken at least one year apart; Quarter 1: quarter rotated images compared to FA images; Quarter 2: quarter rotated images compared to duplicate images; RB vs RC: images compared are with subject's head rotated nominally 12° to left and right, respectively.

	Probe/	With Eye Coordinates			Without Eye Coordinates		
Test	Gallery	Rank 1	10	20	Rank 1	10	20
FA vs FB	1195/1196	95	98	99	94	97	97
FA vs FC	194/1196	82	92	95	80	89	92
Dup 1	722/1196	62	72	79	61	71	79
Dup 2	176/1196	93	97	98	94	95	95
Dup 3	234/ 864	52	71	76	52	68	73
Quarter 1	32/1196	85	91	91	78	87	87
Quarter 2	126/1196	33	60	66	29	60	64
RB vs RC	94/1196	52	67	73	50	65	69

In distinction to security access control, in many situations the person to be identified cannot be expected to cooperate in the generation of a fiducial photo. Much wider application areas, such as automatic monitoring, or automatic passenger tracking and identification at airports, are opened by the PersonSpotter system we have recently developed [Steffens et al., 1997]. It is able to capture, track and recognize a person walking by a camera in real time.

On-line facial expression recognition [Hong et al., 1997] opens vistas on better human-machine communication, for instance for video games, tele-conferencing and computer-based training systems. Recognition of facial and hand gestures [Triesch and von der Malsburg, 1996] can be used to control machines more conveniently.

Fully immersive tele-conferencing requires the creation of a display that renders remote participants with correct direction of gaze independent of their spatial relation to the camera, and the creation of a realistic three-dimensional sound field. For both tasks, heads and facial features of participants (ears in the case of immersive sound) have to be accurately tracked [Maurer and von der Malsburg, 1996].

Video annotation would be an important multimedia application of our technology. Thus, by recognizing faces, facial expression and gestures, including head pose [Elagin et al., 1997], one could identify, characterize and extract human activities from video sequences on the basis of abstract descriptions.

6. Discussion

6.1 Sources of Structural Information

Perhaps the most striking feature of the visual system of higher vertebrates is its great generality and flexibility in recognizing objects and situations. This is in stark contrast to the high degree of specialization of most technical vision systems. One of our motivations for working on face recognition and in the FERET program was to expose ourselves to the requirements of a real-world vision application while working with a minimum of domain-specific structure.

In this connection we would like to return to an issue we raised in the introduction, referring to the type of information sources tapped during the storage and recognition processes. We had classified possible sources into a) individual sample data, b) statistical samples of images of the same object type ("same type" to be taken in a narrower or wider sense), or c) information in the form of first principles not instructed by sample data (at least not by object-specific samples). A previously described system [Lades et al., 1993], which might be considered a forerunner of the system we describe here, had been based entirely on individual data samples extracted from single images and on first principles. Those first principles concerned the form of visual features (Gabor-based wavelets), the data structure of labeled graphs, and a matching procedure permitting in-plane transformations (translation and deformation [Lades et al., 1993] as well as scaling and rotation [Lades, 1995]). The resulting system recognized faces by comparing stored individual samples to full test images. The system was entirely general and could recognize arbitrary objects, not just faces, if only the stored aspect was not too different from the one presented in the test image.

Specific constraints set by the FERET test forced us to build face-specific information into our system. One of these constraints was the great number, several millions, of image comparisons to be performed within a limited time period. This precluded us from comparing image pairs by elastic graph matching with its expensive examination of multiple image jets per stored jet. Instead, we extract from each

image a fixed, parsimonious set of jets, centered on landmarks, and compare each jet to its one correspondent in the other image, see section 2.. Finding landmarks is only possible on the basis of information about faces in general. In our system this "general face knowledge" is constituted by the bunch graph — a small gallery of sample portraits together with semi-automatically identified landmark positions.

Another constraint set by the FERET test was the requirement to be robust with respect to differences in pose. As discussed in section 3.3, this also forced on us face-related system knowledge, in the form of pose-specific graph structures and their correspondences, and in the form of a general mechanism for jet transformation, the parameters of the latter being trained on a set of sample faces.

Although we have to admit that our system presently still relies in several places on hand-constructed scaffolding, such as the manual selection of landmarks in the bunch graph entries, it is perhaps not too far removed from one to be based entirely on first principles plus exposure to samples. This goal is not just of academic value, as vision technology will come into its own only if applications can be trained instead of being manually constructed.

6.2 Comparison to the Vertebrate Visual System

Our extensive usage of Gabor-type wavelets is the most obvious point of similarity to the visual system of higher vertebrates [Jones and Palmer, 1987]. It was Daugman [Daugman, 1988] who first pushed the idea that receptive fields in the primary visual cortical areas are most appropriately described as two-dimensional Gabor functions. It is tempting to liken simple cells [Hubel and Wiesel, 1962] to the sine and cosine components of wavelets (or intermediate phases) and to connect complex cells [Hubel and Wiesel, 1962] with Gabor magnitudes. The great utility of Gabor magnitudes, as experimentally observed in certain stages of the matching process, suggests an evolutionary mechanism for the occurrence of complex cells in our visual system. We have two mutually non-exclusive explanations for this importance of disregarding phase information during the matching process. One is that the inclusion of phases creates the danger of being caught in one of the many local optima when matching a jet to an image. The other is the surmise that perhaps the phase relationships between wavelets of different frequency in a given object location are subject to much stronger variations than amplitude relations.

The great robustness of our face recognition system with respect to image variations is to a very large extent due to properties of the wavelets we employ. Some aspects of this may be easy to understand, such as the gradual response of wavelet responses to image deformation. Another important aspect may be that many common image variations are local both in the spatial and in the frequency domain, thus affecting only some wavelet components and leaving the others intact [Biederman and Kalocsai, 1997]. Other aspects are still obscure, such as the surprisingly small effect of lighting differences on wavelet components.

Another very encouraging aspect of our system is the close correspondence of its behavior under image variation to changes in psychophysical responses [Biederman and Kalocsai, 1997], [Hancock et al., 1997]. The error rates and response times of human subjects degrade under variation of facial expression or pose very much in parallel to the degradation of image-model similarities in the earlier version [Lades et al., 1993] of our system, giving correlation coefficients of 0.90 and higher [Biederman and Kalocsai, 1997].

6.3 Comparison to the Principal Components Approach

The FERET program aims at stimulating the development of alternative face recognition technologies and testing them competitively. Our strongest competitors have based their systems on Principal Component Analysis (PCA) techniques. It is therefore important to compare our system to PCA. In its basic form, the eigenface approach [Turk and Pentland, 1991] considers the pixel array within a rectangle around the face as a vector. Such vectors are collected from a large sample of facial images, all carefully aligned relative to each other. The correlation matrix is formed for this collection of vectors and its eigenvectors, or principal components, are extracted. The original face images can be linearly combined, using the components of the eigenvectors as coefficients, to form *eigenfaces.* A small number of eigenfaces corresponding to the largest eigenvalues are used as feature vectors. These vectors conveniently help to find, encode and compare faces.

In this simple form, the PCA approach has the great advantage over our method of representing faces much more parsimoniously and requiring much less computation. On the other hand, PCA requires the collection of a large number of images before features can be defined. (This is not a problem with the specific job of face recognition, but may turn out to be a severe restriction of flexibility when attempting more general object recognition tasks.) Moreover, the sample images have to be aligned very carefully, imprecision in alignment amounting to a corresponding reduction in effective resolution.

The eigenface approach becomes problematic when facial deformation (due to alteration in expression or pose) becomes important. Two ways have been proposed to deal with this. In one, several smaller windows are defined over landmarks of the face and are independently subjected to PCA in addition to the full face [Pentland et al., 1993], [Penev and Atick, 1996]. During recognition, landmarks are found and compared independently using the local PCAs, perhaps constraining relative positions of landmarks appropriately. In the other approach [Craw and Cameron, 1991], [Vetter and Troje, 1995], [Lanitis et al., 1995] faces are first morphed to an average shape prior to running PCA. PCA may also be performed separately on the shape vectors used during morphing. The essential features of both versions, elastic deformation and local feature vectors attached at landmark nodes, make the PCA approach more similar to ours and suggest perhaps a point of convergence for both methods. A major difference remains, however, the difference in origin of (local) features, object-specific statistical samples in the case of PCA and object-independent wavelets in ours.

Unfortunately we have insufficient information on our competitor's systems to be able to attribute their strengths and weaknesses in the FERET test to specific functional aspects. However, Hancock et al. [Hancock et al., 1997] have compared our system and PCA-based systems with performance of human subjects and concluded: "Comparisons between the systems' performance with faces with and without the hair visible, and prediction of memory performance with and without alteration in face expressions, suggested that the graph-matching system was better at capturing aspects of the appearance of the *face,* while the PCA-based system seemed better at capturing aspects of the appearance of specific *images* of faces" (emphasis original).

In our opinion, the better performance of our system on subtasks of the FERET test with strong image alterations (duplicates, illumination, glasses, pose) and when compared to psychophysical data is to a large part due to properties of Gabor wavelets, which form a well regularized mapping from pixel space to feature space when it comes to slight misalignment and to variation in scale and in illumination. Their structure is determined by principles and not by the particular properties of

statistical samples, and thus they are much less sensitive than the PCA approach to unexpected alterations of images.

Acknowledgement. This work was supported by the Army Research Laboratory, contracts DAAL01-93-K-0109 and DAAL01-96-K-0035. We thank Norbert Krüger, Michael Pötzsch, Michael Rinne and Jan Wieghardt for extensive help with the preparation of the FERET Phase III test, and Jan Vorbrüggen for advice and editorial work on the manuscript. For the experiments we have used the FERET database of facial images collected under the ARPA/ARL FERET program.

References

Bates, R. H. T. and McDonnell, N. J. (1986): *Image Restoration and Reconstruction.* Oxford Unversity Press, Oxford, p. 226–227.

Kalocsai, P., Biederman, I., and Cooper, E. E. (1994): *To What Extent Can the Recognition of Unfamiliar Faces be Accounted for by a Representation of the Direct Output of Simple Cell.* In: Proceedings ARVO, Sarasota, Florida, p.1627.

Biederman, I. (1987): *Recognition-by-Components: A Theory of Human Image Understanding.* Psychological Review 94:115–147.

Biederman, I. and Kalocsai, P. (1997): *Neurocomputational Bases of Object and Face Recognition.* Phil. Trans. Roy Soc.: Biological Sciences, in press.

Craw, I. and Cameron, P. (1991): *Parameterising images for recognition and reconstruction.* Proc. British Machine Vision Conference, Turing Institute Press and Springer Verlag.

Daugman, J. D. (1988): *Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression.* IEEE Trans. on Acoustics, Speech and Signal Processing 36:1169–1179.

Elagin, E., Steffens, J., and Neven, H. (1997): *Automatic Real-Time Pose Estimation System for Human Faces Based on Bunch Graph Matching Technology.* Proc. Intl. Conf. on Automatic Face- and Gesture- Recognition, Nara, Japan (submitted).

Fleet, D. J. and Jepson, A. D. (1990): *Computation of Component Image Velocity from Local Phase Information.* Intl. J. of Computer Vision, 5(1):77–104.

Hancock, P. J. B., Bruce, V, and Burton, A. M. (1997): *A Comparison of Two Computer-Based Face Identification Systems with Human Perceptions of Faces.* Vision Research, in press.

Hong, H., Neven, H., and von der Malsburg, C. (1997): *Online Facial Expression Recognition based on Personalized Gallery.* Proc. Intl. Conf. on Automatic Face- and Gesture- Recognition, Nara, Japan (submitted).

Hubel, D. H. and Wiesel, T. N. (1962): *Receptive Fields, Binocular and Functional Architecture in the Cat's Visual Cortex.* Journal of Physiology, 106–154.

Jones, J. and Palmer, L. (1987): *An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex.* Journal of Neurophysiology, 1233–1258.

Kalocsai, P., Neven, H., Steffens, J., and Biederman, I. (1997): *Statistical Analysis of Gabor-filter Representation.* Proc. Intl. Conf. on Automatic Face- and Gesture-Recognition, Nara, Japan (submitted).

Konen, W. (1996): *Comparing Facial Line Drawings with Gray-Level Images: A Case Study on PHANTOMAS.* Proc. ICANN'96, Springer-Verlag, Heidelberg, New York, 727–734.

Konen, W. and Schulze-Krüger, E. (1995): *ZN-Face: A System for Access Control Using Automated Face Recognition.* Proc. Intl. Workshop on Automatic Face- and Gesture- Recognition, Zürich, 18–23.

Krüger, N., Pötzsch, M., and von der Malsburg, C. (1997): *Determination of Face Position and Pose with a Learned Representation Based on Labeled Graphs.* Image and Vision Computing, 665–673.

Krüger, N. (1997): *An Algorithm for the Learning of Weights in Discrimination Funcions Using a priori Constraints.* IEEE Trans. on Pattern Recognition and Machine Intelligence, 19:764–768.

Lades, M. (1995): *Invariant Object Recognition Based on Dynamical Links, Robust to Scaling, Rotation and Variation of Illumination.* PhD. Thesis, Ruhr-Universität, Bochum, Germany.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W. (1993): *Distortion Invariant Object Recognition in the Dynamic Link Architecture.* IEEE Transactions on Computers, 42(3):300–311.

Lanitis, A., Taylor, C. J., Cootes, T. F. (1995): *A Unified Approach To Coding and Interpreting Face Images.* Proc. Intl. Conference on Computer Vision, Cambridge, 368–373.

Maurer, T., and von der Malsburg, C. (1995): *Single-View Based Recognition of Faces Rotated in Depth.* Proc. Intl. Workshop on Automatic Face- and Gesture-Recognition, Zürich, 248–253.

Maurer, T., and von der Malsburg, C. (1996): *Tracking and Learning Graphs and Pose on Image Sequences.* Proc. Intl. Workshop on Automatic Face- and Gesture- Recognition, Vermont, 176–181.

Penev, J. S., and Atick, J. J. (1996): *Local feature analysis: a general statistical theory for object representation.* Network: computation in neural systems, 7:477–500.

Pentland, A., Moghaddam, B., and Starner, T. (1993): *View-Based and Modular Eigenspaces for Face Recognition.* Technical Report 245, MIT Media Lab Vismod.

Phillips, P. J., Rauss, P., and Der, S. Z. (1996): *FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results.* US Army Research Laboratory Technical Report ARL–TR–995.

Phillips, P. J., Moon, H., Rauss, P., and Rizvi, S. A. (1997): *The FERET September 1996 Database and Evaluation Procedure.* Proc. First Intl. Conference on Audio and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, 395–402.

Phillips, P. J., and Rauss, P. (1997): *Face Recognition Technology (FERET Program).* Proc. Office of National Drug Control Policy, in press.

Pötzsch, M., Maurer, T., Wiskott, L., and von der Malsburg, C. (1996): *Reconstruction from Graphs Labeled with Responses of Gabor Filters.* Proc. ICANN'96, Springer-Verlag, Heidelberg, New York, 845–850.

Rinne, M., Pötzsch, M., and von der Malsburg, C. (1997): *Designing Objects for Computer Vision (FLAVOR).* In preparation.

Steffens, J., Elagin, E., and Neven, H. (1997): *PersonSpotter - Fast and Robust System for Human Detection, Tracking and Recognition.* Proc. Intl. Conf. on Automatic Face- and Gesture- Recognition, Nara, Japan (submitted).

Theimer, W. M., and Mallot, H. A. (1994): *Phase-Based Binocular Vergence Control and Depth Reconstruction Using Active Vision.* CVGIP: Image Understanding, 60(3):343–358.

Triesch, J., and von der Malsburg, C. (1996): *Robust Classification of Hand Postures against Complex Backgrounds.* Proc. Intl. Workshop on Automatic Face- and Gesture- Recognition, Vermont, 170–175.

Turk, M., and Pentland, A. (1991): *Eigenfaces for Recognition.* Journal of Cognitive Neuroscience 3:71–86.

Vetter, T., and Troje, N. (1995): *A separated linear shape and texture space for modeling two-dimensional images of human faces.* Technical Report, MPI for biological Cybernetics, TR15.

Wiskott, L. (1995): *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis.* Reihe Physik vol. 53. Verlag Harri Deutsch, Thun, Frankfurt a. Main.

Wiskott, L., Fellous, J. M., Krüger, N., and von der Malsburg, C. (1995): *Face Recognition and Gender Determination.* Proc. Intl. Workshop on Automatic Face- and Gesture- Recognition, Zürich, 92–97.

Wiskott, L., Fellous, J. M., Krüger, N., and von der Malsburg, C. (1997): *Face Recognition by Elastic Bunch Graph Matching.* IEEE Trans. on Pattern Recognition and Machine Intelligence, 19:775–779.