-> SMC '97 Proc

FACIAL ANALYSIS APPLYING AN ATTENTIVE PROCESSING STRATEGY

Rainer Herpers[1,2] Gerald Sommer[1]

[1] Christian-Albrechts-University, Cognitive Systems Group,
Preußerstr. 1–9, D-24105 Kiel, Germany
[2] GSF – Institute of Medical Informatics and Health Services Research,
Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany
(E-mail: herpers@gsf.de)

ABSTRACT

A technical realization of an attentional mechanism localizing and analyzing prominent facial regions in gray-level images is presented. By adopting the gaze control principles of the HVS for developing an image processing systems, a complex image analysis problem may be decomposed into a number of subproblems which can be solved step by step in a simpler way. The attentive processing strategy starts with the localization of the prominent facial regions based on a saliency representation carrying all the information needed to select and restrict the extend of the prominent facial regions. Subsequently, the detected and now spatially limited regions are classified to evaluate the benefit of applying more detailed analysis methods. During the detailed analysis step the exact positions of anatomical landmarks or keypoints such as eye and mouth corners are determined. Fundamental to the attentive processing strategy is that all processing modules consider and process only prominent and really characteristic image structures. Only those images structures are considered which contribute to the solution of the actual processing task. The attentive processing strategy proposed is able to cope with variations due to the perspective angle or pose, orientation, illumination, and contrast of the studied facial images.

1. INTRODUCTION

Motivated by the eye movement strategies of the human visual system, a computer-based attentional strategy has been developed in order to detect the prominent facial features in portrait images. The attentive processing system is structured into three main parts (fig. 3):

- 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 [4]. Subsequently, the detected and now spatially limited regions are classified to evaluate the benefit of applying more detailed and expensive analysis methods [6]. Furthermore, a first semantic interpretation of the foveated region is processed [7]. During the detailed analysis step the exact positions of relevant keypoints or anatomical landmarks such as eye corners and mouth corners are determined [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 in a saliency representation. For the subsequent classification module an object representation is established which is built of a small number of point representations located only at prominent image positions of the object (such as prominent corners or intersection points). This number of point representations has been connected to maintain the topological relationship. It has been shown that these representations are sufficient for a computation of a reliable classification of the selected facial regions. 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 detail applying stepwise several adaptive detection and tracking methods.

The high degree of efficiency of the processing strategy is achieved by the use of a common 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 embedded into a unified processing strategy based on a common attentive framework [11].

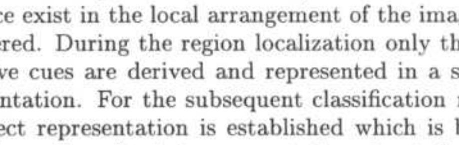


Figure 1. A very enlarged intercanthal distance and an epicanthus are examples of dysmorphic signs in the frontal facial image (from [12, p. 42]).

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 [12] (fig. 1). 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 [12]. 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 it should correspond to the anatomical definition of the keypoint positions.

This work is partially supported by the DFG grants: So 320/1-2 and Ei 322/1-2.

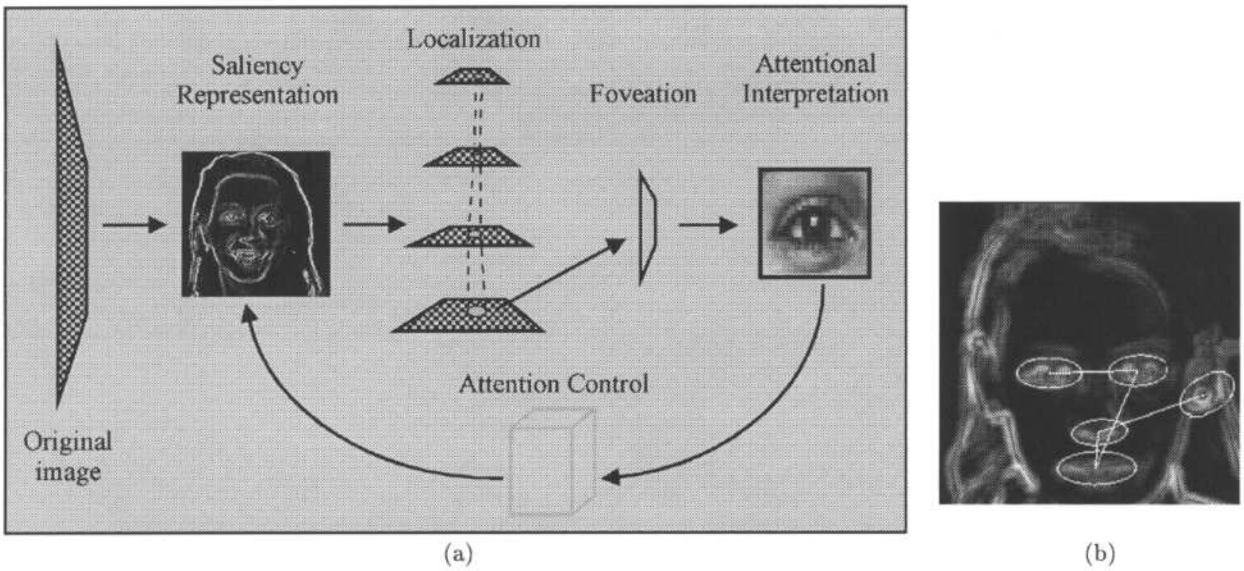


Figure 3. The attentive processing scheme (a). Attentive visual cues are derived from facial grey level images (b) to build up a saliency representation. Prominent facial regions are selected and a decision is derived to apply more detailed processing methods. The attentive processing is supervised by a control strategy essentially based on model knowledge and recently derived information. Applying this procedure, all salient facial image regions can be located, selected (b), and analyzed in detail.

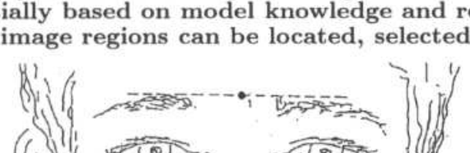


Figure 2. Important anthropometrical landmarks in a frontal face image.

2. ATTENTIVE PROCESSING

In general, the bottom-up visual search task using 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 only on the number of objects to be searched in the image [13]. 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.

In the HVS elementary visual cues such as motion, color, orientation, edge information etc. are derived from the sensed input data during the early stages of the preattentive processing. These features establish a preattentive representation upon which the control of the visual attention is based. The evaluation of this representation provides different salient image regions which may be foveated in a sequential order.

By adopting these gaze control principles for developing artificial image processing systems, a complex image analysis problem P may be decomposed into a number of subproblems P_i that can be solved individually in a simpler way [13].

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

where $n = 2, \dots, N$ 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 probably a dynamic process based on the results obtained from the subproblems processed before. The analysis of the results obtained and the integration of top-down knowledge may be used to determine the next subproblem and the methods to solve them.

2.1. 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. 3). 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. 3b and fig. 4). The saliency representation $S(t)$ is generated from a 'feature representation' $U(I)$ and a time dependent 'control representation' $C(t)$:

$$S(t) := U(I) \times C(t) \qquad (2)$$

The feature representation $U(I)$ is defined as the weighted sum of several filters f_l applied to the image $I(\vec{x})$:

$$U(I) := \sum_l \ a_l \ (f_l \otimes I(\vec{x})) \qquad (3)$$

In the application presented here it is advantageous to use local activity detectors with more blurred responses than those obtained from classical edge detectors [11] (fig. 3b).

The control representation $C(t)$ is needed to control the attentive processing. It is defined as:

$$C(t) := SR(t) \times A(t) \qquad (4)$$

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, locally parameterized 2D Gaussian functions, called the 'suspension functions', are calculated at the center of the already detected regions. The suspension of the already detected regions is represented by a suspension map $SR(t)$ (see the darkened left eye in fig. 4d) [4].

The control representation $C(T)$ enables also the integration of top-down information into the low-level analysis. The knowledge integration is realized by slightly emphasizing the derived salient cues of that 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 in the same way as during a suspension process and encoded in the anticipation map $A(t)$ (see the right eye in fig. 4d).

The content of the saliency representation $S(t)$ can be summarized in two parts, one part resembles the regular attentive or salient cues of the underlying image and therefore, is fixed for all following processing steps. These feature representations $U(I)$ 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.

2.2. Attentive region localization

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 scale hierarchy, employing a maximum search algorithm (fig. 4a). Subsequently, this initial salient map element is expanded dependent on the underlying local saliency distribution to enable the selection of an elliptical region. Elliptical regions have been found to cover better facial regions such as eyes than rectangular ones. For this, a 2D Gaussian function is fitted to the local saliency distribution (fig. 4b) and a certain contour line h is taken to fix the boundary of the elliptical region given by the following ellipse equation:

$$(\vec{x}-\vec{m})^T Cov^{-1}(\vec{x}-\vec{m}) = h = const. \tag{5}$$

where $\vec{m}$ is the expectation of the 2D Gaussian function or the center of the ellipse, Cov is the covariance matrix of all map elements or pixels of the localized region and h a certain contour line. The parameters of the fitted ellipse are updated iteratively to optimally cover the underlying saliency representation at each resolution level. After this iteration the computed region is projected to the next higher resolution level (fig. 4b) and the region adaptation is computed again until the highest resolution level is reached (fig. 4c). The computed elliptical region is selected at the highest resolution level and it is well adapted to the extend and the orientation of the underlying local saliency distribution. To enable the localization of further facial regions the last detected one has to be suspended from the following localizations steps. For this the suspension representation $SR(t)$ is used as mentioned before (fig. 4d). Applying this procedure, all salient facial image regions can be located and further processing modules can be applied.

2.3. Results of the region localization

The results of the computation of salient facial regions are spatially well restricted image areas, which are marked by the number and by the density of included attentive cues. Fundamental to the attentive region localization is the adaptive strategy which is able to foveate dynamically salient regions ranked by their importance.

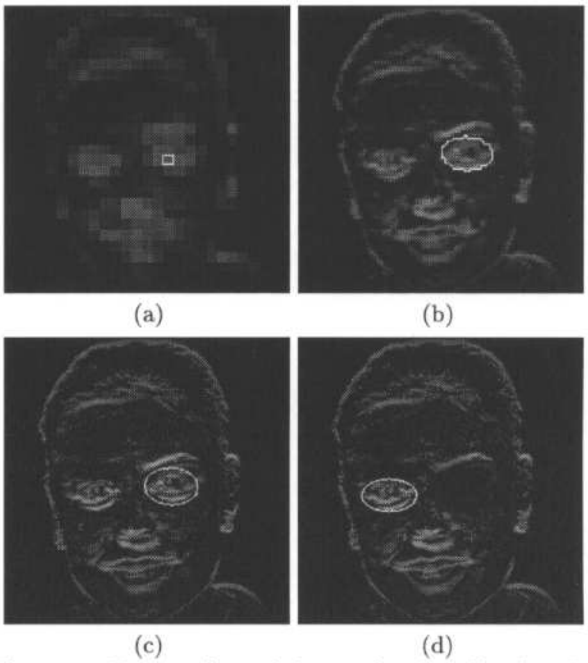


Figure 4. Processing of the region localization by evaluating the multi-scale representation of the saliency map. Maximum search of the element with the highest saliency (a). The maximum element is expanded optimally to the local extension of the feature representation at each scale. Subsequently the computed region is projected to the next higher resolution level (b). Final selection of the first eye region at the highest resolution level (c). After a detailed analysis step, the analyzed eye region is suspended from the following localization steps (d). In addition the local analysis enables an anticipation of the second eye region which is intended to be detected (d).

The region localization has been tested on more than 100 frontal face 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 photographed faces can also differ from image to image (fig. 5). It has been found that within the first 3 foveation steps both eye regions have been detected correctly in more than 98% of the face images considered.

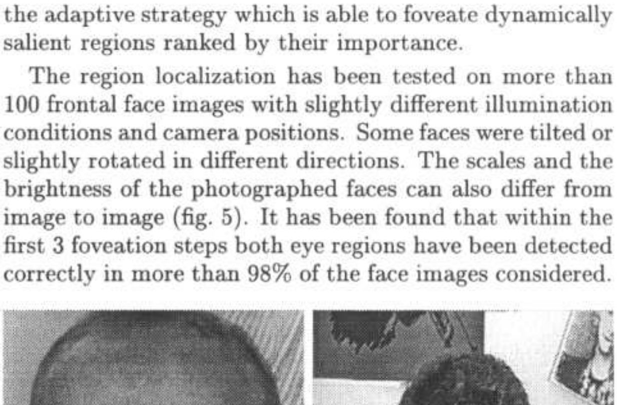


Figure 5. Results of the attentive region localization. The detected regions are spatially well adapted to scale, extend and orientation of the facial regions (a). The attentive localization algorithm is also applied successfully to face images of the face database of A. Pentland[1] although the resolution of the images is too low to apply all subsequent processing modules (b).

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

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 adapted to the particular facial part to be investigated. Therefore, it is important to apply these time consuming methods only to those image parts, for which the methods are developed. Since the eyes contain the most attentive cues in the facial image the attentive region localization developed is able to select the eyes in very early foveation steps (see last chapter). Supposing that just one eye region is located after three foveation steps the complexity of the 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 initialized.

3.1. Point representations

To compute this decision a neural classifier is developed which is motivated by the dynamic link architecture introduced by the Malsburg group [1, 9]. In contrast to these works, where initially identifications of particular subjects are computed, we compute classifications for particular members of a class, e.g. eye regions. Therefore, beside the intraindividual also the interindividual variability has to be considered during the construction of a uniform class prototype. For the recognition of an eye in a facial region a set of point descriptions are established which are located at particular characteristic image positions (e.g. significant corners or intersection points) (see fig. 6).

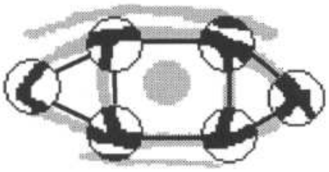


Figure 6. Point descriptions connected by a graph demonstrated at a synthetic eye region. The descriptions at the circled keypoints and the spatial relationship between them are used to clearly represent the characteristic features of eye regions.

The spatial relationship between the several image points are maintained by establishing a 2D graph consisting of nodes for the several point descriptions and of edges for the connections or relations between them. In contrast to the related work, object adapted graphs are used in this application. In detail, an eye region is represented by 12 point descriptions positioned by the nodes of the graph at characteristic object positions or landmarks (fig. 7). Fundamental to this design is that the nodes are positioned only at image positions for which evidence of the local structure exists.

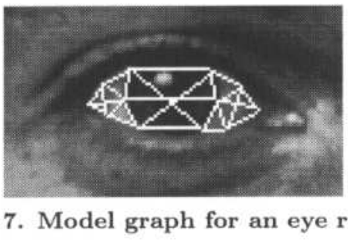


Figure 7. Model graph for an eye region.

Furthermore, we have chosen a new filter set to represent the characteristic image structure at the nodes. To ensure a reliable representation of the underlying image structure at particular image positions consisting mainly of prominent edge and line segments orientation selective edge and line detection filters as well as polar separable filters are applied, which catch selectively the local energy encoded in the surrounding region. For this we have chosen Gaussian edge and line detection filters as well as several polar separable filters also based on Gaussian derivatives [7].

3.2. Graph matching

By applying a two step graph-matching algorithm combined with a simulated annealing approach different facial regions can be distinguished very reliably. In the first processing step the initial starting position is computed while the shape of the graph is kept rigid. In the second step the graph is distorted and the several nodes are moved independently by applying a simulated annealing approach.

Initial starting position

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. 8). No distortion of the graph is allowed only the scale and orientation of the graph is varied to achieve a better invariance properties.

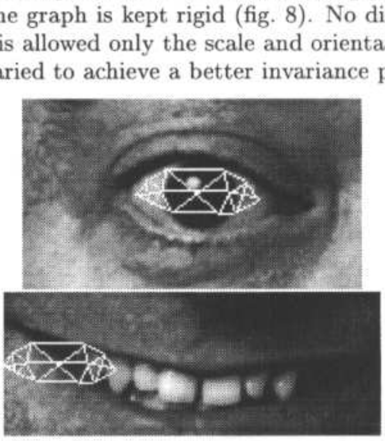


Figure 8. Results of the calculation of the initial starting position of the model graph for an eye region and a non eye region. The calculation will terminate at that position with the greatest similarity to the represented edge information of the model graph.

Simulated annealing

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 [8] is employed to compute the best match of the model graph to a test pattern.

For each node of the graph an alternative position is randomly selected inside a well defined surrounding area. The difference between the similarity of the image structure of the suggested position and the former one determines the acceptance of the new position. This decision problem is realized for each node independently applying a thermodynamic decision rule based on the simulated annealing approach.

Regarding eye regions the adaptation of each node to the corresponding image structure enhances the similar-

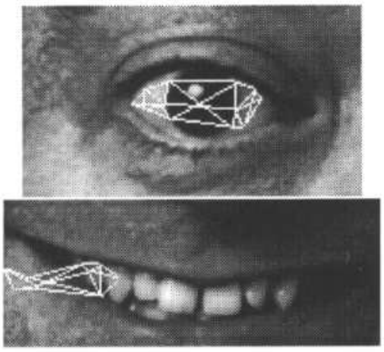


Figure 9. Results of the simulated annealing algorithm demonstrated for an eye region (a) and a non eye region (b).

ity 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 connections (fig. 9). By applying a cost function this property can be used to distinguish reliably between facial regions and others. 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 a little bit worser with 95.5%.

4. STRUCTURAL ANALYSIS

In the third processing step of the attentive processing strategy the foveated regions are analyzed in more detail. The approach starts by detecting the most prominent and reliable features in the facial region. Given our task and image recording conditions these are the strong vertical step edges of the iris in an eye region. 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 [5]. The already detected edges can be checked for their consistency as well as the specific edge structures which are searched for in the next step are given by model assumptions.

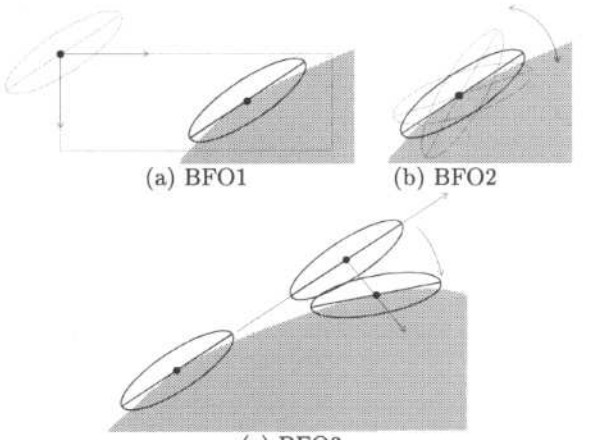
4.1. Basic filter operations

A very flexible filtering scheme is applied which produces many different low-level features providing thus, an expressive data-driven basis for the keypoint detection [5]. 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. 10) that make extensive use of steerable edge and line detection filters [10].

4.2. 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 do not allow a 'classical' computational feasible detection strategy. Therefore, a sequential strategy is needed that uses model knowledge together with the already derived information. The model knowledge also 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. This can be compared to a kind of active concentration process, in which step by step only that information is considered, which is needed to solve the current task.



(a) BFO1

(b) BFO2

(c) BFO3

Figure 10. Basis filter operations. BFO1: Detection of an edge (a). BFO2: Determination of the orientation (b). BFO3: Stepwise tracking (c).

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

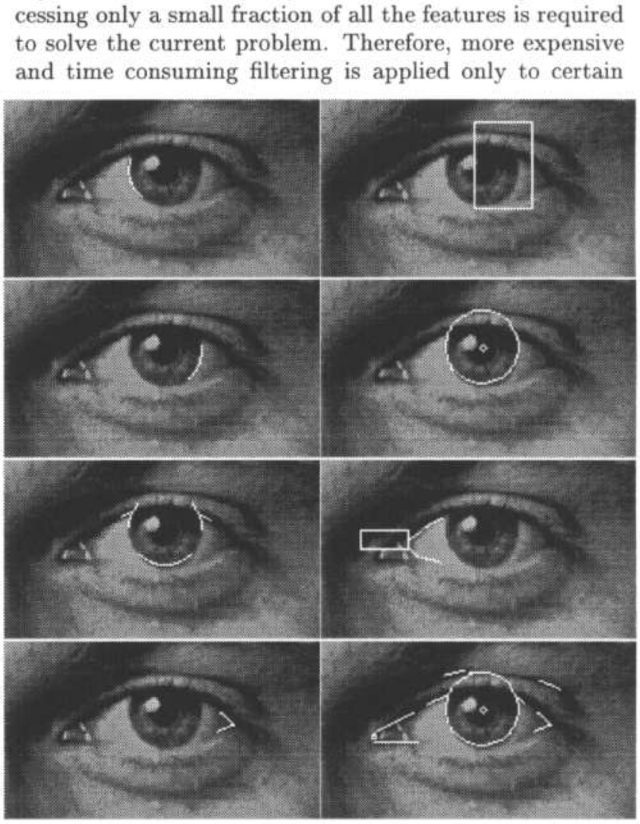


Figure 11. Example steps of the sequential search strategy for the detection of keypoints and prominent structures in an example eye region. First, a prominent vertical bright-to-dark edge is detected (first row). After the detection of the corresponding right edge segment, the final segmentation of the iris is computed (second row). The eyelid edges are searched, tracked, and finally that edge segment which is strongly curved is detected to determine the inner eye corner (third row). For the outer eye corner, also the upper and lower eyelid edges are tracked until they end.

very restricted image areas, where it is really worthwhile.

4.3. Sequential search strategy

The detection of the keypoints in the eye region 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. 11). 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 step of the sequential search. Furthermore, different kinds of edges (white-to-black against black-to-white edges or edges against lines etc.) can be distinguished.

5. RESULTS AND CONCLUSION

The attentive system presented in this work is most closely related to the works of Tsotsos [3] and Califano [2]. In contrast to most of the related work, our approach concentrates on a discussion of a fast data-driven realization of a flexible computer based attentional strategy. Aspects such as the explicit inhibition of non-relevant regions or the neural like modeling were not be addressed in our design. We focus on real-time aspects and the integration of high-level control functions into the low-level analysis processes.

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

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 variations due to the illumination, brightness, and contrast of the studied facial images.

Acknowledgments

We appreciate the support of Prof. Dr. S. Stengel–Rutkowski, of the Kinderzentrum Munich, Institute for Social Paediatrics and Youth Medicine. We are also grateful to our graduate students L. Witta, D. Müller, and K.-H. Lichtenauer. The work is 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.

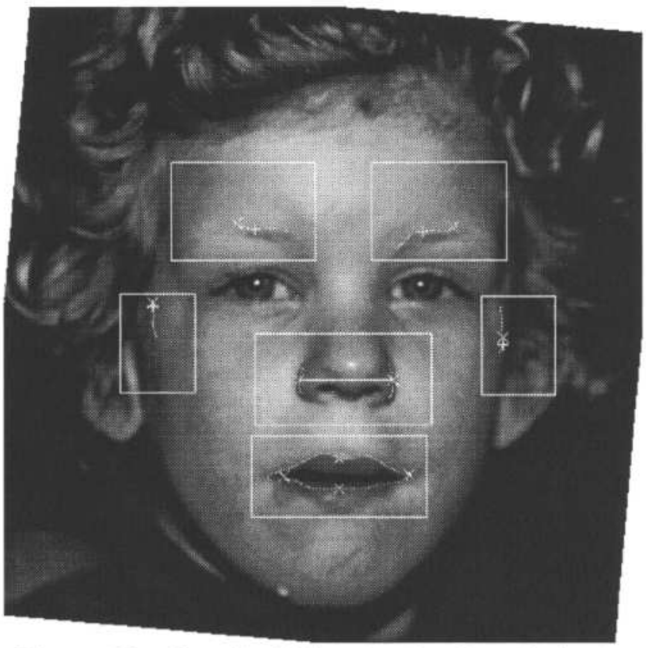


Figure 12. Result of the structural analysis demonstrated for the other facial regions. All important anthropometrical landmarks have been detected concerning figure 2.

[2] A. Califano et al., *Data- and model-driven multiresolution processing*, Computer Vision and Image Understanding, Vol. 63, No. 1, pp. 27-49, 1996.

[3] S.M. Culhane et al., *A prototype for data-driven visual attention*, In: G. Sandini (ed.), Proc. ECCV'92, Springer-Verlag, LNCS 588, pp. 551-560, 1992.

[4] 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.

[5] 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.

[6] R. Herpers et al., *Dynamic cell structures for the evaluation of keypoints in facial images*, to be published in Int. Journal of Neural Systems, special issue NN in CV Applications, 1997.

[7] R. Herpers et al., *Invariant classification of image parts using a dynamic grid of point representations*, to be published in Proc. of 3rd int. Conf. on Engineering Applications on Neural Networks, EANN'97, Stockholm, 16.- 18.06.1997.

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

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

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

[11] 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.

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

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