DISCRIMINATION OF FACIAL REGIONS BASED ON DYNAMIC GRIDS OF POINT REPRESENTATIONS*

R. HERPERS[†,‡] and G. SOMMER

Institute of Computer Science and Applied Mathematics, Cognitive Systems Group
Christian-Albrechts-University of Kiel, Preußerstr. 1-9, 24105 Kiel, Germany
[†] *GSF – National Research Center for Environment and Health*
Institute of Medical Informatics and Health Services Research, MEDIS
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

The application of an elastic graph-matching approach to discriminate facial image regions is presented. In contrast to the dynamic link architecture introduced by the Malsburg group, our application is not an identification task but a classification task. Therefore, our approach differs in several important aspects:

(1) the choice of the filter set,
(2) the selection of the positions of the nodes of the graph to represent the characteristic image information,
(3) the generation of a representative reference pattern needed for the calculation of the classifications, and
(4) a new two-step graph-matching approach based on the simulated annealing technique.

The approach was tested on facial regions taking the eye region as an example target. A classification performance for the verification of eye regions of more than 93% was achieved.

Keywords: Computer vision, classification, discrimination, dynamic graph matching, image processing.

1. INTRODUCTION

The classification and the discrimination of 2-dimensional patterns or more precisely 2-dimensional projections of 3-dimensional objects are still difficult tasks in computer vision. The high complexity of the task due to the high degree of freedom of the problem does not allow for a general purpose approach which is universally applicable. The difficulties arise from the variability of the patterns to be classified under real world conditions. Sometimes no invariances which unambiguously characterize a target are present or can be computed successfully applying known approaches. But biological visual systems perform this complex task very accurately and effortlessly.

Successful computational approaches have to be able to cope with the high degree of variability of real world input pattern which might be an arbitrary projection of the 3-dimensional scene. In this context, the development of approaches which search for invariances or invariant features under various projection angles and unknown scales is still a central objective of many computer vision research groups.

*This work is partially supported by the DFG Grants So 320\1-2, Ei 322\1-2, and He 2967\1-1.
‡Corresponding author. Department of Computer Science, 6 Kings's College Rd., University of Toronto, Toronto, Ontario, M5S 1A4, Canada. E-mail: herpers@vis.toronto.edu

International Journal of Pattern Recognition and Artificial Intelligence Vol. 12 No. 4 (1998) 381–405

In many of the approaches reported in the literature, several, sometimes very restrictive preconditions have to be fulfilled to ensure a successful classification. Motivated by biological visual systems neural network approaches have been developed in the last twenty years to enable flexible and adaptive representations of complex target patterns. But only relatively few successful solutions based on neural network approaches have been reported which have been generally tested on real world problems.[1,4–6,15,17,19,20]

The dynamic graph-matching approach developed by the Malsburg group and which can be interpreted also as a neural network approach[3,16,21] has presented both a flexible and a reliable solution to integrate a high degree of variability into the representation of a target of interest. By applying this approach the classification task is based on the evaluation of a defined set of invariant features linked with a framework which ensure the relative localization of the features. The relationship between the several features is maintained by a feature representation performed using a flexible, labeled graph. More details concerning this approach will follow in the subsequent section.

The discrimination module developed here is used as an essential component in building an image processing system to classify facial images.[7–9,13] Face recognition applications have become popular in the last ten years because of the high complexity of the pattern *face* and because of the outstanding performance of biological visual systems in recognizing faces. But the computer based recognition of facial images recorded under arbitrary preconditions has not been successfully solved until now. As mentioned above, the detection of invariant features is the main obstacle in this research field. To get past this obstacle, an attentive processing strategy has been developed which decomposes the entire classification problem into a set of better solvable subproblems.[13] The computation of these subproblems is then based on particular image regions which have been localized in advance and which have been spatially limited to the extent of the underlying object (Fig. 1).[7,13] The discrimination module, the objective of this contribution, should provide a decision about the benefit of applying more detailed and expensive analysis methods to the located spatially-limited facial regions. In the subsequent detailed analysis step, the exact positions of relevant keypoints or anatomical landmarks such as eye corners and mouth corners are determined[8] and evaluated.[11]

Firstly, the original dynamic graph-matching approach developed by the Malsburg group is introduced and discussed in more detail. Subsequently, the improvements and redeveloped methods used for this application are named and the motivations are discussed briefly. In Sec. 2, the idea of point representation together with the filtering scheme used in this contribution is introduced. Section 3 discusses the details of the improved object adapted graph followed by the graph matching in Sec. 4, which presents the cost function as well as details concerning the determination of the initial starting position and the subsequent local graph matching. Finally, in Sec. 5, classification results are presented and discussed.

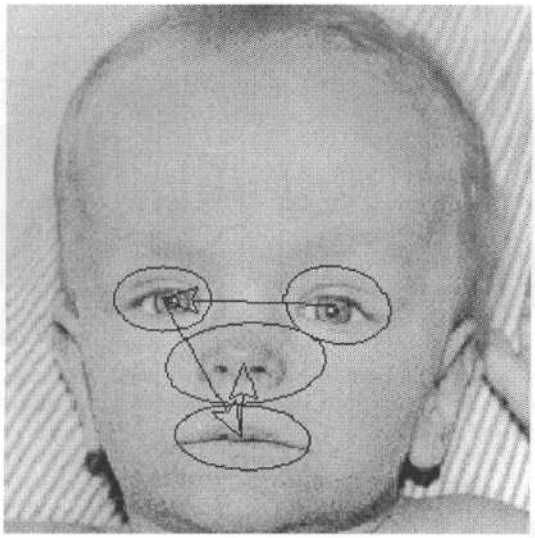


Fig. 1. Result of the attentive region localization. The main facial regions, such as the eye, mouth, and nose region, have been located in the first 4 localization steps. The computation is based on a saliency representation which carries the main salient features of the facial image.

1.1. Malsburg's Dynamic Graph-Matching Approach

In Refs. 3, 16 and 21 a dynamic graph-matching approach has been developed to identify faces from examples represented in a database. The underlying idea of this approach is that the representation of a face is computed at a set of particular image positions which obviously characterize an individual (Fig. 2). A set of Gabor wavelets is used for the descriptions of the faces at particular image positions. The projection coefficients of these filters (called jets) contain the gray-level distribution at the considered image points and their surrounding regions. The application of different scaled and rotated Gabor filters at several image positions is particularly suited to the representation of textured image structures. Therefore, the image positions to be represented do not have to be selected so precisely. Neighboring point descriptions (4 per neighborhood) are connected by a *homogeneous* grid or a graph as demonstrated in Fig. 2 (left). The connections of the graph ensure the spatial relationship between the point descriptions. But it should be emphasised that a large proportion of the point descriptions or nodes of the graph are calculated at non-characteristic image positions (e.g. in the hair, on the cheeks, or on the forehead).

For the computation of an identification of an individual the graph is matched stepwise to the image to be examined (Fig. 2 (right)). The aim is to compute an optimal point correspondence between the represented nodes of the investigated image and the stored representation. Finally, a threshold determines the success of the identification.

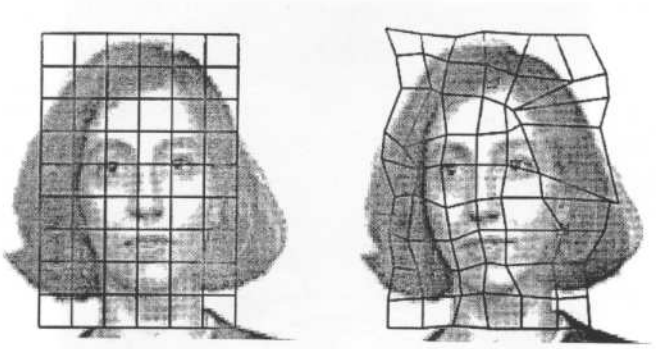


Fig. 2. Face identification applying point descriptions and graph matching. At each node of the grid the gray-level distribution of the local surroundings is represented in the parameters of the so-called jets (from Ref. 16).

1.1.1. Disadvantages

The main disadvantage of using the Malsburg approach for a multi class classification problem is that the computing effort increases drastically with the number of stored or represented classes respectively in sample images. This may decrease the flexibility and robustness of the approach. However, this approach is designed for the identification of images from example images stored in a database. Only comparisons between two individuals or between two images of the same individual are computed. Therefore, only the problem of the intra-individual variability is addressed. The inter-individual variability is used to calculate discriminations between individuals. Hence, for a multi class problem, a new representation which represents the class specific characteristic more than individual ones, has to be developed.

Furthermore, we think that the point representations essentially used in this approach should be located at characteristic image positions and not at arbitrary image positions defined by the intersection points of a homogeneous grid. Besides this aspect, the exact scaling of the used grid in the referred work is not addressed satisfactorily. The consequence of the representation only at characteristic image positions is that an object adapted graph has to be developed. Furthermore, at the characteristic positions, more specific than textured image structures have to be computed and represented. For this, we have selected a special filter set which might be more suitable to represent characteristic edge and line structures than Gabor wavelets do.

1.2. Classification of Facial Regions

A classification module which is able to discriminate between different classes of images of facial parts is required for the image processing system developed for our face processing application. In other words, an N-class problem has be solved. But for the special attentive approach used in this processing system, the classification problem can be reduced to a two-class problem because it is observed that the eye

regions are detected reliably by the attentive localization strategy in early fixation steps (Fig. 1).[7,13] This means that there is an increased probability for an early detection of eye regions in contrast to non-eye regions. This behavior is caused by the highly salient features within the eye regions (more prominent edges and line segments with a more dense distribution) which is also in accordance with the behavior of biological visual systems which fixate the eye region first.[22] This detection result might be essentially used as *a priori* knowledge for the design of an efficient classifier. Assuming that the probability of the detection of an eye region is high during the first fixation step only, this detection has to be proven. Therefore, the task of the classification module is restricted to verifying the hypothesis of the fixation of an eye region against the fixation of a non-eye region. In a subsequent processing step within the face-processing application, the facial regions, in particular the eye regions, are investigated in more detail to determine exactly the position of several anatomical landmarks.[8,9]

1.2.1. Redevelopment of the dynamic graph-matching approach

A direct transfer of the Malsburg approach to our verification task is obviously not possible for several important reasons. In contrast to their approach, our application is not an identification task but a classification task. Therefore, our approach differs in several important aspects:

1. the choice of the filter set,
2. the selection of the positions of the nodes of the graph to represent the characteristic image information,
3. the multiple connection structure of the object adapted graph (irregular grid),
4. the generation of a representative reference pattern, and
5. a new two-step graph matching based on the simulated annealing approach.

As already discussed in the previous section the main reasons for the redevelopments are the different kinds of classification problems that we have to tackle here. For our application, a representation of a class of images (e.g. eye regions) is needed which should be as independent of individual variabilities as possible. This stands in contrast to the approach of the Malsburg group which has to represent individual facial images. Therefore, a more uniform representation is needed which might be found by applying object adapted graphs. Using object adapted graphs which are adapted to the salient edge and line structures of the object, it is obvious to use also a special filtering scheme which is specially suited to represent edge and line structures. Therefore, we have decided to apply special edge and line detection filters rather than Gabor wavelets which are mainly suited to represent textured structures. For reasons of the object adapted graph, a new connection structure also has to be developed because a uniform connection structure has proven to be not so successful. The generation of a representative reference pattern is required by the representation of a class of images in contrast to the representation of individuals. Finally, the convergence of the graph-matching process is significantly improved by the application of a simulated annealing approach in which different temperature

levels ensure the graph flexibility during the final adaptation needed to find a good approximation of a global minimum.

1.2.2. Previous investigations

The design of the model graph used is the result of several investigations comparing the performance of the classification of different anatomical landmark positions. These investigations were based on the classification of keypoints and anatomical landmarks in the eye region as reported in Refs. 10 and 11. Based on the good performance of the keypoint classification approach by applying point descriptions, the hypothesis of the approach presented here is formulated. So, multiple point descriptions or '*nodes*' $\mathcal{N}$ positioned at particular characteristic image locations $\vec{x}_j$ $(j = 1, \ldots, J)$ and connected topologically may be sufficient for a reliable verification of a complex image pattern (Fig. 3). This results in a special graph which is adapted to the characteristic structures of the class of objects to be classified. Fundamental to this processing is that the point descriptions are mainly computed at really characteristic image positions which are composed of significant edge and line segments.

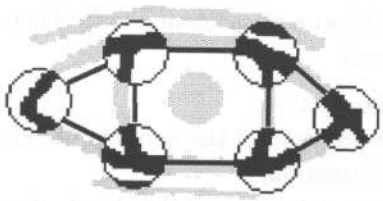


Fig. 3. The idea of connected point descriptions demonstrated at an eye region. The descriptions at the circled keypoints and the spatial relationship between them may be sufficient to clearly represent the underlying object.

2. POINT DESCRIPTIONS

The underlying assumption of the approach presented here is that the common information of the class of image regions is given by a characteristic arrangement of particular edge and line segments as well as particular junctions and intersections (Fig. 3). To this end, a representation has to be found which encodes the underlying edge and line structures at the considered image positions.

The point descriptions or labels $L(\vec{x})$ are established from a set of projection coefficients $l_i(\vec{x})$ achieved by projecting a particular image position and its surrounding region $I(\vec{x})$ to several filters f_i $(f_i \in \mathcal{F})$.

$$l_i(\vec{x}) := f_i \otimes I(\vec{x}) \qquad \text{for } i = 1, \ldots, S\,. \tag{1}$$

The dimension of each label $L(\vec{x})$ is equal to the number i of applied representation filters f_i. In the following, the particular labels L used for the several point descriptions at the different image positions $\vec{x}_j$ will be denoted by $L_j(\vec{x})$ or abbreviated by the term 'label of j' L_j (with $j = 1 \ldots J$). The index j is related

to a particular semantic image structure or anatomical landmark.[a] In other words, this index j should be understood as a semantic index rather than as a position dependent index related to a fixed image position.

2.1. Filter Selection

The main goal for the composition of a set $\mathcal{F}$ of representation filters f_i is a clear and obvious description of the image structure characterizing the common underlying information of the entire class of images to be classified. Only the characteristic image information for a reliable and robust classification of each possible image should be encoded. Here and in the following we will understand the image structure of interest to be a characteristic arrangement of edge or line structures (as demonstrated in Fig. 3) rather than the more general term of image texture.

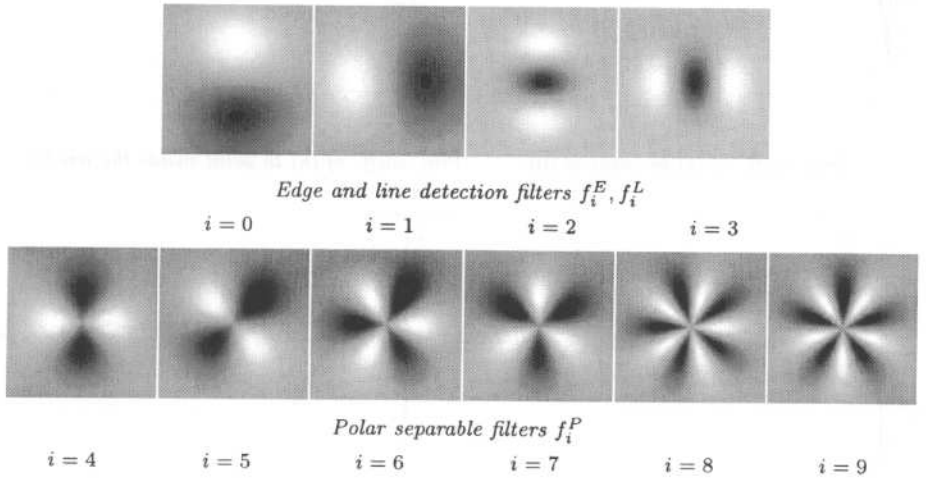


Fig. 4. Set of filters which are applied to each node of the graph to describe the underlying edge and line structure. First row: Gaussian edge and line detection filters (filter index $i = 0 \ldots 3$) ($\sigma = 6$). Second row: polar separable filters (filter index $i = 4 \ldots 9$) with different angle-dependent oscillations ($m = 2$, $m = 3$ and $m = 5$, respectively). For the several orthogonal partners of the polar separable filters the original filters are turned by adding $\frac{\pi}{4}$ for $i = 5$, $\frac{\pi}{6}$ for $i = 7$, and $\frac{\pi}{10}$ for $i = 9$ to φ. The kernel size for all filters is 27×27 pixel.

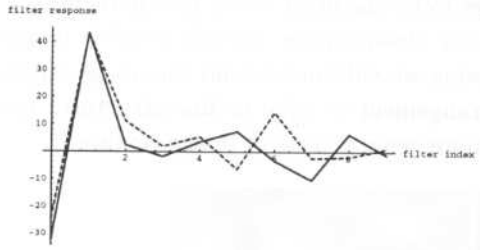
In general, for the task to be solved in this application, a complete description of the considered image positions which allow for an error free reconstruction of the complete gray-level distribution is not worthwhile. Moreover, the characteristic edge and line information should be encoded as clearly and obviously as possible to ensure a reliable and robust discrimination between different structures (see Fig. 5).

An additional goal which has to be considered during the choice of appropriate filters is to achieve a high degree of similarity between the representations and therefore, also between the filter results. Even if the edge and line structures which are to be compared are slightly different but derived from images with a similar

[a] e.g. As the index for the point representation of the inner eye corner in contrast to the one for the outer eye corner.

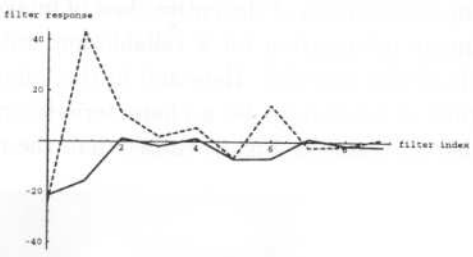
Reference point (a) *Similar point* (b) *Three different points* (c)

Original eye region with marker for particular considered position(s)



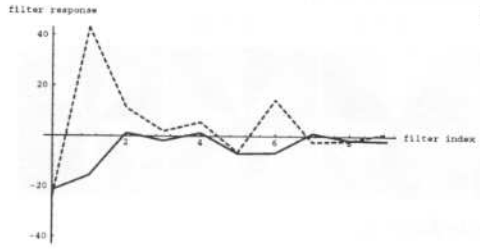
(d)

Proj.-coeff. of (a) *to point of* (b).



(e)

Proj.-coeff. of (a) *to point inside the iris* (c)



(f)

(g)

Proj.-coeff. of (a) *to point upon the eyelashes* (c). *Proj.-coeff. of* (a) *to point on the skin* (c)

Fig. 5. Projection coefficients [proj.-coeff.] of the reference point (a) (dotted line in any case of (d)–(g)) calculated in comparison to different image positions (b), (c) (black line in (d)–(g) respectively). The left eye region shows the reference point (a), the second eye region shows an example of a test pattern with a similar underlying image structure (b), and the right eye region shows three different examples with non similar image structures (c). The first graph (d) shows the comparison of the reference pattern (a) to the test pattern (b). The filter results for the edge detection filters (filter index $0, 1$) are prominent while those of the line detection filters are poor (filter index 2, 3) for both points. The filter results for the polar separable filters (filter index $4 \ldots 9$) show that the orientation of the image structure is slightly different. But in general, the main character of the represented image structure from the two compared image positions is qualitatively the same. The second graph (e) shows the comparison with the test point inside the iris, the third graph (f) with the test point inside the eyelashes, and the fourth (g) with the test point on the wrinkled skin below the iris. For each demonstrated image position the derived representation can be distinguished very reliably from the reference structure because the several filter responses show qualitatively and quantitatively very different results. In the last case of the wrinkled skin below the iris (c) no clear underlying image structure can be encoded. Therefore, the projection coefficients have very poor values (g).

general expression, the filter responses should represent appropriately the image structure, [compare Figs. 5(a), (b) and (d)]. That means that an invariant representation has to be found which encodes edge and line structures with different qualitative expressions e.g. bright edges in relation to sharp ones. Therefore, Gaussian derivative edge and line detection filters are used for a reliable representation of the edge and line structures of interest.[b]

To ensure a reliable representation of the image structure $I(\vec{x})$ at particular image positions $\vec{x} = (x, y)^T$, the following orientation selective edge and line detection filters (f^E, f^L) are applied (see Fig. 4, first row):

$$F^E(x,y) := -\frac{1}{\sqrt{2\pi}\sigma^3}\ x\ e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

$$F^L(x,y) := \frac{1}{\sqrt{2\pi}\sigma^3}\ (\frac{x^2}{\sigma^2} - 1)\ e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

where σ is a scaling value.[c]

However, a representation based only on these four line and edge detection filters ($f_1^E, f_2^E, f_3^L, f_4^L$ see Fig. 4, first row) is neither sufficient for a clear description of any important image structure nor for a numerical calculation of the distinctiveness. Additional features have to be taken into account (e.g. curvatures, different orientations, and other more complex structural information apart from the clear or optimal edge structures) to enable a clear discrimination between all the important image positions.

Several polar and cartesian separable filters based on Gaussian derivatives have been investigated to find an appropriate filter set $\mathcal{F}$ for the representation of the local image structures. It turned out that polar separable filters were superior to cartesian filters. More complex filters have been investigated but they did not represent the characteristic information as reliably as required for our classification task. They encoded gradually increased individual features which were inappropriate for our classification task. Also filters with too many radial or angular oscillations decrease the representation performance.

Three polar separable filters $(f^P_{4,\ldots,9})$ which selectively catch the local energy encoded in the surrounding region have been added to the set of representation filters but with only a small number of angle-dependent oscillations (see Fig. 4, second row). The polar separable filters used are defined by the following equation:

$$F^P(r,\varphi) := r\ e^{-\frac{r^2}{2\sigma^2}} \cos(m\varphi) \tag{4}$$

with radius $r(x,y) := \sqrt{x^2+y^2}$ and angle $\varphi(x,y) := \arctan\frac{y}{x}$, m counts the number of angle-dependent oscillations and σ is a scaling value.

[b]The use of a uniform normalization of the filter kernels (L^1-norm concerning the size of the filter template of all filters $f_{1,\ldots,9}$) in addition to a selection of a high scaling value ($\sigma = [5\ldots8]$) of the Gaussian derivative filters will ensure the necessary comparability of the labels computed in cases of differently expressed edge or line structures (see also Sec. 3.2.).

[c]Large scale edge and line detection filters are more advantageous than small ones because they are not so selective. Therefore, they will provide comparable filtering results in cases of different expressed edge and line structures (for more details see Refs. 13 and 18).

3. LABELED GRAPH

Based on the labels L calculated for each characteristic image position by the filter set $\mathcal{F}$, the labeled graph $\mathcal{G}$ is defined as a set of nodes $\mathcal{N} := \{N_1, \ldots, N_J\}$ and a set $\mathcal{E}$ of edges, where each member or element $E_{j,k}$ denotes the edge between node N_j and node N_k $(j \neq k)$.

$$\mathcal{G} := \{(\mathcal{N}, \mathcal{E})\} \tag{5}$$

Each node N is labeled by its position $\vec{x}$ and a S-dimensional label L:

$$N := \{(\vec{x}, L) | \vec{x} = (x, y), L = l_1, \ldots l_S\}\,. \tag{6}$$

Each component $l_i, i = 1, \ldots, S$ of the label L is the response (projection coefficient) of the ith filter from a given set of filters $\mathcal{F}$ (Fig. 4).

Several nodes of the graph are connected whereby multiple edges between the nodes are necessary to maintain the spatial relationship between the nodes (Fig. 6). Topologically neighboring nodes are connected by edges and are labeled with the distance.[d] The distance between a node N_j and a node N_k is defined as the Euclidean distance between their positions $\vec{x}_j$ and $\vec{x}_k$:

$$d_{j,k} := \|\vec{x}_j - \vec{x}_k\|\,. \tag{7}$$

3.1. Synthetic Reference Graph

A synthetic reference pattern $\mathcal{G}^R$, called *reference graph*, which is calculated from an average of several examples is introduced to provide a representative model. Superpositions of several individual local structures such as eye corners result in a more robust reference pattern than a single individual one. Therefore, a number of examples are selected manually to generate a representative reference pattern which describes the common structure of all class members.

$$\hat{L}_j^R(\vec{x}) := \frac{1}{n} \sum_{h=1}^{n} L_{j,h}^R(\vec{x}) \tag{8}$$

where n is the number of examples of corresponding labels $L_{j,h}^R(\vec{x})$ of the reference pattern related to a particular kind of landmark j (fixed). To establish also reliable reference values for the edges between the nodes of the model graph, distance ranges with a mean value and a standard deviation are introduced, which are also based on the same set of n examples.

[d] It has been shown that for each node of a graph the number of edges should be nearly equal. The number of edges can be reduced without any significant influence to the classification result only for nodes which have a lower importance for the representation (e.g. the eye corners). In the example, shown in Fig. 6, multiple connections (4-5) between the nodes are evident. This has the advantage that the spatial relationship is saved very carefully.

3.2. Similarity Function

To quantify the similarity between corresponding labels $\hat{L}_j^R$ and L_j^T the normed scalar product of the two feature vectors is calculated as follows:

$$S(\hat{L}_j^R, L_j^T) := 1 - \frac{\langle \hat{L}_j^R, L_j^T \rangle}{\|\hat{L}_j^R\| \cdot \|L_j^T\|} \tag{9}$$

where the index ^{R} stands for the representation of the reference pattern and the index ^{T} marks a representation of a test pattern. By defining this similarity function the domain of the function $S(\vec{x}, \vec{y})$ lies between 0 for identical feature vectors and 2 for feature vectors with exactly opposite directions ($Def(S) = [0 \ldots 2]$).[e] Another advantage of the presented definition is that it is invariant with respect to global changes of the image contrast and to the brightness of the image which may occur under varying recording conditions.

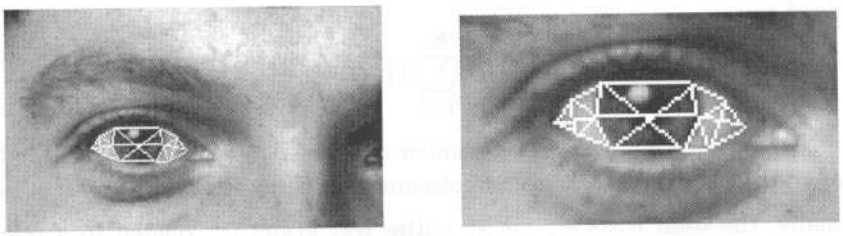


Fig. 6. Adapted graph for the class of eye regions demonstrated at one sample eye region. Left: the original image part with the model graph. Right: the enlarged graph of the eye to document the spatial relationship between the nodes.

4. GRAPH MATCHING

To match an individual pattern G^T to the reference pattern G^R a two-step graph-matching approach has been developed. But beforehand the two-step graph-matching algorithm will be introduced, a cost function has to be defined in advance.

4.1. Cost Function

To evaluate the success or the quality of a graph-matching process a cost function has to be formulated with which the adaptation of the graph can be quantified. The total cost C_g of a graph of a test pattern $\mathcal{G}^T$ has two components.

4.1.1. Costs due to nodes

The **first** component quantifies the similarity of each label L^T of the nodes of a graph in relation to the reference labels L^R based on the similarity function S [see Eq. (9)]. In the following, this cost component will be denoted by 'cost due to

[e] This issue is advantageous for an efficient and robust calculation.

nodes' C_n. The cost function C_n for the similarity[f] (or distinctiveness) of an entire graph is defined as the sum of the similarities of the labels L_j^T of all constituent nodes N_j or as:

$$C_n := \sum_{j=1}^{J} S(L_j^R, L_j^T). \tag{10}$$

4.1.2. Costs due to connecting edges

The **second** component quantifies the changes in the spatial relationship between the nodes. All the connecting edges $E_{j,k}$ between the nodes are evaluated and the sum of all evaluated connecting edges gives the cost component C_c due to the connecting edges. The calculation of the cost function C_c is based on the relative length of the actual connecting edge $d_{j,k}^T(t)$ of a node pair (N_j, N_k) of a test pattern $\mathcal{G}^T$ just considered relative to the length of the connecting edge $d_{j,k}^R$ in the reference graph $\mathcal{G}^R$ [see also Eq. (7)]. The component of the costs due to the connecting edges is defined as:

$$C_c := \sum_{j=1}^{J} \frac{1}{K_j} \sum_{k=1}^{K_j} f\left(\frac{d_{j,k}^R - d_{j,k}^T(t)}{d_{j,k}^R}\right) \tag{11}$$

where K_j (with $1 \leq K_j \leq J$) is the number of edges $E_{j,k}$ (with $k \in 1, \ldots J$, j fixed) related to the node N_j and f is a monotonous increasing function (e.g. $f(z) := |z|$).

Finally, the total costs C_g for an entire test graph $\mathcal{G}^T$ related to a reference graph $\mathcal{G}^R$ are given by

$$C_g := C_n + \lambda \; C_c \tag{12}$$

where λ is a weighting factor[g] ($\lambda \geq 0$) to control the quality of the matching process.

4.2. Determination of the Initial Starting Position

To initialize the adaptation of the subsequent graph-matching process to a test pattern the reference graph $\mathcal{G}^R$ has to be shifted roughly to an appropriate starting position in advance. This processing is necessary to enable a reliable selection of the best starting position independent of any translation or shift of the pattern of interest in the image part (considering both shifts in the X- and Y-directions) (Fig. 7). During the computation of the initial starting position the shape of the graph is kept rigid. No individual movement of the nodes of the graph is allowed during this initial matching phase, only the scale of the graph is varied to achieve a higher scale invariance. To do this, the model graph is scaled or better adapted to a number of distinct sizes by keeping the general shape rigid and adopting the point representations at the several node positions.

[f]The index n of the cost function should denote the kind of costs due to the similarity between corresponding nodes or more precisely due to the different projection coefficients $l_{i,j}$ of the feature vectors L_j^T.

[g]Low values of λ will result in strongly distorted graphs while higher values will maintain the global relationship between the nodes.

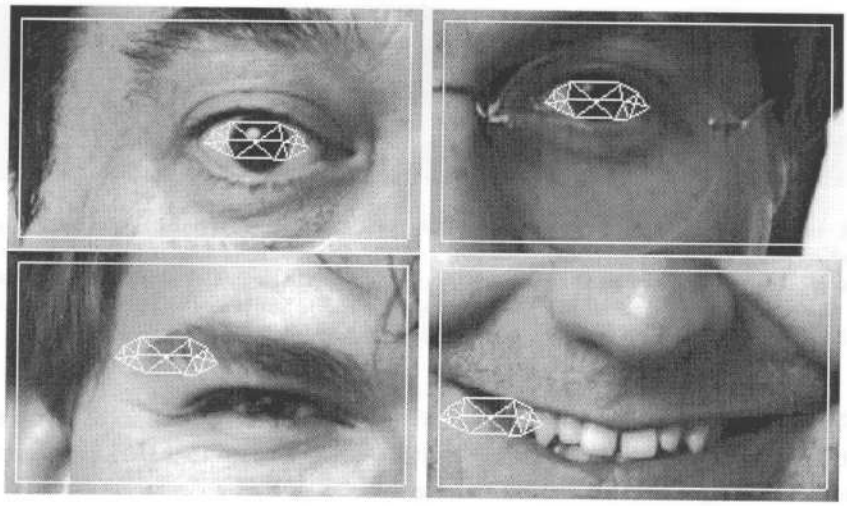


Fig. 7. Calculation of the best starting position for the model graph in four examples. In the first row, two successful examples for eye regions are shown. Even if spectacles cover the eye region, the computation of the best starting position may succeed (first row right). The computation of the best starting position may fail if the remaining characteristic information of the eye region is too poor, here caused by a "mimic" expression (screwed eye) (second row left). If a non-eye region is computed (second row right) the calculation will terminate at a position with the greatest similarity to the represented edge information.

For the determination of the initial starting position, the graph similarity $C_g(x_i)$ is calculated for all image positions in the considered test image ($\forall \vec{x} \in I(\vec{x})$) and the image position with the minimum costs $C_g(\vec{x})$ indicates the best starting position (Fig. 7). The component of the total costs C_g due to the connecting edges is equal to 0 ($C_c = 0$) because the graph is moved undistorted. The same is valid when the initial scale is varied.[h] In other words, the determination of the initial starting position is based only on the similarity of the point representations $C_g := C_n$ (that means $\lambda := 0$, refer to Eq. (12)) in each case.

4.2.1. Invariance properties

A degree of rotation invariance can be integrated into the determination process if different rotated examples of the model graph $\mathcal{G}^R$ are used. The computation effort increases drastically, however, if different scales are also simultaneously checked. Therefore, in the presented realization no additional effort is made to achieve further rotation invariance, since firstly it can be assumed that the facial regions and in particular the eye regions are in an approximately horizontal orientation and secondly, the algorithm developed is already able to successfully detect slightly rotated eyes (in a range of $\pm 10°$) using just one horizontal model graph (Fig. 8).

[h]When testing different initial scales to find the best starting position the scale of the reference graph is also adapted so that the connection costs will remain at 0.

Fig. 8. Invariance property of the graph-matching approach with regard to rotations. Stepwise altered rotation angle 5° (left), 10° (middle), and 15° (right). Rotations within the range ±10° are acceptable without the need to consider any rotation of the model graph. Even in isolated cases above this empirical threshold, the determination of the initial starting position may succeed but the subsequent local matching may fail (right).

For the scale invariance properties of this initial determination of the starting position it was found that size changes within a range of up to ±10% are acceptable without checking different scaled model graphs (Fig. 9).

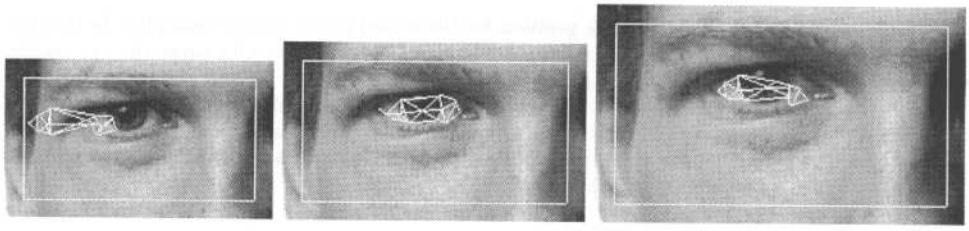


Fig. 9. Invariance property of the graph-matching approach with regard to scale changes. Stepwise varied scale −20% (left), −10% (middle), and +10% (right). Size changes within a range of ±10% may be acceptable without checking differently scaled model graphs. Object sizes above this threshold will result in failures (left).

Furthermore, the approach is invariant with respect to noise. Noise levels above 1 dB are acceptable for a successful determination of the initial starting position and for a successful computation of the subsequent local matching (Fig. 10). Even in images with drastically added noise (SNR = 0.5 dB), the determination of the initial starting position may succeed but the subsequent local matching may fail [Fig. 10 (right)].

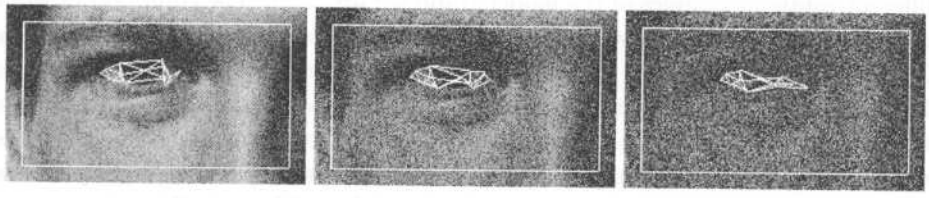


Fig. 10. Invariance property of the graph-matching approach with regard to noise. Example image with added noise, Signal/Noise-Ratio of 2 dB (left), 1 dB (middle), and 0.5 dB (right). The determination of the initial starting position succeeds in each case while the computation of the subsequent local matching will fail with a $SNR < 1$ dB (right).

4.3. Local Matching

After the determination of the initial starting position, the nodes of the graph are moved relative to their neighbors to achieve an optimal correspondence to the underlying image structure for each constituent node. Theoretically, by moving n nodes of a graph each independently in two directions and considering also an adapted scale and the right orientation of the entire graph a $2n+2$-dimensional optimization problem has to be solved. The solution of such a high dimensional problem considering on average graph sizes of more than 10 nodes, cannot be calculated exactly in reasonable computing time.

For this optimization problem, a heuristic numerical approach is applied which is able to determine an acceptable approximation of the optimal solution. In the realization, a simulated annealing approach[14] is employed to compute such an approximation.

4.3.1. Simulated annealing

The simulated annealing approach is a well-known method to determine efficiently a global minimum in a high dimensional search space which approximates well the optimal solution.[14] The underlying idea is adapted from a thermo-dynamic model in molecular physics. The model describes the controlled cooling of matter dependent on a temperature τ. At high temperatures the molecules are able to move fast and therefore, cover greater distances. The freedom of movements of the molecules become restricted stepwise as temperatures decrease until the matter has cooled down completely. At the end of the cooling process there is a systematic order between the molecules. But this is only possible if the temperature has been high enough in advance to enable the molecules to move large distances.

Adapting this property to a high dimensional search problem, a high temperature at the beginning $\tau(0)$ allows the accommodation of stages with higher energy levels or in other words allows local maxima to be overcome in the high dimensional search space.[i] Another important aspect of this approach is that the temperature has to be reduced gradually and in controlled decreasing steps to avoid a too sudden freezing of the current state. This increases the likelihood of finding a local minimum which best approximates the optimal solution.

For each node N_j an alternative position $\vec{x}(t+1)$ is selected randomly inside a well-defined surrounding region $\mathcal{R}$ (e.g. $\mathcal{R}$ is a 7×7 surrounding region) and the expected total costs caused by the entire graph $\hat{C}_g(\vec{x}(t+1))$ (expected graph) are calculated. The cost difference $\Delta C(\vec{x}(t))$ between the costs expected $\hat{C}_g(\vec{x}(t+1))$ of the suggested position $\vec{x}(t+1)$ and the most recent previous position $\vec{x}(t)$ determines the acceptance.

$$\text{Cost difference} \qquad \Delta C_t := \hat{C}_g(\vec{x}(t+1)) - C_g(\vec{x}(t)) \,. \tag{13}$$

If the suggested position $\vec{x}(t+1)$ is accepted for the node position $N_j(\vec{x}, L)$,

[i] The probability of making a move which may increase the cost is greater than zero. It is proportional to the temperature τ.

the next node N_{j+1} is considered by applying the same procedure; if the suggested position is not accepted, another suggestion $\vec{x}(t+2)$ will be selected and the testing procedure is iterated.

To also enable an acceptance of stages with higher energies or costs, the decision of acceptance of a suggested position is of a stochastic nature. If ΔC is greater than 0 ($\Delta C > 0$) the probability of acceptance should be lower (but not equal to 0) than if ΔC is lower than or equal to 0 ($\Delta C \leq 0$). The probability of acceptance p is realized by applying a sigmoid function (Fig. 11), which is dependent on the cost difference ΔC and the temperature τ (equal to the Bolzmann-function).

$$p(\Delta C_t, \tau) := 1 - \frac{1}{1 + e^{-\frac{\Delta C_t}{k\tau}}} \tag{14}$$

where t is the time value and k is a constant factor to control the relation between the cost difference ΔC and the temperature τ.

With decreasing temperatures $\Delta\tau$, the probability of acceptance p of states which have higher costs than the stage before is decreased.

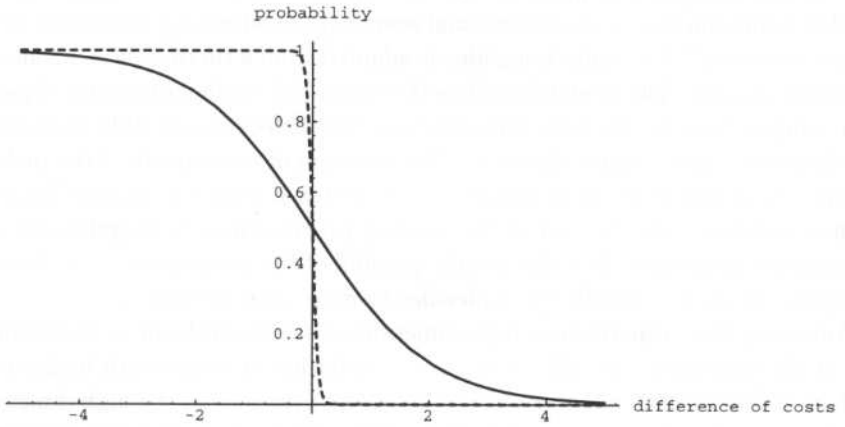


Fig. 11. Probability of acceptance p of a suggested position in relation to the difference of cost ΔC. For cost differences greater than 0, ($\Delta C > 0$) the probability of acceptance is not equal to 0 if the temperature is high enough. This ensures that during the local adaptation process even positions which would produce higher costs would be accepted (black line: temperature $\tau = 1$, dotted line: $\tau = 0.05$).

Furthermore, this selection and testing procedure is iterated additionally for a particular temperature step τ. During this iteration period, the temperature τ is kept fixed to find the optimal position of the nodes relative to each other. In the realization, discrete temperature steps are used and the change of one temperature step to the next lower one is processed if an equilibrium state is reached. This equilibrium can be determined either by the convergence of the time-dependent cost function or can be approximated by a number of discrete matching steps.[j] The

[j] In the realization, a number of discrete matching steps (within a range of 100–200 steps) are chosen because the matching can be performed with reduced computation effort by maintaining simultaneously qualitatively comparable results.

equilibrium state indicates the decrease in the temperature τ. The reduction in the temperature τ during the simulated annealing process is modeled nonlinearly to achieve a smooth fading of the temperature[k] (Fig. 11) and is defined as follows.

$$\tau(t+1) := q \cdot \tau(t), \quad \text{with} \quad 0 < q < 1 \tag{15}$$

The starting temperature $\tau(0)$ is determined empirically. It depends on the problem to be tackled and the scale of dependent value[s] to be optimized (cost functions C_g, C_n, and C_c) (see Fig. 12).

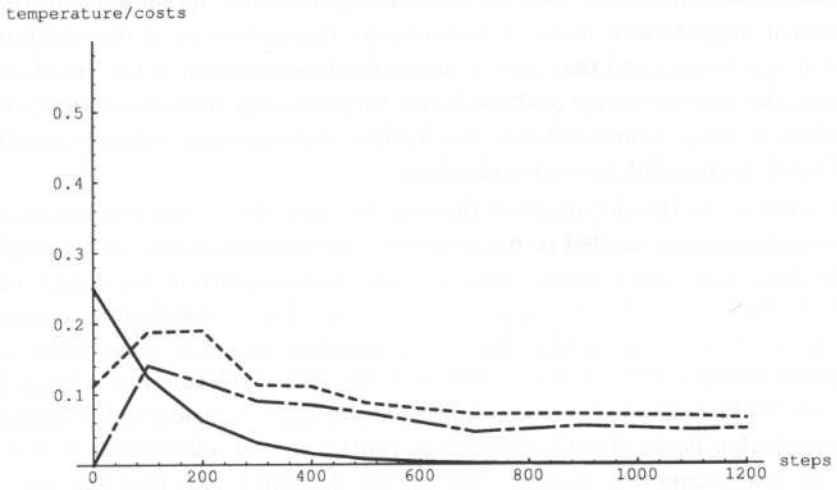


Fig. 12. Graph showing the relationship between the gradient of temperature, costs, and the number of steps. The running of the temperature τ (black line) together with two cost functions, total costs C_g (dotted line), and connection costs C_n (bar dotted line) shows the development of the optimization process. The temperature starts at step $t = 0$ with an initial value of $\tau(0)$ and decreases nonlinearly (discretely after each 100 steps). The costs increase drastically at the beginning due to the high temperature and the rapidly increasing cost component due to the distorted connecting edges. After a number of matching steps (ca. 300 steps) the matching process begins converging because a new spatial order emerges. The total costs and the costs due to the connecting edges decrease very slowly at a very low temperature until no further change occurs indicating the termination of the matching process (1200 steps).

After a total number of matching steps, within a range of 1000–6000 steps, the simulated annealing process converges on a local minimum. It is assumed that the computed minimum is a good estimate of the optimal solution. The total costs calculated for the best match are taken to verify the image regions of interest.

4.3.2. Restriction and modeling of the selection area

As mentioned before, within a well-defined surrounding region $\mathcal{R}$ of each node N_j of the considered image position $\vec{x}(t)$, one alternative image position $\vec{x}(t+1)$ is selected. The selection is not completely randomized because the 2-dimensional

[k] In the realization, values for $q = 0.5$ up to $q = 0.7$ have been found to provide good results.

random selection process is additionally modeled by a probability function. The idea behind the probability modeling of the randomized selection is that even at high temperatures a controlled distortion of parts of the graph should be performed. By applying a 2-dimensional Gaussian function, positions very close to the recent starting point $\vec{x}(t)$ will be selected with a higher probability than locations which are further away. The disadvantage is that this constraint may slightly restrict the flexibility and even the speed of the computation of the local graph-matching process.

To optimize the computation of worthwhile candidates for new positions, a $n \times n$ surrounding area $\mathcal{R}$ of each node N_j is defined upon which the subsequent 'random' selection of suggestions is based. The reason for this restriction of the selection area is that it can be assumed that after a successful determination of the initial starting position, the optimal image position is not very far away from the actual position. Therefore, a range which restricts the furthest distance may enhance an efficient selection of meaningful position candidates.

In addition to the definition of the cost function due to the connecting edges, other constraints are needed to maintain the spatial relationship of the graph during the local matching process. This is particularly important for object adapted graphs as they are used in this application because of their non-homogeneous topology. To avoid the situation that during the matching processes a particular node is positioned where it will cause an inverted neighborhood relationship, several limitations have to be defined. For example, a node which lies centrally and is surrounded by several other nodes should maintain its central spatial relationship in each case. With no constraints it is possible that during the graph matching the node N_i is moved to the left and subsequently its neighboring node N_j lying to its left is moved to the right so that the spatial relationship changes (node N_j changing from a neighbor on the left-hand side to a neighbor on the right-hand side). This may cause unwanted distortions and additional intersections of the connecting edges with other neighbors. Therefore, the area $\mathcal{R}$ in which a node may be moved is additionally restricted by the relative positions of its direct (connected) neighbors.

These constraints avoid additional intersections of connecting edges or generally undesired distortions of the graph but they also restrict the 'flexibility' of the graph-matching processes. In contrast to the approach introduced in Ref. 16 in which the high dimensional optimization problem is solved applying a testing scheme comparable to a simulated annealing approach using only the temperature $\tau = 0$, the approach presented here is able to efficiently produce a controlled distortion of the graph and reliable matching results. This is only made possible by these additional constraints.

Some results of the local-matching algorithm are demonstrated in Fig. 13. Two examples of eye regions are shown in which the local matching has optimally adapted the model graph to the underlying structure (compare Figs. 7 and 13, first row each). Furthermore, two examples are shown in which the graph is not adapted to an eye but to another facial structure (Fig. 13, second row). The graph is visibly more distorted than the adapted graphs of the eyes. This will generate higher costs

due to more distorted connecting edges. The total costs after completion of the local matching procedure are significantly higher for non-eye regions than for eye regions. The main reason is that the cost component due to the connecting edges is significantly increased while the cost component due to the nodes converges after the local matching at nearly the same level.

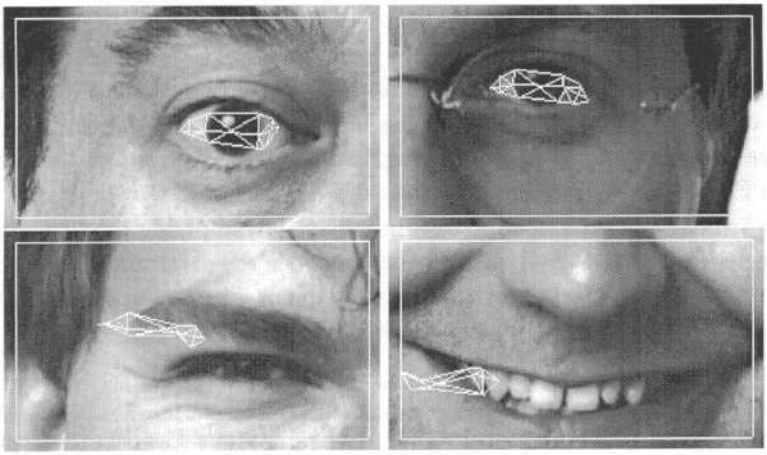


Fig. 13. Results of the local-matching algorithm. The adaptation of the model graph to the individual structures is shown for the images from Fig. 7. For the eye regions the local adaptation enhances the similarity between the several point representations so that the total costs C_g will decrease drastically (first row) by maintaining the global spatial relations between the nodes. For non-eye regions the spatial relations will be distorted more severely which causes high costs for the connecting edges C_c. During the adaptation of non-eye regions the final costs will also decrease but they will converge at a much higher level (second row).

5. CLASSIFICATION RESULTS

The graph-matching approach as it has just been introduced, has been tested on a database of 36 facial image regions.[l] The test set was divided into three subsets:

1. Class I, subset of eye regions satisfying the preconditions[m] ($n = 13$),
2. Class II, subset of eye regions not exactly satisfying the preconditions[m] ($n = 9$),
3. Class III, subset of non-eye regions ($n = 14$).

The application of the object adapted graph requires that different preconditions have to be fulfilled by the examples of the images to be classified. To derive or test the limitations of the approach presented here, several images not exactly fulfilling the preconditions have been included in the dataset. These are for example, screwed eyes or eyes with heavy shadows so that the characteristic information which has

[l]The set of test samples used is independent of the data which are used to establish the reference graph $\mathcal{G}^R$ introduced in Sec. 3.

[m]The exact definition of the preconditions is beyond the scope of this paper but we refer the interested reader to Ref. 13. However, the main aspects or underlying ideas of the definition used will follow in the subsequent text.

If the model graph had been successfully superimposed on the eye region of a member of subclass I or II, this member remained in its class. If the determination of the initial starting position failed, the region was added to the non-eye region class III.

The local matching was computed for all images and the classification was performed based on the set of examples. Fifteen of sixteen eye regions with successfully superimposed model graphs were verified correctly. It is remarkable that even 4 eyes not fulfilling completely the preconditions (members of subset II) were also verified successfully as eyes. This gave a discrimination performance of 93.75%. Only one eye region of subset II could not be recognized correctly. Furthermore, no non-eye region was misclassified as an eye region.

After the reallocation and the calculation of the local matching together with the classification, a sensitivity of 100% and a specificity of 95.5% were obtained.

The qualitative properties, sensitivity and specificity, are understood in this context as follows: Sensitivity is the proportion of eye regions which are correctly classified as eye regions. That is, all classified eye regions are eyes in reality and are successfully determined during the calculation of the initial starting position ($100 - f_a$; f_a = false-acceptance-rate) and that no non-eye region is misclassified as an eye. Specificity is understood as the proportion of non-eyes which are not classified as eyes that means the proportion of regions which are not classified as eyes and are not eyes in reality ($100 - f_r$; f_r = false-rejection-rate). A graphical summary of the classification results is presented in Fig. 14.

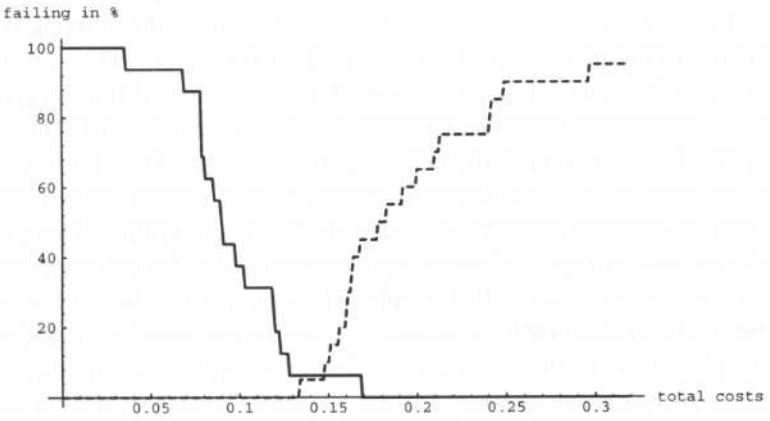


Fig. 14. Success of the classification of the reallocated test set dependent on the choice of the discrimination threshold. Between the false-rejection-rate (subset of eye regions) (black line) and the false-acceptance-rate (subset of non-eye regions) (dotted line) the best discrimination can be achieved if the threshold is set equal to 0.13. In this case only one eye region was misclassified and no non-eye region was classified as an eye region.

starting position fails on an eye region the subsequent processing will run as on a non-eye region and therefore, will also result in failure.

would have to be solved as discussed in the section 'local matching' (Sec. 4.3). A heuristic solution of this high dimensional problem can only be achieved if a good initial starting position can be assumed. This leads to a vicious circle and illustrates that an enhancement of the performance in the determination of the initial starting position is very difficult to achieve.

One possible solution to this problem could be to integrate several components of the degrees of freedom into even the first processing step. Some preliminary investigations have been made with different sizes of the model graph and differently oriented graphs (compare Figs. 8 and 9). The same would be possible for independently scaled graphs. For example, screwed eyes may be better detected during the first processing step if the size of the graph is varied in Y-direction while the size in X-direction is kept fixed. But such additional global distortions of the entire graph will very quickly increase the dimensions of the search space, even for the determination of the initial starting position so that a complete generation of the optimal solution may rapidly become impossible.

To summarize, the quality of the entire classification module is mostly dependent on the performance of the determination of the initial starting position. If the determination of the initial starting position fails for an image part from the class of interest the subsequent local matching process is not able to correct this failure. Therefore, additional effort has to be directed towards an enhancement of this component of the classification module. The subsequent local-matching algorithm succeeds in each case. That means that in the case of a successful determination of the initial starting position the subsequent adaptation is able to produce the necessary discrimination between an object belonging to the class of interest and an object not belonging to the class.

ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of Dirk Müller, a graduate student at the Technical University of Munich. We are also grateful to Dr. M. Michaelis from the Medis-Institute at the GSF Research Center for providing the filtering scheme and for many helpful discussions. The work has been partially supported by DFG grants So 320/1-2, Ei 322/1-2, and He 2967/1-1.

REFERENCES

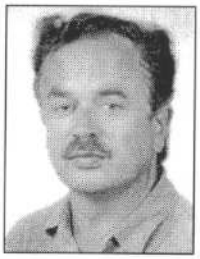
1. H. Bouattour, F. F. Soulie and E. Viennet, "Solving the human face recognition task using neural nets", *Proc. Artificial Neural Networks* **2**, Elsevier, 1992, pp. 1595–1598.
2. J. Bruske and G. Sommer, "Dynamic cell structure learns perfectly topology preserving map", *Neural Comput.* **7**, 4 (1995) 845–865.
3. J. Buhmann, J. Lange and C.v.d. Malsburg, "Distortion invariant object recognition by matching hierarchically labeled graphs", *Proc. Int. Joint Conf. Neural Networks*, IJCNN, IEEE, 1989, pp. 155–159.
4. G. W. Cottrell and M. Fleming, "Face recognition using unsupervised feature extraction", *Proc. Int. Neural Network Conf.*, 1990, pp. 322–325.
5. A. Fuchs and H. Haken, "Pattern recognition and associative memory as dynamical process in a synergetic system", *Biol. Cybern.* **60** (1988) 17–22 and 107–109.

6. B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, "Sexnet: A neural network identifies sex from human faces", *Adv. Neural Inf. Proc. Syst.* **3**, ed. R. P. Lippmann (1990) 572–577.
7. R. Herpers *et al.*, "GAZE: An attentional processing strategy to detect and analyze the prominent facial regions", *Proc. Int. Workshop on Autom. Face- and Gesture-Rec.*, ed. M. Bichsel, Zurich, Switzerland, 1995, pp. 214–220.
8. R. Herpers, M. Michaelis, K.H. Lichtenauer and G. Sommer, "Edge and keypoint detection in facial regions", *Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition*, IEEE Comput. Soc. Press, Killington, Vermont, 1996, pp. 212–217.
9. R. Herpers, M. Michaelis, L. Witta and G. Sommer, "Context-based detection of keypoints and features in eye regions", *Proc. 13th Int. Conf. Pattern Recognition*, IEEE Comput. Soc. Press, Vienna, Austria, Vol. B, 1996, pp. 23–28.
10. R. Herpers, L. Witta, J. Bruske and G. Sommer, "Evaluation of local image structures applying a DCS network", *Solving Engineering Problems with Neural Networks, Proc. 2nd Int. Conf. EANN96*, eds. A. B. Bulsari *et al.*, London, 1996, pp. 305–312.
11. R. Herpers, L. Witta, J. Bruske and G. Sommer, "Dynamic cell structures for the evaluation of keypoints in facial images", *Int. J. Neural Syst., Special Issue on Neural Networks for Comput. Vision Appl.* **8**, 1 (1997) 27–39.
12. R. Herpers, D. Müller, M. Michaelis and G. Sommer, "Invariant classification of image parts using a dynamic grid of point representations", *Neural Networks in Engineering Systems, Proc. Int. Conf. EANN 1997*, eds. A. B. Bulsari *et al.*, Stockholm, 1997, pp. 41–44.
13. R. Herpers, "GAZE: A common attentive processing strategy for the detection and investigation of salient image regions", PhD Thesis, Christian-Albrechts-Universität, D-24105 Kiel, Germany, 5/1997 and Tech.-Rep. GSF-Bericht 16/97, D-85764 Oberschleißheim, Germany, 1997.
14. S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, "Optimization by simulated annealing", *Science* **220** (1983) 671–680.
15. M. Kosugi, "Robust identification of human face using mosaic pattern and bpn", Article No. 0-7803-0557-4/92, 1992.
16. M. Lades *et al.*, "Distortion invariant object recognition in the dynamic link architecture", *IEEE Trans. Comput.* **42** (1993) 300–311.
17. A. J. Luckmann and N. M. Allinson, "A multiple resolution facial feature location network", *Proc. Sec. Int. Conf. Visual Search*, Durham, 1990, pp. 1–11.
18. M. Michaelis, "Low level image processing using steerable filters", PhD Thesis, Christian-Albrechts-Universität, D-24105 Kiel, Germany, 1995.
19. J. L. Perry and J. M. Carney, "Human face recognition using multilayer perceptron", *Proc. Int. Conf. Neural Networks* **2**, 1990, pp. 413–416.
20. T. Poggio and F. Girosi, "Networks for approximation and learning", *Proc. IEEE* **78**, 9, 1990, pp. 1481–1497.
21. L. Wiskott, J. M. Fellous, N. Krüger and C. v. Malsburg, "Face recognition and gender determination", *Proc. Int. Workshop Autom. Face- and Gesture-Rec.*, ed. M. Bichsel, Zurich, Switzerland, 1995, pp. 92–97.
22. A. L. Yarbus, *Eye Movements and Vision*, Plenum Press, New York, 1967.

Rainer Herpers received a diploma degree in computer science at the RWTH Aachen in 1991 and the PhD degree in electrical engineering from the University of Kiel in 1997. Since 1992 he has been a member of the Institute of Medical Informatics and Health Services Research at the GSF-Research Center for Environment and Health and of the Computer Science Institute at the Christian-Albrechts University of Kiel, Cognitive Systems Group.

Dr. Herpers is currently working on a two year DFG grant as visiting research scientist at the Vision Group of the Department of Computer Science of the University of Toronto, Canada.

His research interests include image analysis, face processing, and applications of neural networks methods.



Gerald Sommer received a diploma in 1969 and a Dr. rer. nat. degree in 1975 in physics from Friedrich-Schiller University, Jena. He received a Dr.-Ing. habil. degree in 1988 from the Technical University Ilmenau. He worked from 1969–1991 in the physics, medicine, technology and mathematics departments at the University of Jena. From 1991–1993 he was head of the division of medical image processing at GSF-MEDIS, Neuherberg. In February 1993, he joined the Computer Science Institute of Christian-Albrechts University, Kiel, as Professor of cognitive systems.

His current research interests include the behavior based design of autonomous systems.