Pattern Recognition 30(6):837-846 (1997)

Phantom Faces for Face Analysis*

Laurenz Wiskott†
Institut für Neuroinformatik
Ruhr-Universität Bochum
D–44780 Bochum, Germany
http://www.neuroinformatik.ruhr-uni-bochum.de

Abstract

The system presented is part of a general object recognition system. Images of faces are represented as graphs, labeled with topographical information and local features. New graphs of faces are generated by an elastic graph matching procedure comparing the new face with a composition of stored graphs: the face bunch graph. The result of this matching process can be used to generate composite images of faces and to determine facial attributes represented in the bunch graph, such as sex or the presence of glasses or a beard.

Keywords: face analysis, sex discrimination, facial attributes, phantom faces, Gabor wavelets, elastic graph matching, bunch graph.

1 Introduction

The system presented here has not primarily been designed for face processing or even sex identification. It is rather part of a larger effort to develop a general recognition system, which can be applied to faces [1] as well as to any other class of objects [2]. It has also been shown that it can deal with many different aspects of object recognition such as rotation and scaling in the image plane [2], rotation in depth [3], or partial occlusions [4]. The object representation has the form of labeled graphs; the edges are labeled with geometrical information (distances), and the nodes are labeled with local features (jets). The *jets* are based on a Gabor wavelet transformation, which is a general and biologically motivated image preprocessing procedure. Graphs representing objects can be matched onto new images by maximizing a similarity function taking into account spatial distortions and the similarities of the local features.

The conceptional basis for this algorithmic system is Dynamic Link Matching, a fully neural system for robust object recognition [5, 6, 7, 8]. Dynamic Link Matching enriches conventional neural networks with the equivalent of pointers. These are realized as *dynamic links*, synapses rapidly switching on the basis of signal correlations. The graphs become layers of neurons and the matching is achieved by a fast selforganization process establishing regular connection patterns between these layers. The matching process can provide the system with invariance under translation, rotation, and scaling and robustness against distortions, e.g. caused by rotation in depth.

For the matching, a more general representation has been developed, the *bunch graph*, a composition of graphs of identical structure. In this structure, each node is labeled with a set of jets. By this means, the matching can choose between alternative jets for each node to achieve higher precision and be more generally applicable. The matching is described in detail elsewhere [9]. In this paper I focus on analyzing the matching result, generating composite or phantom faces, and determining facial attributes.

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

†Current address: Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, San Diego, CA 92186-5800, http://www.cnl.salk.edu/CNL, wiskott@salk.edu

2 Face Representation

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

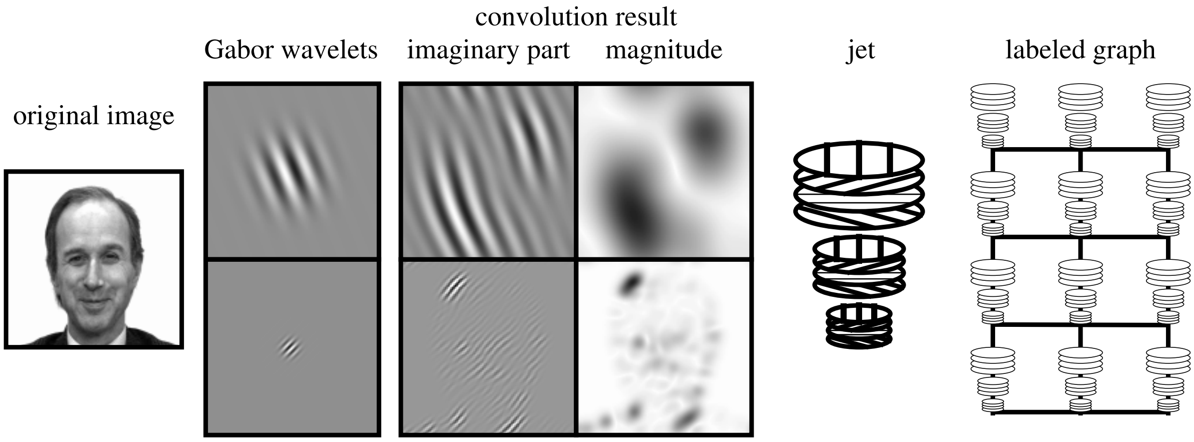

Figure 1: The graph representation of a face is based on a wavelet transform, a convolution with Gabor kernels of different size and orientation. The phase varies according to the main frequency of the kernels (see imaginary part) while the magnitude varies smoothly. The set of coefficients of the transform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

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

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

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

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

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

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

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

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

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

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

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

where $\vec{k}_j$ is the characteristic wave vector of the respective Gabor kernel and $\vec{d}$ is an estimated displacement vector which compensates for the rapid phase shifts. $\vec{d}$ is determined by maximizing $\mathcal{S}_\phi$ in its Taylor expansion around $\vec{d} = 0$, which is a constraint fit of the two-dimensional $\vec{d}$ to the 40 phase differences $\phi_j - \phi'_j$ (a detailed description of the estimation of the displacement vector $\vec{d}$ can be found elsewhere [9]).

The jets and the similarity function are robust against changes in lighting conditions in two respects. Firstly, since the kernels are DC free, the jets are invariant with respect to general offsets in the image grey values. Secondly, since the similarity function $\mathcal{S}_\phi$ is normalized, it is invariant with respect to contrast variations.

The first graph has to be defined manually, but additional graphs can be generated automatically by comparing the stored graph with a new image using an elastic graph matching process.

3 Elastic Graph Matching

Assume a graph $\mathcal{G}^\mathcal{M}$ of a face has been defined for a particular image by selecting a set of image locations $\{\vec{x}_n^\mathcal{M}\}$ as nodes and connecting the nodes by edges. The nodes are labeled with the Jets $\mathcal{J}_n^\mathcal{M}(\vec{x}_n^\mathcal{M})$ and the edges are labeled with the distance vectors $\Delta\vec{x}_e^\mathcal{M} = \vec{x}_n^\mathcal{M} - \vec{x}_{n'}^\mathcal{M}$, with edge e connecting node n' with n ($n = 1, ..., N$; $e = 1, ..., E$). This graph serves as a stored model. For a new image, a graph $\mathcal{G}^\mathcal{I}$ of identical structure can be defined by selecting an arbitrary set of locations $\{\vec{x}_n^\mathcal{I}\}$ (the number of locations must be the same as for the model graph), adopting the model graph structure (i.e. the information as to which edges connect which nodes), and deriving the node and edge labels from the new image as above for the model image.

Since we are not interested in an arbitrary image graph, but in an image graph that is most similar to the model graph, we have to select a set of node locations which maximizes the similarity function

$$\mathcal{S}_\mathcal{M}(\mathcal{G}^\mathcal{I},\mathcal{G}^\mathcal{M}) = \frac{1}{N}\sum_n \mathcal{S}_\phi(\mathcal{J}_n^\mathcal{I},\mathcal{J}_n^\mathcal{M}) - \frac{\lambda}{E}\sum_e \frac{(\Delta\vec{x}_e^\mathcal{I} - \Delta\vec{x}_e^\mathcal{M})^2}{(\Delta\vec{x}_e^\mathcal{M})^2},$$

which takes into account the average jet similarities and the spatial distortion of the image graph relative to the model graph (the relative strengths of these two terms is determined by λ). We apply a heuristic schedule of varying the node locations $\{\vec{x}_n^\mathcal{I}\}$ to maximize $\mathcal{S}_\mathcal{M}$ and refer to this process as *elastic graph matching*. It is described in greater detail elsewhere [1, 9]. The result is an image graph $\mathcal{G}^\mathcal{I}$ with the structure of the model graph and the nodes located at the respective facial landmarks. This graph then can be used for further processing such as comparison to a large gallery of faces, without need for doing the matching repeatedly.

4 Face Bunch Graph

The elastic graph matching works well if the new face is similar to the one represented by the model graph, but it fails if the faces look different. As a solution to this problem, one could match a number of different graphs to an image separately and accept only the result of the best match. But due to the large variety of faces, this would require a large number of different model graphs in order to become reliable. Instead I combine a representative set of individual model graphs into a stack-like structure which can be matched as a whole, see Figure 2. Each model has the same graph structure and the nodes refer to the same facial landmarks, termed hereafter fiducial points. (Actually, since the model graphs have a rectangular grid, only the left and right eye node and the height of the mouth nodes are in exact register. The other nodes are not perfectly aligned due to variations in facial geometry.) All jets provided by nodes referring to the same fiducial point are bound together and represent various instances of this local face region. I refer to this set of jets as a *bunch.* A *bunch graph* has bunches as node labels instead of single jets. The geometry, i.e. the set of edge labels, is simply the average over all models constituting the bunch graph.

When matching a bunch graph to a new image, the so called *local expert* which is the best fitting jet in each bunch, is used for comparison and positioning (of course you have to check all jets in a bunch to find the local expert). Thus, the similarity function $\mathcal{S}_{\mathcal{M}}$ changes to

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

where $\mathcal{G}^{\mathcal{B}}$ denotes a bunch graph containing M models. Since for each node a different model may provide the local expert, one takes full advantage of the combinatorial power of the bunch graph. For example, from the bunch graph the left-eye jet can be taken from model 3, the nose jet from model 28, the left-corner-of-the-mouth jet from model 11, etc. By using a face bunch graph instead of a single graph, the elastic graph matching can find the fiducial points more reliably and for a much larger variety of faces. In addition, the information as to which model provides the local expert at which node can be used for analyzing a face with respect to attributes which are more abstract than the identity of the person.

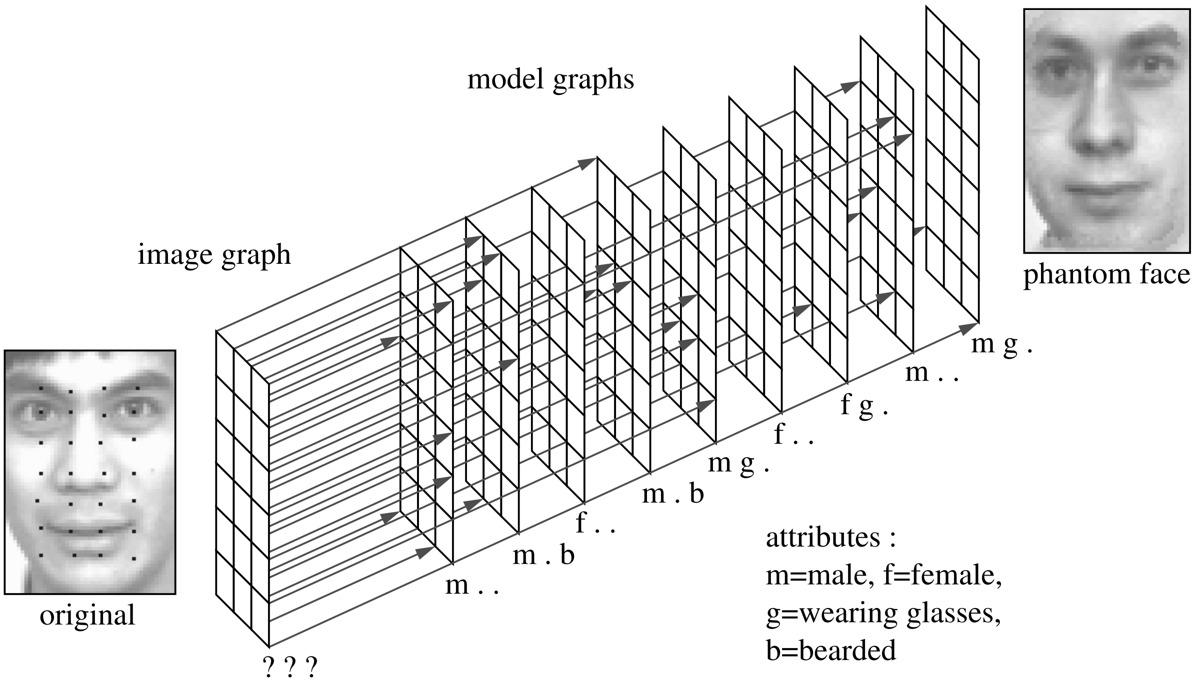

Figure 2: The stack structure of the face bunch graph. It is shown how the individual nodes of an image graph will fit best to different model graphs. Each model graph is labeled with known attributes, on the basis of which the attributes of the new face can be determined.

5 Phantom Faces

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

First I am going to reconstruct the face image on the basis of the matching results, generating a phantom face resembling the original. Each node of the graph has associated with it an expert model, i.e. the one model providing the local expert for that node. For reconstruction, I take local grey value patches directly from the expert model images at the respective node locations; the reconstruction is not based on Gabor wavelet coefficients. These patches are centered on the image graph nodes and joined with smooth transitions. This is done by a weighted average over grey values from different patches, each patch having a Gaussian weight function centered on the node. This simple method gives a good reconstruction of the original face, see

Figure 3. Such a phantom face looks natural, though it is typically composed of patches from about ten to twenty different models. It is striking how similar the phantom faces are to the originals, but Figure 2 also reveals limitations. The original was half Asian while the bunch graph contained only Caucasian faces. Thus the phantom face is a Caucasian version of the original.

6 Determination of Facial Attributes

Since the phantom face looks so similar to the original, it can be expected that more abstract attributes can also be inferred from the expert models. If, for example, the local experts are taken mostly from female models, one can expect that the phantom face will look female and consequently that the original face was probably a female as well. This also holds for other attributes, such as facial hair or glasses. If the expert models for the lower half of the image graph are mostly bearded, then the original face was probably bearded as well, and similarly for glasses. One only has to label all models in the face bunch graph with their respective attributes, decide which region of the face is relevant for a certain attribute, and then compare which attribute was most often provided by the experts in that region.

7 Node Weights

I have refined this simple idea by applying a standard Bayesian approach to determine automatically which nodes are most reliable. For each node n I introduce the stochastic variable X_n which can assume the values 1 or 0 depending on whether the respective expert model is of a particular attribute class (male, bearded, with glasses) or not (female, beardless, no glasses). X is the random variable for the image face itself. A sample of this stochastic variables is denoted by x_n and x respectively. Given an image with a certain value x of X, one can determine the probability $P(x_1, ..., x_N|x)$ of a certain combination of node labels. I make the strong assumption that the probabilities for the individual nodes are independent of each other $P(x_1, ..., x_N|x) = \prod_n P(x_n|x)$. The Bayes a posteriori probability for a new image having the attribute x given the node labels x_n then is

$$P(x|x_1, ..., x_N) = \frac{P(x)\prod_n P(x_n|x)}{P(1)\prod_n P(x_n|1) + P(0)\prod_n P(x_n|0)}.$$

The decision whether the attribute is present or not, i.e. $x = 1$ or $x = 0$, is then based on whether $P(1|x_1, ..., x_N) > P(0|x_1, ..., x_N)$ or not. This can easily be transformed into the more illustrative weights formulation in which the decision is made on the basis of a weighted sum over the nodes of one attribute. If one takes into account the formula for the Bayes a posteriori probability and the fact that the x_n may assume the values 0 and 1 only, one obtains:

$$\begin{aligned} P(1|x_1, ..., x_N) &> P(0|x_1, ..., x_N) \\ \Longleftrightarrow \quad \sum_n \ln\left(\frac{P(x_n|1)}{P(x_n|0)}\right) &> \ln\left(\frac{P(0)}{P(1)}\right) \\ \Longleftrightarrow \quad \sum_n x_n \left(\ln\left(\frac{P(1_n|1)}{P(1_n|0)}\right) - \ln\left(\frac{P(0_n|1)}{P(0_n|0)}\right)\right) &> \ln\left(\frac{P(0)}{P(1)}\right) - \sum_n \ln\left(\frac{P(0_n|1)}{P(0_n|0)}\right) \\ \Longleftrightarrow \quad \sum_n x_n\beta_n &> \theta, \end{aligned}$$

with

$$\begin{aligned} \beta_n &= \ln\left(\frac{P(1_n|1)P(0_n|0)}{P(1_n|0)P(0_n|1)}\right), \\ \theta &= \ln\left(\frac{P(0)}{P(1)}\right) - \sum_n \ln\left(\frac{P(0_n|1)}{P(0_n|0)}\right). \end{aligned}$$

The weights β_n are shown in Figure 4.

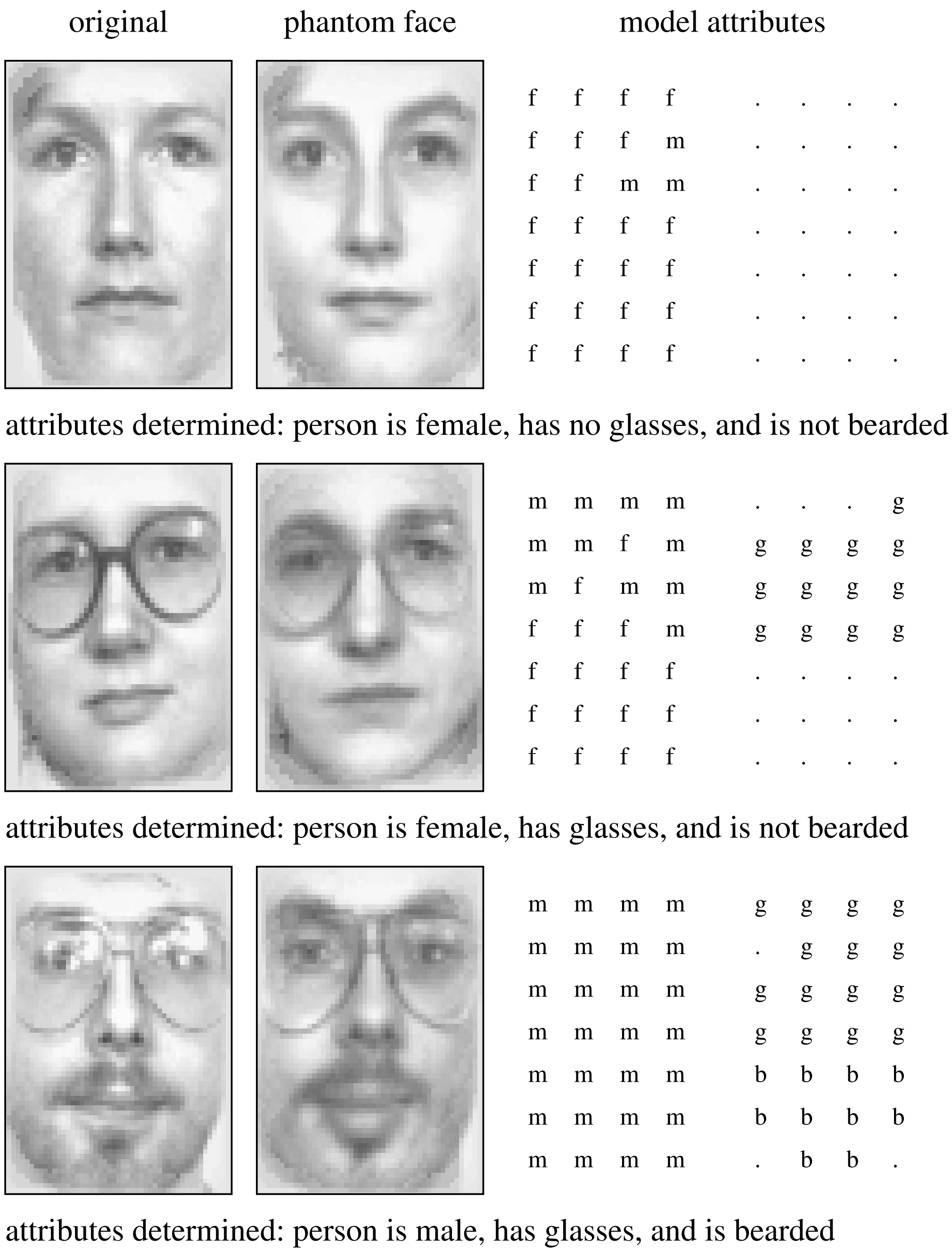

Figure 3: Shown is the original and the phantom face for three different persons. Notice that the phantom image was generated only on the basis of information provided by the match with the face bunch graph; no information from the original image was used. That is the reason why certain details, such as the reflections on the glasses or the precise shape of the lips of the top image are not reproduced accurately. The fields of labels on the right indicate the attributes of the models which were used as local experts for the individual nodes; m: male, f: female, b: bearded, g: glasses.

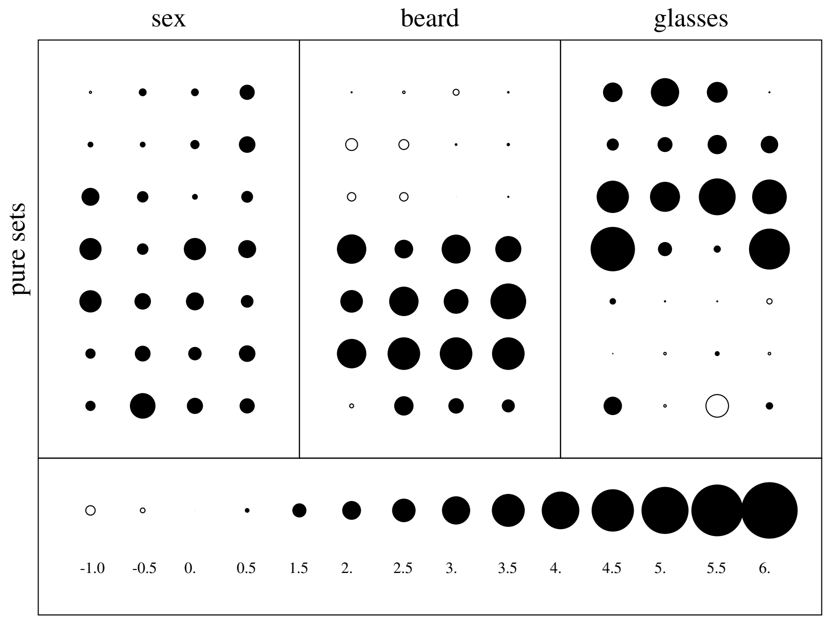

Figure 4: Weights β_n of the nodes. Radii grow linearly with the weights, white circles indicate negative values. From left to right for sex identification, for beard detection, and for glasses detection. The weights were determined on pure sets (see Section 8).

The conditional probabilities $P(x_n|x)$ are not known but have to be estimated on the basis of relative frequencies $F(x_n|x)$ evaluated on a training set of images for which the attributes are known. I avoided having conditional probabilities of 0 or 1 by enforcing at least one sample in each class. The absolute probabilities were chosen to be $P(1) = P(0) = 0.5$ in order to exclude prejudices about the test set composition derived from the training set.

8 Results

In the test runs I used a gallery of 111 neutral frontal views, 65% of which were male, 19% were bearded, and 28% had glasses. Each of the 111 faces was analyzed while the remaining 110 models constituted the face bunch graph. All faces were normalized in size while maintaining the aspect ratio. The influence of hair-style was reduced by a grey frame around the faces (see visible region in Figure 6). The 111 model graphs of 7×4 rectangularly arrayed nodes were positioned manually; the image graphs were generated automatically. Results are shown in Table 1 for the three attributes sex, beard, and glasses.

	sex	beard	glasses
complete sets (111/111/111)	0.924 ± 0.033	0.941 ± 0.027	0.963 ± 0.023
small sets (68/45/51)	0.875 ± 0.050	0.935 ± 0.054	0.919 ± 0.053
pure sets (68/45/51)	0.831 ± 0.050	0.857 ± 0.052	0.885 ± 0.043

Table 1: Rates of correct attribute determination: The complete set of 111 faces included 65% male faces, 19% bearded faces, and 28% faces with glasses; pure sets varied only in the considered attribute; small sets were of same size as the respective pure sets, but randomly drawn from the complete set. It can be seen to which extent the performance degrades for pure sets and how much of the degradation is due to the reduced face bunch graph size.

The sets of faces were always split randomly into a training set and a test set of same size. On the training sets the conditional probabilities were determined and on the test sets the performance was tested.

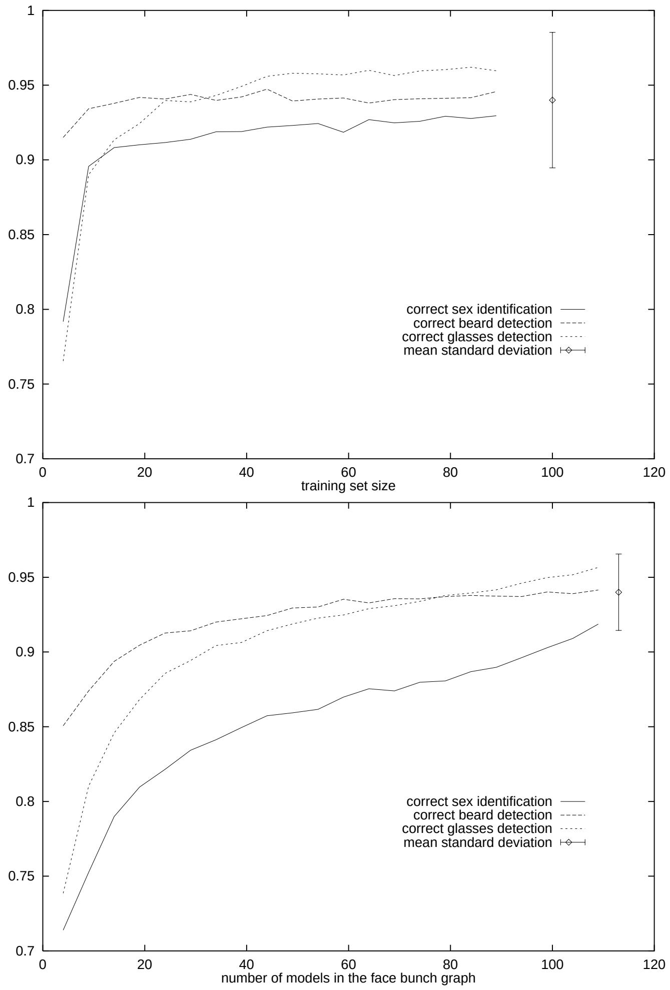

Figure 5: These graphs show the dependence of the mean rates of correct attribute determination depending on training set size (top graph) and bunch size (bottom graph). For the top graph, test set size was constantly 20 samples and the bunch size was 110, while the training set size varied from 4 to 89. For the bottom graph, training and test set had their standard size of 55 and 56 respectively, while the bunch size varied from 4 to 109. For each data point 500 different samples of the training and test set were chosen randomly to get a reliable mean performance.

For each figure, several hundred sample sets were drawn randomly. The standard deviation is shown in the table as well.

Besides the results for the complete sets, results on small and pure sets are given as well. In a pure set, faces differ only in one of the three attributes, i.e. 68 beardless faces with no glasses for sex identification, 45 male faces without glasses for beard detection, and 51 beardless males for glasses detection. The performance degrades significantly. Part of this degradation on pure sets is due to the fact that attributes such as sex and beard can no longer cooperate. But part of the degradation is due to the reduced bunch graph size. For comparison, results on small sets of the same size as the pure sets randomly selected from the complete set are shown as well. This gives an impression on what fraction of the degradation is due to the reduced bunch graph size and not to the fact that pure sets were used.

Figure 5 shows the dependence of performance on training set size (number of models to determine the node weights) and the bunch size (number of models constituting the bunch graph). The graphs show that only a small number of training samples is required to get close to maximum performance. At least for sex identification and beard detection 20 samples seem to be enough; for glasses it is 40. On the other hand performance has not yet saturated with bunch size for glasses detection and sex identification. Thus one can expect that results would be improved significantly if more models were used in the bunch graph.

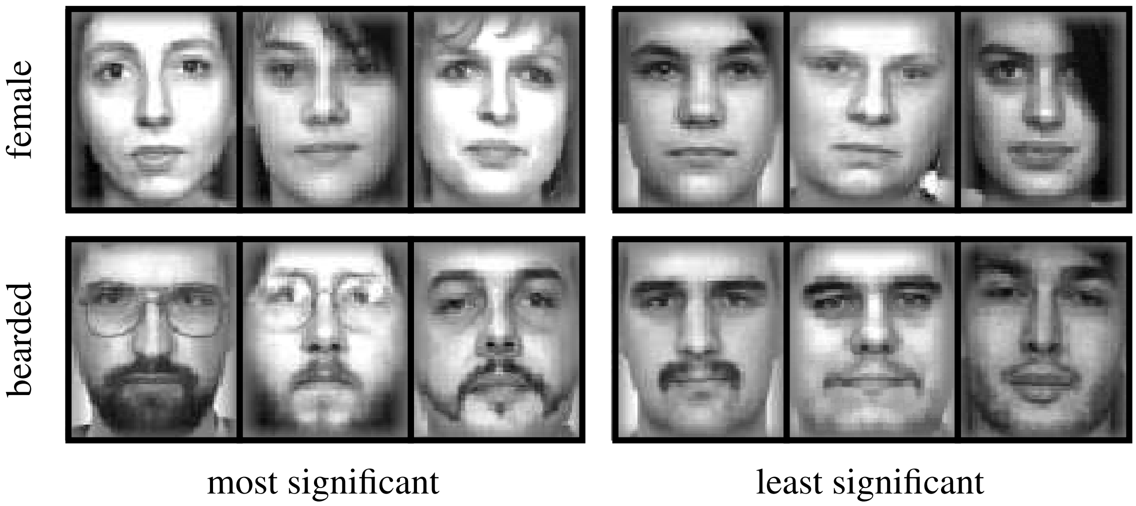

Figure 6: Female and bearded sample faces ordered by their attribute significance as judged by the system. Two of the three least significant females (last and third last) were the youngest ones in the gallery, of an age were facial sex is not yet fully developed. For the bearded males the order correlates well with contrast and extension of the beard.

Figure 6 shows several sample faces with respect to their significance for the attributes 'female' and 'bearded'. For the bearded males the order correlates well with contrast and extension of the beard. For the females it is conspicuous that the three least significant females included the two youngest ones in the gallery (last and third last), of an age where facial sex is not yet fully developed. For glasses no such obvious order was obtained. For the beardless persons one female was misclassified as being bearded because of her smile. Since smiles were not represented in the bunch graph, it was best captured by jets representing beards.

9 Discussion

I have demonstrated a simple and general approach for the determination of facial attributes. No extensive training is required. The system generalizes from a composite representation of single sample faces, the face bunch graph, by combining subparts into new composite or phantom faces. Abstract attributes can be transferred to a new face in a simple manner. The classification performance relies on what is represented in the face bunch graph: one cannot expect that with a Caucasian bunch graph the system performs well on Asian people, for example. I assume, however, that with an appropriate bunch graph, other attributes such as age, ethnic group, or facial expression can be detected.

9.1 Comparison with Other Systems

For the automatic classification of faces with respect to their sex, two different approaches have been used, one is based on extracted geometrical features, while the other is more holistic and works on the grey value images directly.

A system based on geometrical features was presented by Brunelli & Poggio [11]. Frontal view face images were automatically normalized with respect to rotation and scaling. Then 18 different geometrical features such as pupil-to-nose vertical distance, nose width, chin radii, and eyebrow thickness were automatically extracted, providing one 18 dimensional vector per image. No hair information was used. The data basis contained 168 images of 21 males and 21 females. A hyper basis function network was trained on the data sets of all minus one person and tested on the excluded ones. The mean performance on the training sets was 92% and on the test sets 87.5%.

A similar approach was used by Burton et al. [12], but the geometrical features were extracted manually. By means of a discriminant function analysis they selected 12 out of 18 distances between fiducial points.

Reference	Method	# train. images	Performance on test set [%]	train. set [%]
Brunelli & Poggio (1993)	18 geometrical features, hyper basis function network	4×41	87.5	92.0
Burton et al. (1993)	12 geometrical features (manually extracted), discriminant function analysis 16 geometrical 2D and 3D features	179		85.5 93.9
Golomb et al. (1991)	grey value images (manually aligned, limited hair information), back-propagation	80	91.9	
Gray et al. (1995)	grey value images (manually aligned), perceptron	2×80	81.0	
O'Toole et al. (1991)	grey value images (manually aligned(?), hair information), principal component analysis	80/167		74.3
Wiskott (this paper)	grey value images, Gabor wavelet transform, bunch graph incl. faces with facial hair (19%) and glasses (28%)	34/67 55/110	83.1 92.4	

Table 2: Methods, and performances of the different systems discussed.

The performance on the training set of 91 male and 88 female faces was 85.5%. No generalization results on a test set are given. (They also made experiments including profile images providing some 3D information. In addition, more complex features such as ratios or angles were extracted. By using 16 out of 27 features, they improved the training set performance to 93.9%.)

In contrast to the systems based on geometrical features there are systems which process the grey value images directly, usually after applying a normalization step, in which the faces were aligned and the grey values were corrected for a common contrast or offset.

Golomb et al. [13] employed a standard back-propagation network for sex identification. In 90 images of faces (45 beardless male, 45 female) the eyes were located manually and the images then rotated and scaled automatically to a standard format of 30×30 pixels. Images were compressed by an encoder back-propagation network with 40 hidden units. The output of these 40 units served as input for a sex identification network, the SexNet, trained with the back-propagation algorithm as well. 8 tests were performed with a training set of 80 images and 10 test images. Mean performance and standard deviation were 91.9%±8.6%. The system used limited hair information.

On the same database Gray et al. [14] applied a simple perceptron. With a similar experimental setup (normalized images of 30×30 pixels, training on 2×80 images (incl. flipped ones) and testing on 2×10 images), but with no hair information the mean performance was 81%±12%.

O'Toole et al. [15] applied principal component analysis to aligned 151×225 pixels images of 167 Caucasian and Japanese faces. The images were cropped to eliminate clothing, but hair information was retained. A simple criterion based on the reconstruction coefficients of the first four eigenvectors yielded a performance of 74.3% correct sex identification. No clear distinction between training and test set was made.

A comparison between the different systems shows that the system presented here achieves a comparable level of performance (see Table 2). But first of all it is interesting for conceptional reasons.

Firstly, it uses a different form of object representation. It is neither based on geometrical features nor on full face images. Instead it integrates both kinds of information by referring to fiducial points and evaluating local grey value distributions represented by jets. Though the localization of the nodes on the fiducial points is not currently used, the precisions seems to be high enough to extract useful geometrical features if object adapted graphs are used instead of the regular ones here [1]. The coding of the local grey value distribution by jets is general and biologically motivated.

Secondly, it relies much less on training (cf. to back-propagation, perceptron, and principal component analysis) or manual selection (cf. to manually defined or even manually extracted geometrical features) than all other systems. Actually, if one uses the same weight for all nodes, i.e. without any training, the performance on sex identification is almost as good as with learned weights. The determination capabilities rely almost purely on a simple piecewise combination of most representative samples, the local experts.

Thirdly, it is part of a general approach to object recognition and not specifically developed for the task of sex identification. As already mentioned in the Introduction, the same general system has been applied to recognition of individual faces and other objects, and it can easily be adapted to deal with rotation, scaling, and occlusions. This makes the system fully automatic and potentially applicable under a wide range of different conditions. The eigenfaces approach has also been used for face recognition [16], but the other systems are specific to the task of sex identification rather than part of a general effort to model object recognition.

One disadvantage of the system presented here is that it allows less visualization of the facial features by which the sex identification is made. The geometrical approaches can point to certain distances or ratios which are typical for a particular sex. The holistic approaches can visualize a typical male or female face, at least to a certain degree. Our approach can only provide a hint on which regions in a face are more discriminative than others. However, it might be possible to extend the Bayesian analysis to single coefficients of the jets and by this means obtain more accurate information which might allow reconstruction of a face which emphasizes features which are relevant for sex identification[1].

Another drawback of the system presented is that it is slow. The Gabor transformation and the elastic graph matching require about one minute on a Sun SPARCstation 10–512 with a 50 MHz processor. The code has not been optimized with respect to speed. However, it requires little time for training.

9.2 Future Perspectives

So far the attribute labels of the bunch graph models are binary and defined by hand. A male face can therefore be misclassified because it looks actually female or because there is a similar female face in the bunch graph that is male in appearance. The attribute labels should vary continuously from one extreme, e.g. male, to the other, female. Taking this into account should improve the determination performance. This gradual labeling would have to be learned from a binary one. It might also be possible to derive reasonable attribute classes for a given bunch graph autonomously.

Beside improving the system it would be interesting to see how it performs on other attributes or other classes of objects. With appropriate bunch graphs the system should be able to discriminate between different emotional expressions or races or it might be able to estimate the age of a face. Another class of objects would, for example, be domestic mammals. The system would then have to discriminate between dogs, horses, and sheep, despite the large variety, which can be within each group.

Acknowledgements

Many thanks go to C. von der Malsburg for his support and helpful comments. I would also like to thank Irving Biederman, Jean-Marc Fellous, Norbert Krüger, and Thomas Maurer for fruitful discussions.

References

[1] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition and gender determination," in *Proc. Int'l Workshop on Automatic Face- and Gesture-Recognition, IWAFGR'95, Zurich* (M. Bichsel, ed.), pp. 92–97, MultiMedia Laboratory, University of Zurich, June 1995. 1, 3, 10

[2] M. Lades, *Invariant Object Recognition with Dynamical Links, Robust to Variations in Illumination.* PhD thesis, Fakultät für Physik und Astronomie, Ruhr-Universität Bochum, D-44780 Bochum, 1995. 1

[3] T. Maurer and C. von der Malsburg, "Single-view based recognition of faces rotated in depth," in *Proc. Int'l Workshop on Automatic Face- and Gesture-Recognition, IWAFGR'95, Zurich* (M. Bichsel, ed.), pp. 248–253, MultiMedia Laboratory, University of Zurich, June 1995. 1

[1] This idea goes back to a comment by Irving Biederman.

[4] L. Wiskott and C. von der Malsburg, "A neural system for the recognition of partially occluded objects in cluttered scenes," *Int'l J. of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 935–948, 1993. 1

[5] C. von der Malsburg, "The correlation theory of brain function," internal report, 81-2, Max-Planck-Institut für Biophysikalische Chemie, Postfach 2841, 3400 Göttingen, FRG, 1981. Reprinted in E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks II*, chapter 2, pages 95–119. Springer-Verlag, Berlin, 1994. 1

[6] E. Bienenstock and C. von der Malsburg, "A neural network for invariant pattern recognition," *Europhysics Letters*, vol. 4, no. 1, pp. 121–126, 1987. 1

[7] W. Konen and J. C. Vorbrüggen, "Applying dynamic link matching to object recognition in real world images," in *Proc. Int'l Conf. on Artificial Neural Networks, ICANN'93* (S. Gielen and B. Kappen, eds.), (London), pp. 982–985, Springer-Verlag, 1993. 1

[8] L. Wiskott and C. von der Malsburg, "Face recognition by dynamic link matching," in *Proc. Int'l Conf. on Artificial Neural Networks, ICANN'95, Paris*, (Paris), pp. 347–352, EC2 & Cie, Oct. 1995. (ISBN 2-910085-19-8). 1

[9] L. Wiskott, *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*, vol. 53 of *Reihe Physik*. Thun, Frankfurt am Main, Germany: Verlag Harri Deutsch, 1995. (PhD thesis). 1, 3

[10] R. L. DeValois and K. K. DeValois, *Spatial Vision*. Oxford Press, 1988. 2

[11] R. Brunelli and T. Poggio, "Caricatural effects in automated face perception," *Biol. Cybern.*, vol. 69, pp. 235–241, 1993. 9

[12] A. M. Burton, V. Bruce, and N. Dench, "What's the difference between men and women? Evidence from facial measurement," *Perception*, vol. 22, pp. 153–176, 1993. 9

[13] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, "SexNet: A neural network identifies sex from human faces," in *Proc. Advances in Neural Information Processing Systems 3* (D. S. Touretzky and R. Lippman, eds.), pp. 572–577, SanMateo, CA: Morgan Kaufmann, 1991. 10

[14] M. S. Gray, D. T. Lawrence, B. A. Golomb, and T. J. Sejnowski, "A perceptron reveals the face of sex," *Neural Computation*, vol. 7, pp. 1160–1164, 1995. 10

[15] A. J. O'Toole, H. Abdi, K. A. Deffenbacher, and J. C. Bertlett, "Classifying faces by race and sex using an autoassociative memory trained for recognition," in *Proc. 13th Annual Conf. of the Cognitive Science Society* (K. J. Hammond and D. Gentner, eds.), Hillsdale (NJ): Lawrence Erlbaum, 1991. 10

[16] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991. 11