An Attentive Processing Strategy for the Analysis of Facial Features*

R. Herpers[1,2] and G. Sommer[1]

[1] Institut für Informatik, Lehrstuhl für Kognitive Systeme, Universität Kiel
Preußerstr. 1-9, 24105 Kiel, Germany, Email: herpers@gsf.de
[2] GSF–Institut für Medizinische Informatik und Systemforschung, MEDIS
Ingolstädter Landstr. 1, 85764 Oberschleißheim, Germany

Abstract. Facial landmarks such as eye corners, mouth corners or nose edges are important features for many applications in face recognition. The exact detection of these landmarks, however, is not an easy task because of the high individual variability of facial images and therefore, of the tremendous complexity of all the low-level features existing within the image. For instance, a precise and reliable detection of the eye corners has not been successfully solved until now. However, the knowledge of the exact position of these landmarks in the facial image is important for many matching and face processing tasks. For the classification and discrimination of dysmorphic facial signs a precise and reliable detection of a certain set of anatomical facial landmarks is particularly necessary. For this, an attentive processing strategy has been developed which puts the focus of the processing on only those salient image areas which are really needed to solve the several subtasks. The fundamental idea of the approach presented is to concentrate the artificial attention upon only a small fraction of the existing low-level features within a spatially well restricted image area.

Keywords. Attentive Vision, Face Recognition, Image Processing

1 Introduction

Motivated by the eye movement strategies of the human visual system, a computer-based attentional strategy has been developed to detect the prominent facial features in portrait images. The attentive processing system is structured into three main components (fig. 1):

- the *localization of salient regions*,
- the *classification of foveated regions* and
- the *structural analysis of foveated regions.*

In a first processing step, the most salient facial regions such as the eye, nose, and mouth region are localized, based on a saliency representation, which is established by deriving several attentive visual cues [2]. Subsequently, the detected and now spatially limited regions are classified to evaluate the benefit of applying more detailed and expensive analysis methods [6, 7]. Furthermore, a first semantic interpretation of the foveated region is processed.

*This work is partially supported by the DFG, Grants So 320\1-2 and Ei 322\1-2.

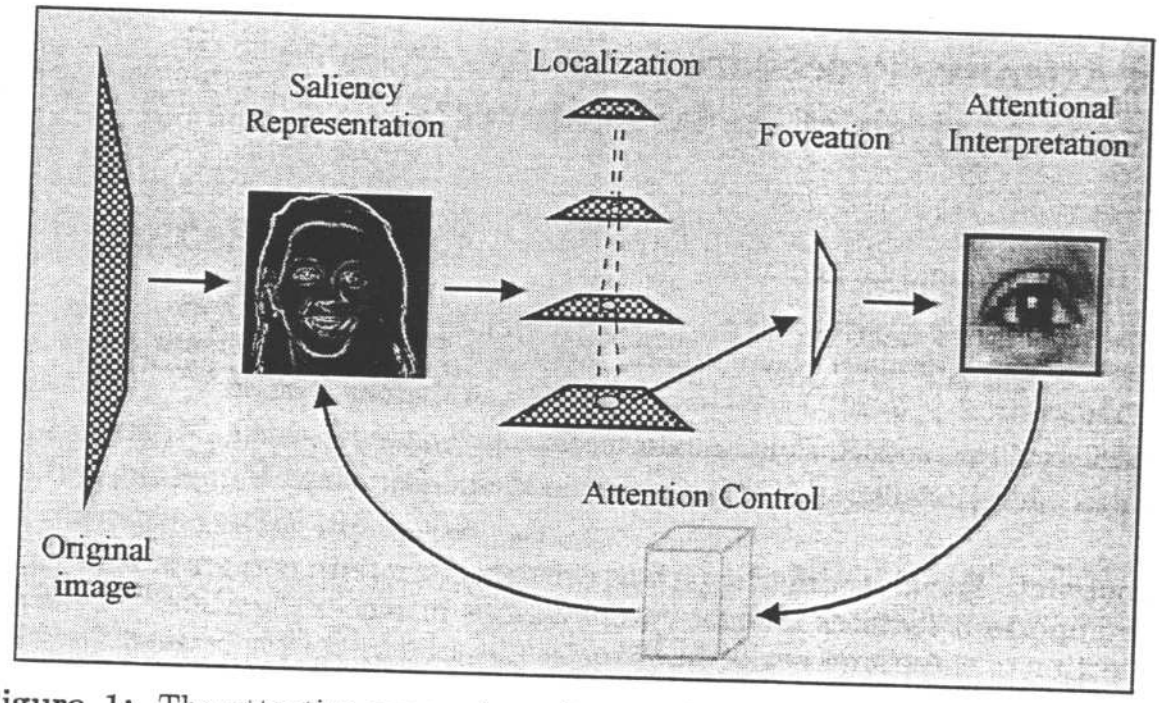


Figure 1: The attentive processing scheme. Attentive visual cues are derived from facial grey level images to build up a saliency representation. Prominent facial regions are selected and a decision is made to apply more detailed processing methods. The attentive processing is supervised by a control strategy essentially based on model knowledge and recently derived information. Applying this procedure, all salient image regions are located, selected, and analyzed in detail.

During the detailed analysis step the exact positions of relevant keypoints or anatomical landmarks such as eye corners and mouth corners are determined [3] and evaluated [5].

Fundamental to the approach presented is the common use of the processing principle: the consideration and evaluation of only salient image features. In other words, based on this processing strategy only those image features are processed and analyzed further for which evidence exist in the local arrangement of the image area considered. During the region localization only the most attentive cues are derived and represented. For the subsequent classification module an object representation is established which is composed of a small number of point representations located only at prominent image positions of the object such as prominent corners or intersection points. During the third processing step, where a structural analysis of the foveated facial region is performed, only prominent image structures such as characteristic edge and line segments are detected and considered in more detail applying stepwise adaptive detection and tracking methods.

A high degree of efficiency is achieved using an uniform or common comprehensive filtering scheme during all processing steps. While in the first processing step only simple features are required, in the subsequent processing steps more detailed information is needed. The filtering scheme developed is incorporated into a unified processing strategy based on a common attentive framework [12].

The attentional processing mechanisms developed are used as an essential component in building an image processing system to classify facial images of children with dysmorphic signs. Dysmorphic signs in facial images are minor anomalies which, by definition, do not lead to functional disturbances [13, 14] (fig. 2). In this context, dysmorphic signs in faces play an important role in syndrome identification because in most cases they are visible and prominent in the facial image. It has been shown, that particular phenotypic combinations are typical for distinct dysmorphic syndromes [13, 14]. Therefore, the detection of particular keypoint positions (fig. 2) in facial images is of high diagnostic value. For this, the localization of the keypoints or landmarks should be very accurate, reproducible, and should correspond to the anatomical definition of the keypoint positions.

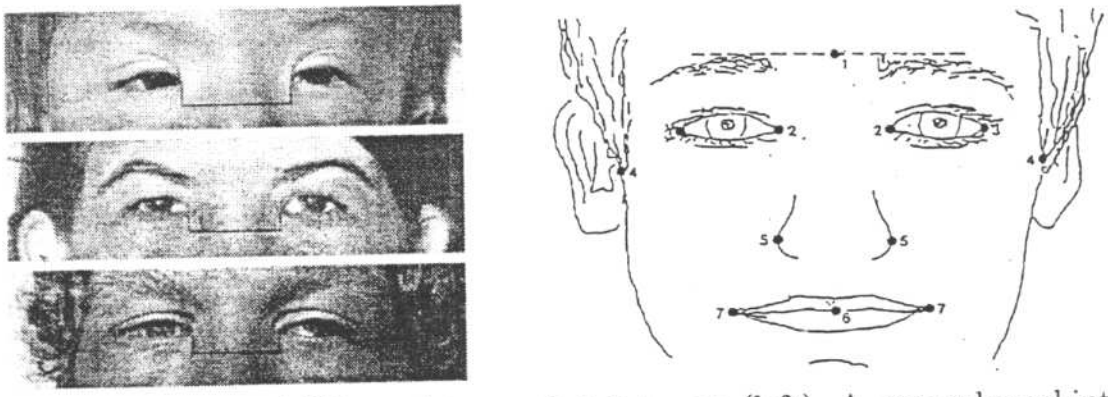


Figure 2: Example of different intercanthal distances (left). A very enlarged intercanthal distance and an epicanthus are typical examples of dysmorphic signs in the frontal facial image (first eye strip) (from [14, p. 42]). Important anthropometrical landmarks in a frontal face image (right).

2 Localization of salient regions

In general, the bottom-up visual search task in real world images is an NP-complete problem depending on the image size and the number of objects to be searched. However, a task-directed search based on selective attentional mechanisms can be computed in linear time complexity which depends on the number of objects to be searched in the image [15]. These complexity considerations suggest that attentional mechanisms are necessary to successfully solve an image analysis problem in real-time. Thus, image processing with attention control reduces computational load and focuses the computation to only those image regions which are important to solve the current task.

In the HVS elementary visual cues such as motion, color, orientation, edge information etc. are derived from the sensed input. These features establish an attentive representation upon which the control of the visual attention is based. The evaluation of this representation provides different salient image regions which are foveated in a sequential order. By adopting these gaze control principles for developing artificial image processing systems, a complex image analysis problem P is reduced to a number of subproblems P_i that can be solved individually in a simpler way [8]:

$$P(I, M) := \{P_1(I_1, M_1), \ \ldots \ , P_n(I_n, M_n)\} \tag{1}$$

where $n = 2, \ldots, N$ is the number of subproblems, $I_i \subseteq I$ are image parts of the image I and $M_i \subseteq M$ are model assumptions and other parameters. In other words, only that information, which is really essential for a given task, has to be extracted and processed further. Irrelevant image regions may be excluded as far as possible to avoid additional processing effort. The processing of several subproblems P_i is a dynamic process based on the results obtained from the subproblems processed before (fig. 1). The analysis of the results obtained and the integration of top-down knowledge may be used to determine the next subproblem and methods to solve it.

Saliency representation

The attentive processing strategy is based on a saliency representation $S(t)$ carrying all the information which is necessary to compute the selective attention (fig. 1). In the realization this saliency representation is a 2D saliency map $S(t)$ in which the spatial distribution of the salient cues from the underlying image is encoded (fig. 3). The saliency representation $S(t)$ is generated from a 'feature representation' $U(I)$ and a time dependent 'control representation' $C(t)$; $S(t) := U(I) \times C(t)$. The feature representation $U(I)$ is defined as the weighted sum of several filters f_l applied to the image $I(\vec{x})$:

$$U(I) := \sum_l a_l \ (f_l \otimes I(\vec{x})) \tag{2}$$

The control representation $C(t)$ is needed to control the attentive processing. It is defined as $C(t) := SR(t) \times A(t)$. The first component of the control map $C(t)$ is the suspension map $SR(t)$ generated to suspend already foveated and analyzed regions from further processing steps and the second component is the anticipation map $A(t)$.

To enable the computation of additional salient regions, previously selected regions have to be suspended from the subsequent processing. Therefore, a locally parameterized 2D Gaussian function, called 'suspension functions', is calculated at the center of the already detected region. The suspension of the already detected regions is represented by a suspension map $SR(t)$ (see the darkened right eye in fig. 3c) (for more details see [2, 8]).

Control representation

The control representation $C(T)$ also enables the integration of top-down information into the low-level analysis process. The knowledge integration is realized by slightly emphasizing the derived salient cues of those image regions which are intended to be localized in the next localization steps. Therefore, a spatially restricted 2D Gaussian function is calculated for each expected region and encoded in the anticipation map $A(t)$ (see the left eye in fig. 3c).

The content of the saliency representation $S(t)$ can be summarized in two parts, one part contains the regular attentive or salient cues of the underlying image and therefore, it is fixed for all following processing steps. This feature representation $U(I)$ can be processed in advance applying fast and parallel convolution methods. The second part of the saliency representation $S(t)$ the control representation $C(t)$ influences the selection of the salient features during the processing of the several attentive regions and, therefore, is time variant. It can be viewed as an integration process or an instrument for decreasing the saliency of regions which are already detected and for increasing the saliency of regions which are intended to be localized based on model assumptions and a priori knowledge.

Evaluation of the scale hierarchy

The saliency representation $S(t)$ is represented in a scale hierarchy to provide a high degree of invariance and robustness during the evaluation of the spatially distributed salient image cues. It forms the basis of all subsequent localization steps. The selection of a salient facial region starts at the coarsest scale of the hierarchy, employing a maximum search algorithm (fig. 3a). Subsequently, this initial salient map element is expanded dependent on the underlying local saliency distribution. For this, a 2D Gaussian function is fitted to the local saliency distribution (fig. 3b) and a certain contour line h is taken to fix the boundary of an elliptical region given by the following ellipse equation:

$$(\vec{x} - \bar{m})^T Cov^{-1} (\vec{x} - \bar{m}) = h = const. \tag{3}$$

where $\bar{m}$ is the expectation of the 2D Gaussian function and Cov is the covariance matrix of all map elements or pixels of the localized region.

(a) (b) (c)

Figure 3: Processing of the region localization by evaluating a multi-scale representation of the saliency map. Maximum search of that element with the highest saliency (a). The maximum element is expanded optimally to the local extension of the feature representation at each scale (figs. inverted). Subsequently the computed region is projected to the next higher resolution level (b). To compute further attentive regions, regions already considered are suspended from the following localization steps (right eye in (c)). In addition, the integration of model knowledge enables an anticipation of that region which is to be detected (left eye in (c)).

After an iteration step to optimally match the ellipse to the underlying saliency representation the computed region is projected to the next higher resolution level (fig. 3b) and the region adaptation is computed again until the highest resolution level is reached. The computed region is selected at the highest resolution level and it is well adapted to the extent and the orientation of the underlying local saliency distribution (fig. 4). To enable the localization of further facial regions all previously detected regions have to be suspended from the following localizations steps. For this the suspension representation $SR(t)$ is introduced as mentioned before (fig. 3c). Applying this procedure, all salient facial image regions can be located and further processing modules can be applied to them.

Results of the region localization

The region localization has been tested on more than 100 frontal facial images with slightly different illumination conditions and camera positions. Some faces were tilted or slightly rotated in different directions. The scales and the brightness of the facial images recorded also differed from image to image (fig. 4). Within the first 3 foveation steps both eye regions were detected correctly in more than 98% of the face images considered.

Figure 4: Results of the attentive region localization. The detected regions are spatially well adapted to scale, extent and orientation of the facial regions. The attentive localization algorithm was also applied successfully to face images of the face database of A. Pentland[1] although the resolution of the images was too low to apply all subsequent processing modules (right).

3 Region classification

After the attentive localization of the different prominent facial regions, it has to be decided whether the foveated region should be investigated in more detail or not. The subsequent detailed analysis methods (described in chapter 4) are very precise and specially developed for the particular facial part to be investigated. Since the eyes contain the most attentive cues in the facial image they are selected during early foveation steps [8]. Assuming that just one eye region is located after three foveation steps the complexity of the

[1] available at ftp: vismod.www.media.mit.edu/pub/images

general classification task can be reduced to the verification of the question "does the foveated region contain an eye region ?". After the successful classification of one eye region further processing steps can be chosen and model knowledge about the spatial relations between the facial regions can be integrated.

Point representations

To make this decision a neural classifier has been developed which is motivated by the dynamic link architecture introduced by the Malsburg group [1, 10]. In contrast to their work, where initially particular subjects are identified, we have to compute a discrimination task which distinguishes between different classes of facial regions. Therefore, besides the intraindividual also the interindividual variability has to be considered during the construction of the uniform classifier. For the recognition of an eye in a facial region a set of point descriptions is established. These point descriptions are located at particular characteristic image positions (e.g. significant corners or intersection points) (see fig. 5).

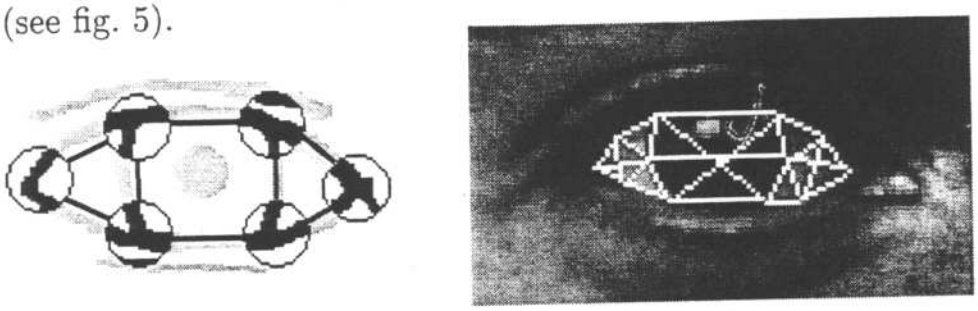


Figure 5: Point descriptions connected by a graph demonstrated for an artificial eye region (left). The descriptions at the circled keypoints and the spatial relationship between them are used to clearly represent the characteristic features of the considered class of region. Model graph used for eye regions (right).

The spatial relationship between the several image points is maintained by establishing a 2D graph consisting of nodes for the several point descriptions and of edges for the connections between them. In contrast to the related work, object adapted graphs are applied. In detail, the eye pattern is represented by 12 point descriptions positioned by the nodes of the graph at characteristic object positions or landmarks (fig. 5). Fundamental to this design is that the nodes are positioned only at image positions for which evidence of the local structure exists. To ensure a reliable representation of the underlying image structure at the characteristic image positions orientation selective edge and line detection filters as well as polar separable filters all based on Gaussian derivatives are applied [6, 7].

Graph matching

By applying a two step graph-matching algorithm different facial regions can be distinguished very reliably. In the first processing step the initial starting position is computed to fit the undistored model graph to the most appropriate position. In the second step the graph is distorted and the nodes are moved independently by applying a simulated annealing approach.

For the determination of the initial starting position, the similarity of the graph is calculated for all image positions in the considered image part and that image position which matches best indicates the initial starting position. During the computation of the best starting position the shape of the graph is kept rigid (fig. 6). No distortion of the graph is allowed, only the scale and orientation of the graph is varied, to achieve better invariance properties.

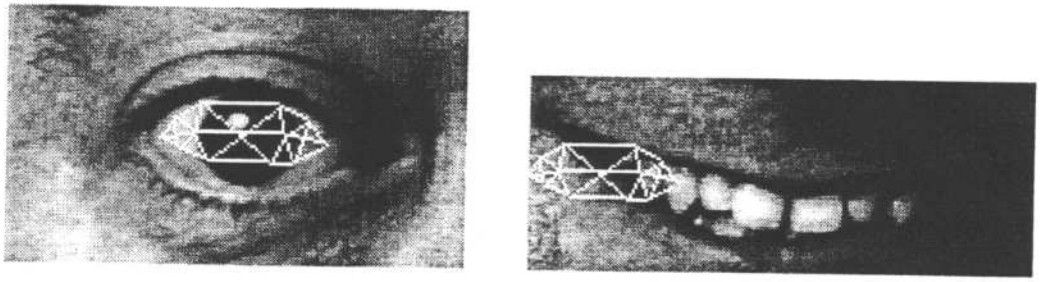


Figure 6: Results of the calculation of the initial starting position of the model graph for an eye region (left) and a non eye region (right). The calculation will terminate at that position with the greatest similarity to the represented edge information of the model graph.

After the determination of the initial starting position the model graph has to be adapted independently of its topology to achieve an optimal correspondence for each constituent node. The solution of such a high dimensional problem, considering graph sizes with 12 nodes, cannot be calculated completely. 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 [9] is employed to compute the best match of the model graph to the test patterns (for more details we refer the reader to [6, 7]).

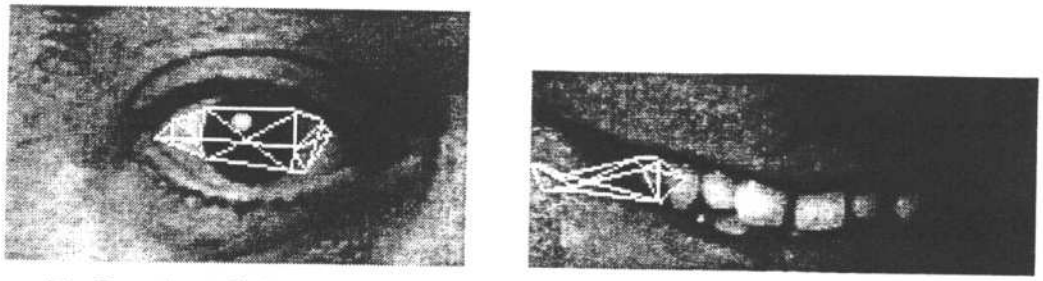


Figure 7: Results of the simulated annealing algorithm demonstrated for an eye region (left) and a non eye region (right).

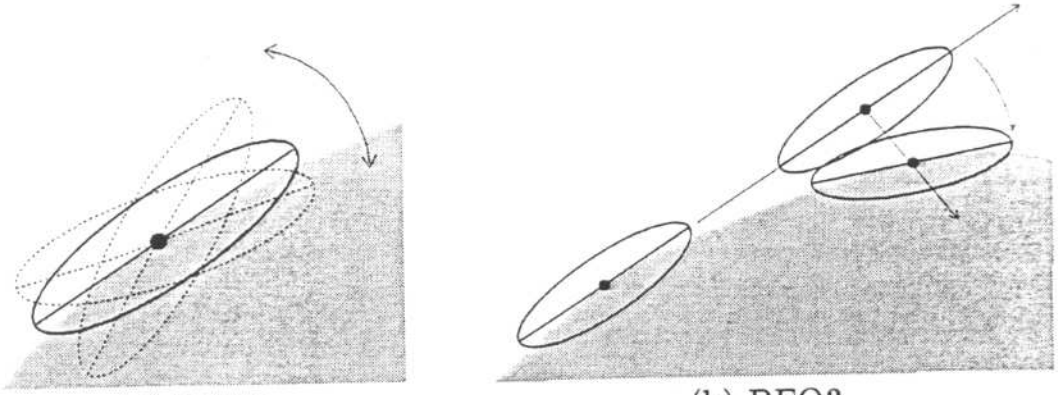
Regarding eye regions the adaptation of each node to the corresponding image structure enhances the similarity to the model representation by maintaining simultaneously the spatial relationship between the nodes. For non eye regions the spatial relations will be distorted more severely, which causes higher distortions of the connecting edges (fig. 7). By applying a cost function this property can be used to distinguish reliably between facial regions. With this classification module all eye regions can be classified successfully as eye regions (sensitivity of 100%). For non eye regions the classification performance is not quite as good (95.5%).

4 Structural analysis

In the third processing step of the attentive processing strategy the foveated regions are analyzed in more detail. In the following we will take the eye region as an example, but figure 10 shows that characteristic facial features and anatomical landmark positions may also be successfully computed for other facial regions. The approach starts by detecting the most prominent and reliable features in the facial region considered. Given our task and image recording conditions these are the strong vertical step edges of the iris in eye regions. Subsequently, the complete iris and the eyelids are tracked to finally detect the eye corners. At each step the detection and tracking is controlled by integrated model knowledge [3, 4]. Additionally the assumptions are used to check the consistency of previously detected edges and to predict the edge structures searched for in the next processing step.

Basic filter operations

A very flexible filtering scheme is applied which produces many different low-level features thus providing an expressive data-driven basis for the keypoint detection [3, 4]. The detection of the image structures in the facial regions is based on a sequential search and tracking of the characteristic edges and line segments. The detection and tracking is realized by three different **basic filter operations** (fig. 8) that make extensive use of steerable edge and line detection filters which are used several times during the processing [11].



(a) BFO2 (b) BFO3

Figure 8: Examples of the basis filter operations. BFO2: Determination of the orientation (a). BFO3: Stepwise tracking (b).

Integration of model knowledge

The approach presented for the keypoint detection essentially uses model knowledge to establish a sequential search strategy for a stepwise detection of the main characteristic structures in the image part considered. The derivation of a large number of features and the integration of model knowledge are mutually dependent.

On the one hand, the large number of features that are derivable by the filtering scheme does not allow a 'classical' computational detection strategy.

Therefore, a sequential strategy is introduced that integrates model knowledge into the low-level analysis processes. The model knowledge provides the 'global overview' for the confusing wealth of features. We call such a strategy 'attentive' because it focuses only on those types of features, which may be present as dictated by the model, or in other words, which are expected to be present.

On the other hand, an attentive processing requires a flexible and probably complete representation of all included features. Therefore, the filtering scheme must allow for an efficient on-line choice of the type and the quality of the features needed. In other words, there is a potential need for many features, but during the processing only a small fraction of all the features is required to solve the current problem. Therefore, more expensive and time consuming filtering is applied only to certain image positions where it is really worthwhile [11].

Sequential search strategy

The detection of the keypoints in the facial regions is achieved by a sequential search or tracking of the edge and line structures where each step consists of several applications of the basic filter operations (fig. 9). The selection of the different operations and their parameters for each step is controlled by the already derived information together with the model knowledge. The model knowledge used consists of the relevant edge and line structures, their scales and geometrical relations of the considered facial region. This information is used to search for reliable edges at specific positions, orientations, and scales in each processing step. Furthermore, different kinds of edges (white-to-black versus black-to-white edges or edges against lines etc.) can be distinguished very reliably.

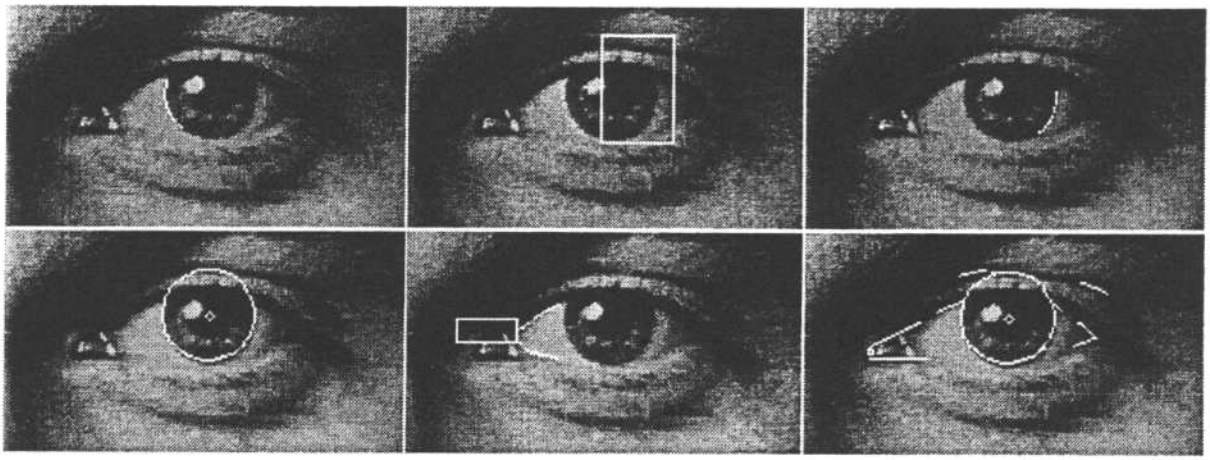


Figure 9: Example steps of the sequential search strategy for the detection of keypoints and prominent structures in an example eye region. First, a prominent vertical bright-to-dark edge is detected (first row). After the detection of the corresponding right edge segment, the final segmentation of the iris is computed. The eyelid edges are searched, tracked, and finally that edge segment which is strongly curved is detected to determine the eye corners (second row).

5 Conclusion

Fundamental to the attentive processing strategy presented is that all processing modules consider and process only prominent and really characteristic image structures. Only those image structures are considered which are relevant to the following processing task. Beginning with the region localization these are distributed edge and line structures. In the following processing steps more detailed structures need to be detected and therefore, more sensitive and expensive methods are applied locally.

The presented approach combines a cyclic as well as a hierarchical procedure which is essentially supported by appropriate model knowledge. The proposed strategy realizes an efficient integration of high-level information into mainly low-level processes.

The attentive processing strategy presented has a notably high degree of scale invariance, achieved using different resolution levels adapted to the requirements of the particular processing module and to the task to be solved. The detection and analysis of the prominent facial regions is independent of the exact position and the orientation of the face and the facial components within the image. The search algorithms proposed are able to cope with different variations due to the illumination, brightness, and contrast of the studied facial images.

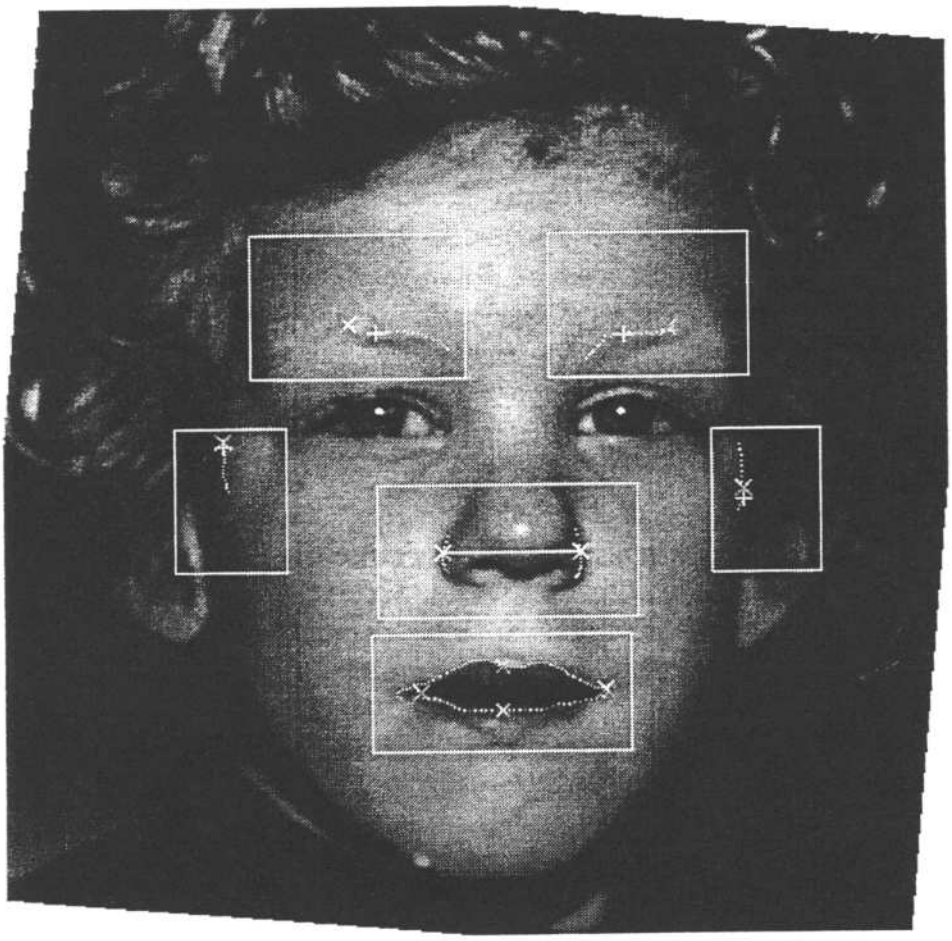


Figure 10: Result of the structural analysis demonstrated for the other facial regions. All important anthropometrical landmarks which are relevant to our medical application (see fig. 2) have been detected.

Acknowledgments

We appreciate the support of Prof. S. Stengel–Rutkowski of the Institute for Social Paediatrics and Youth Medicine. The work is partially supported by the DFG grants So 320/1-2 and Ei 322/1-2.

References

[1] J. Buhmann et al., *Distortion invariant object recognition by matching hierarchically labeled graphs*, Proc. Int. Joint Conf. on Neural Networks, IJCNN, IEEE, pp. 155-159, 1989.

[2] R. Herpers et al., *GAZE: An attentional processing strategy to detect and analyze the prominent facial regions*, In: Proc. of the Int. Workshop on Autom. Face- and Gesture-Rec., Zurich, M. Bichsel (ed.), pp. 214-220, 1995.

[3] R. Herpers et al., *Edge and keypoint detection in facial regions*, In: Proc. of the 2. Int. Conf. on Automatic Face and Gesture Recognition, Killington, IEEE, pp. 212-217, 1997.

[4] R. Herpers, et al., *Context Based Detection of Keypoints and Features in Eye Regions*, In: Proc. of the 13th int. Conf. on Pattern Recognition, 13th. ICPR, Vienna, Austria, IEEE Computer Society Press, Vol. B, pp. 23-28, 1996.

[5] R. Herpers et al., *Dynamic cell structures for the evaluation of keypoints in facial images*, In: Int. Journal of Neural Systems, Special Issue Neural Networks for Computer Vision Applications, pp. 27-39, 1997.

[6] R. Herpers et al., *Invariant classification of image parts using a dynamic grid of point representations*, In: Proc. of 3rd int. Conf. on Engineering Applications on Neural Networks, EANN'97, Neural Networks in Engineering Systems, A.B. Bulsari et al. (eds.), pp. 41-45, 1997.

[7] R. Herpers et al., *Discrimination of facial regions based on dynamic grids of point representations*, to be published in Int. Journal of Pattern Recognition and Artificial Intelligence, Special Issue on "Neural Networks in Computer Vision Applications", 1998.

[8] R. Herpers, *GAZE: A common attentive processing strategy for the detection and investigation of salient image regions*, PhD thesis, Christian-Albrechts-Universität, Kiel, Germany, 1997.

[9] S. Kirkpatrick et al., *Optimization by simulated annealing*, Science, Vol. 220, pp. 671-680, 1983.

[10] M. Lades et al., *Distortion invariant object recognition in the dynamic link architecture*, IEEE Trans. on Computers, Vol. 42, pp. 300-311, 1993.

[11] M. Michaelis, *Low level image processing using steerable filters*, PhD thesis, Christian-Albrechts-Universität, Kiel, Germany, 1995.

[12] M. Michaelis et al., *A common framework for preattentive and attentive vision using steerable filters*, In: Proc. of the CAIP'95, Prague, V. Hlavac et al. (eds.), Springer-Verlag, LNCS 970, pp. 912-919, 1995.

[13] S. Stengel-Rutkowski et al., *Anthropometric definitions of dysmorphic facial signs*, In: Hum. Genet., Vol. 67, pp. 272-295, 1984.

[14] S. Stengel–Rutkowski et al., *Chromosomale und nicht-chromosomale Dysmorphiesyndrome*, Enke Verlag, Stuttgart, 1985.

[15] J.K. Tsotsos, *Analyzing vision at the complexity level*, Behavioral and Brain Sci., Vol. 13, pp. 423-469, 1990.