A common framework for preattentive and attentive vision using steerable filters

Markus Michaelis[1], Rainer Herpers[1], and Gerald Sommer[2]

[1] GSF-Medis, 85764 Oberschleißheim, Germany, Email: (michaeli,herpers)@gsf.de
[2] Institut für Informatik, Christian-Albrechts-Universität, D-24105 Kiel, Germany

Abstract. Motivated by human eye movement strategies a computer based 'attentional mechanism' for processing face images is developed that comprises preattentive and attentive processing strategies. Both parts use a common filtering framework based on steerable filters. During the preattentive processing, prominent facial regions like the eye or the mouth region are localized. The selected regions are analyzed in more detail in the attentive processing step. A variety of complex features are derived applying an efficient filter method based on steerable filters.

1 Introduction

In 'classical' computer vision systems the first processing step commonly consists of the convolution with a few simple filters, e.g. for edge detection. However, the restriction to only one kind of simple features like edges has the drawback that information is lost or that it is not represented in an explicit way. In this paper we propose an approach based on an attentional strategy and steerable filters which allows flexible, efficient, and explicit representations of complex features.

The explicit representation of a huge number of different local image structures causes a data explosion that cannot be calculated, stored or analyzed in practice. To cope with this situation, an attentional strategy is applied. The first observation within such a strategy is that for many tasks the major part of the image can be ruled out as being insignificant by applying simple and fast methods. Motivated by human eye movement strategies [11], the prominent regions of an image are detected in a preattentive processing step and the attention is focused on the spatially limited areas. The more detailed and expensive attentive processing can therefore be restricted to these image regions.

Computer vision systems incorporating attentional strategies have been investigated e.g. in [7, 10]. However, these authors treat only the preattentive processing and they use no explicit representations of the features the attention is based on. In [2] the preattentively focused regions are analyzed at an enhanced resolution. The idea is that high resolution is necessary in some regions whereas it is computationally prohibitive to process the whole image with high resolution.

The problem tackled in this paper is the following: The data explosion is not only due to an enhanced resolution and the image size but also by the huge

[1] *This work is supported by DFG grants So 320/1-1 and Ei 322/1-1.*

Hlaváč, Šára (Eds.): CAIP '95 Proceedings, LNCS 970

number of local degrees of freedom (local image structures, orientations, scales, etc.). However, all these degrees of freedom potentially have to be considered and represented to support higher processing levels with appropriate low-level information. Therefore, the number of different filters is large and the goal is to calculate more complex and time consuming responses only for positions and parameters where it is rewarding.

The attentive processing of the salient regions has to exploit previously derived information to restrict the number of filters, their quality and parameter ranges to a minimum. The decision, which filters and parameters are of interest, can only be taken at run-time. In this paper we propose a filtering scheme that supports such a strategy where the use of expensive filters is controlled by preceding simpler filters. The attentive part therefore is sequential and has a top-down control while the preattentive part is parallel and bottom up. The filtering scheme is essentially based on steerable filters. The filtering of the preattentive and the attentive processing can be combined in a common framework based on one set of basic filters.

2 Steerable filters

It has been shown by several authors (e.g. [3, 8, 5, 6]) that it is advantageous for various tasks in computer vision to have the response of a certain filter $F(\mathbf{x})$ in a continuum of orientations (θ) or scales (σ). For the efficient calculation of such continuous responses, all deformed filters are reconstructed from a small number of so called **'basis functions'** $A_k(\mathbf{x})$. 'Steerability' then refers to the following reconstruction formula with superposition coefficients b_k that depend on the deformation:

$$F_{\theta,\sigma}(\mathbf{x}) \;=\; \sum_k b_k(\theta,\sigma) A_k(\mathbf{x}) \qquad (1)$$

Steerable filters were introduced in [3] and [8]. In this paper we use the singular value decomposition approach to steerability that has first been proposed by Perona [8]. For a detailed discussion of steerability, especially for the case that several deformations (orientation and scale) are steered simultaneously we refer to [6]. Following we show as an example for steerability the filters our attentive processing scheme is based on.

A second and first derivative of Gaussian with an aspect ratio of 2 (fig. 1a,b) are steered in orientation and two octaves in scale. The optimal basis functions are polar separable (fig. 1e-h). These basis functions have no interpretation or meaning themselves, they are only used for reconstructing the filters. However, they are some sort of local 'activity' detectors what is exploited for the preattentive processing.

With 30 basis functions for the first derivative and 40 for the second (the filters are steered separately, i.e with different basis functions), we obtain very good approximations with less then 3% L^2 error. The following properties of the basis functions support the efficient adaption of the speed and quality of the steered filters: (1) The basis functions are orthogonal. Thus it is easy to add new basis functions for a better reconstruction without changing the coefficients

b_k for the old basis functions. **(2)** All deformed filters can be reconstructed, even if only a small number of basis functions are used.

Hence, we can start with a small number of basis functions and add more basis functions only where it is necessary. Using only 13 and 10 basis functions we have about 25% and 22% L^2 error but the properties of these 'poor' approximations (fig. 1c,d) are qualitatively those of the original filters so that they are sufficient for many tasks

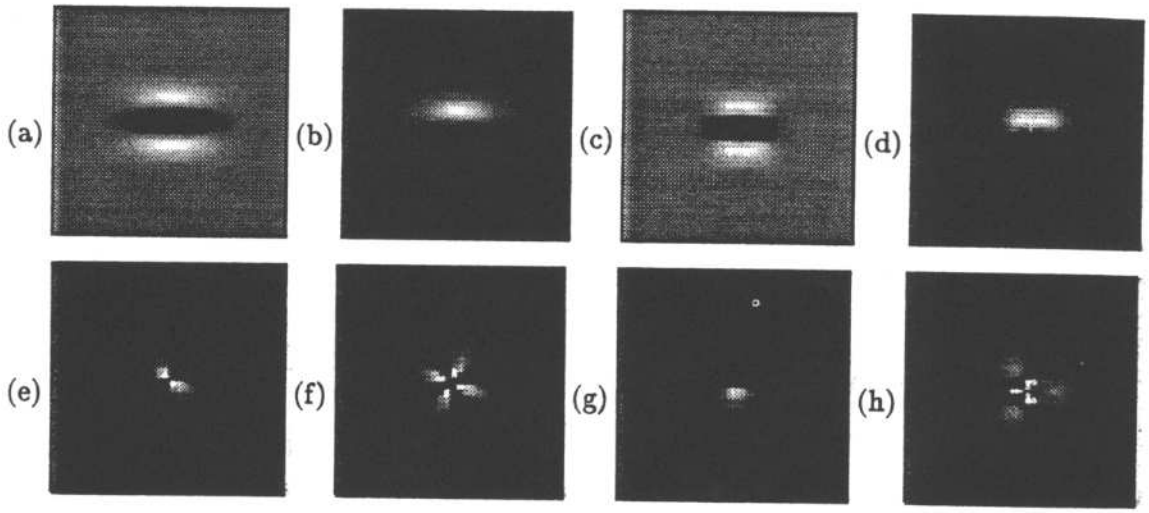

Fig. 1. Filters that are steered in orientation and scale (a),(b). Approximations of the filters with 13 and 10 basis functions (c),(d). Examples of the basis functions (e)-(h).

3 Preattentive processing

Preattention in biological systems is based on the analysis of several simple features which are derived from the visual input [9]. At an early stage of the visual processing this feature representation carries all the information to control the 'visual attention'. For a preattentive foveation in artificial attention systems the salient stimuli of the image objects of interest have to be derived. In the case of detecting prominent facial regions only the edge or local activity information is sufficient for their localization [1].

This preattentive part of the image processing strategy has a datadriven and parallel character and can be compared with the parallel preattentive processing in biological visual systems. 'Parallel' implies that a small set of simple features are derived for the whole image without any local adaption to the data. The different processing steps of the preattentive localization are shown in fig. 2a.

Saliency representation: The preattentive search of salient regions is based on a saliency representation S where the brightness at a pixel-position encodes its saliency (fig. 2b). It is computed from a feature representation F and a control representation C: $S = F \times C$. In general, F can represent any salient features. In the case of face recognition, however, it is sufficient to represent the local activity in the image, detected by fast and simple filters. It is advantageous to use filters

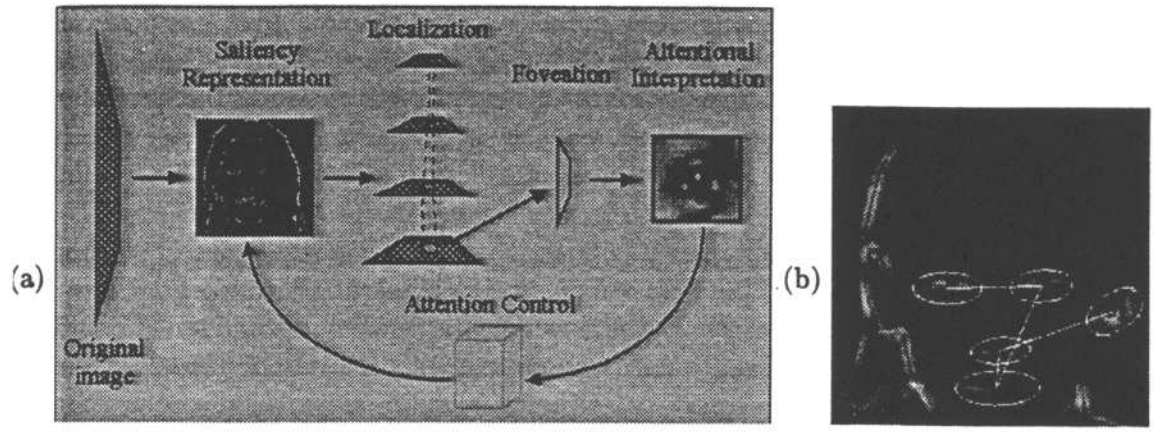

Fig. 2. Flow chart of the preattentive localization (a). Saliency representation (b) derived by the basis functions of the steerability scheme (fig. 1). The scan path shows the first 5 foveated regions after applying the region-adaption algorithm.

with more smeared responses than edge detectors have. Such filters are provided by the steerability scheme (section 2). The basis functions (fig. 1e-h) are viewed as local activity detectors in our application.

The representation C is introduced to control the sequential foveation of the different salient regions. Already detected regions are suspended from the following processing by decreasing the saliency in their local neighborhood. The saliency representation S is scaled in a multiresolution pyramid. This pyramid is the basis for all following localization processes (fig. 2).

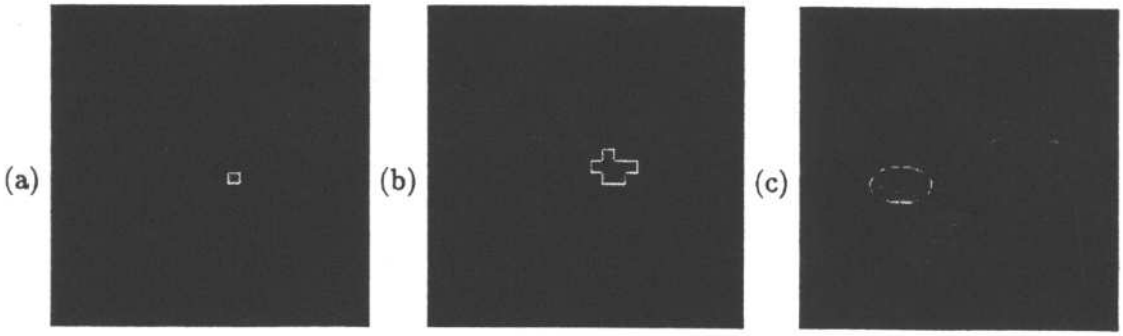

Fig. 3. Preattentive localization of two prominent facial regions (left and right eye) by evaluating the saliency maps in the multi-scale representation.

Localization process: The localization process starts with a maximum search at the lowest resolution level of the scale hierarchy (fig. 3a). The detected maximum element is expanded to a preliminary region by applying an iterative region-growing algorithm (fig. 3b). The region is expanded such that it wraps in size and orientation the bright blob in the saliency representation (fig. 3c). This coarsely localized region is projected to the next higher resolution level of the saliency representation and the region-adaption algorithm is applied again. At the finest resolution level an optimized wrapping of the underlying salient region

is gained. It is a compromise between detecting single bright pixels and melting regions that we would like to be separated (e.g. nose and mouth). To compute the next salient region, the already foveated region is suspended from the following processing steps (fig. 3c, dark blob at the already detected right eye). The procedure of localization and subsequent suspension is iterated until all salient image regions are detected.

Invariance properties: The proposed preattentive strategy shows good invariance properties. The robustness of the developed attentive mechanism is demonstrated in fig. 4. The prominent facial regions are detected independently of their orientation. Regions in differently scaled images can be detected without changing the control parameters. In addition, the proposed search algorithm is able to handle relatively variable preconditions as illumination, brightness and noise (fig. 4). No special preprocessing of the image data is necessary, like normalization or segmentation. For more details see [4].

Fig. 4. Demonstration of the invariance properties of the developed localization algorithm. The localization of the regions is invariant with respect to rotation, noise and illumination (from left to right).

4 Attentive processing

In this section we propose a filtering scheme for the attentive processing. A variety of features and degrees of freedom have to be considered for the detailed analysis of the preattentively localized regions. For this, a collection of different filters are needed. Due to the large number of filters, it is unreasonable to convolve the regions with all filters as it is done in 'classical' schemes. The idea for the processing in the attentive part is that special filters are applied only for certain positions and parameters where it is rewarding. For this, a priori knowledge and information which is derived by preceding simpler and faster filters are exploited. We therefore propose an efficient and flexible filtering scheme based on steerable filters that supports such a strategy.

We start with the filters from fig. 1, steered in orientation and scale, that we refer to as 'simple' filters. The regions are convolved with the first 13 and 10 basis functions of the even and odd simple filter. Then we can reconstruct the

simple filters for all scales and orientations and pixel-positions. By superposing appropriate simple filters from different positions we can synthesize many different 'complex' filters. Examples are shown in fig. 5. Note that even the one-sided filter in fig. 1 is a complex filter because it needs simple filters from different positions if it is steered in orientation or scale.

The 23 basis functions are quite a few filters but (1) some of them already are available from the preattentive processing, (2) they are reused for many different complex filters, and (3) their size is much smaller than the size of the complex filters. More basis functions for more accurate reconstructions of the simple filters are added only where the responses of the 'fast' versions are ambiguous.

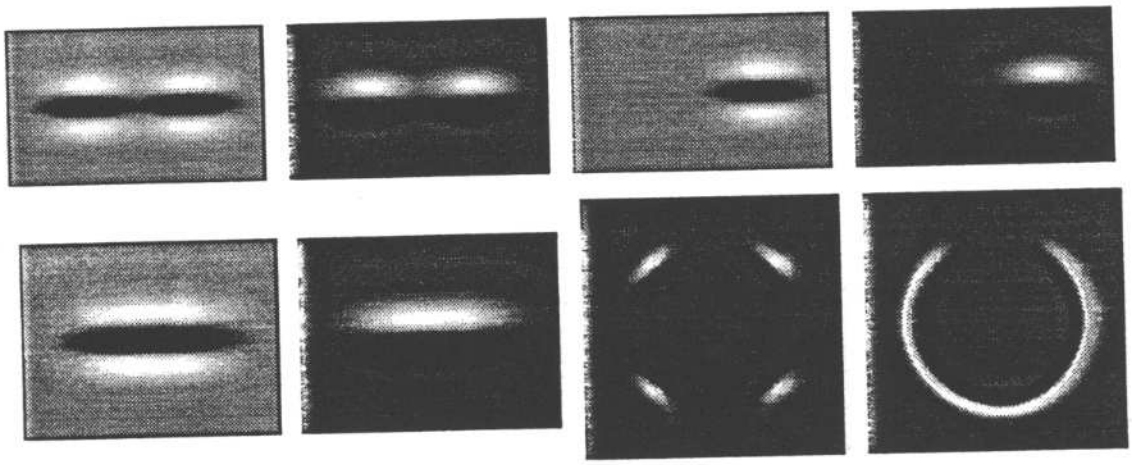

Fig. 5. Depicted are some complex filters, all composed from the same simple constituent filters from fig. 1. The upper row shows a double lobed and a one sided filter (rotated about a shifted center). The lower row shows a filter with a higher orientation selectivity and a circular filter. The circular filter is shown in a 'fast' and 'slow' version with 90° and 10° spacing between the constituent filters.

It is obvious that the complex filters can also be steered in orientation, scale and various other parameters (radius of the circular filter etc.). However, the synthesis and steering of the filters will not be perfect because spatially the simple filters are only available on a discrete grid. At larger scales or if no details of the response are of interest, the quality is sufficient. For a more accurate version we interpolate positions not on the grid by the four neighboring pixel-positions. Then the complex filter is about 5 times slower but it is only applied for some positions and parameters, so that the overall performance is usually close to that of the 'fast' version.

This filtering scheme is very efficient for the following reasons: (1) The basis functions are used for many different complex filters. (2) Several of the basis functions have already been convolved with the region in the preattentive step. (3) Only fast 'low quality' reconstructions of the simple filters with a few basis functions are applied in general. More basis functions are added only where the response is ambiguous. This is supported by the properties of the basis functions (section 2). (4) The complex filters can be adapted in speed and quality by

varying the number of simple filters (for the circular filter, interpolation etc.).

One example for the performance is the following. The double lobed filter from fig. 5 has been applied in [5] with a 'direct' steering of this filter. The proposed method is, dependent on the details, at least 10 times faster for the same filter. It is even more efficient if many complex filters are used because they are all based on the same basis functions.

Example of application: We now show examples where the proposed filtering method is applied. Figure 6 shows an eye region of a face image. Within facial images, a large response of a circular filter (fig. 5) is characteristic for the iris. Therefore, we use the circular filter to detect and localize the eye. First, the fast version with four constituent filters and 13/10 basis functions is applied. The hypothesis of an eye region is rejected if there is no response above a certain threshold. In the other case, the accurate filter is applied for those positions, where the response is above the threshold. This is usually the case for only about 1% of the pixels of the foveated region. Hence, we get the performance of the perfect filter with the costs of the fast filter. The responses of both filters are shown in figure 6.

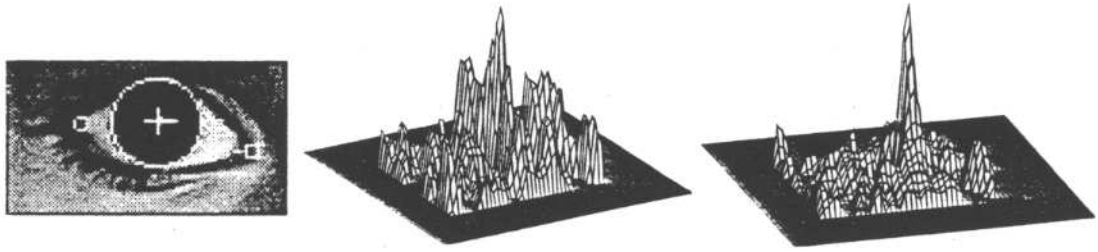

Fig. 6. Eye region with eye-corners and iris (left). Response of the circular filter to this region; 'fast' filter (90° spacings, 13/10 basis functions) middle, accurate filter (10° spacings, 40/30 b. fct.) right. The response is set to zero at the border of the region.

Another example is the characterization of the right eye corner using a one sided filter (fig. 7). The expensive interpolated filter is only applied for those orientations where the fast filter has a (significantly) non-zero response. This example has also been treated in [5] but the presented filtering method is much faster. It is also more flexible because it supports an easy variation of parameters like the shift of the center of the one sided filter or the combination with other complex filters.

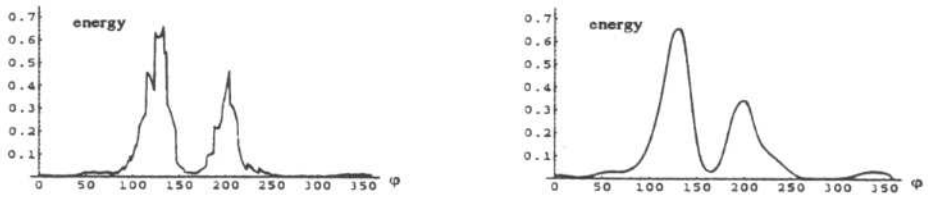

Fig. 7. Orientational signature at the right eye corner for the one-sided filter. The non-interpolated response (left) and the 4 nearest neighbor interpolated and low-pass filtered response (right) are depicted. The jaggedness is an artifact from the pixel discretization. The interpolated response has no visible difference to the true response.

5 Conclusions

In this paper we argued that computer vision systems that exploit a variety of complex features need attentional strategies to cope with the huge amount of data. We demonstrated that the prominent regions in face images can be detected by a preattentive localization mechanism applying only simple and fast filters. The attentive processing of the detected regions uses a variety of different complex filters. We proposed a filtering scheme based on steerable filters for the efficient calculation of their responses. All filters of the preattentive and attentive processing are based on the same set of basis functions. The multiple reuse of these basis functions for the different filters makes this scheme very efficient. The speed and accuracy of the filters can easily be adjusted by the number of basis functions that are used.

We are currently developing a computer based system that automatically detects and localizes a set of keypoints in face images (eye corners, mouth ends etc.). This system is based on the attentional strategies presented in this paper.

Acknowledgements: We would like to thank Lars Witta and Holger Kattner for their help. The work is supported by DFG grants So 320/1-1 and Ei 322/1-1.

References

1. R. Brunelli and T. Poggio, *Face recognition: features versus templates*, IEEE PAMI-15, 1042-1052, 1993.
2. A. Califano, R. Kjeldsen and R.M. Bolle, *Data and model driven foveation*, Proc. 10th. ICPR 90, 1990.
3. W.T. Freeman and E.H. Adelson, *The design and use of steerable filters for image analysis*, IEEE PAMI-13, 891-906, 1991.
4. R. Herpers, H. Kattner, H. Rodax and G. Sommer, *GAZE: An attentive processing strategy to detect and analyze the prominent facial regions*, Proc. Int. Workshop on Automatic Face- and Gesture-Recognition, Zurich, Switzerland, 1995.
5. M. Michaelis and G. Sommer, *Junction classification by multiple orientation detection*, ECCV'94, Stockholm, Vol.I, Eklundh (Ed.), 101-108, 1994.
6. M. Michaelis, *Low level image processing using steerable filters*, PhD thesis, Christian-Albrechts-Universität, D-24105 Kiel, Germany, 1995.
7. B. Ohlshausen, C. Anderson, and D. v. Essen, *A neural model of visual attention and invariant pattern recognition*, CNS Memo 18, Caltech Pasadena, 1992.
8. P. Perona, *Steerable-scalable kernels for edge detection and junction analysis*, ECCV'92, 3-18, 1992.
9. A. Treisman, *Preattentive processing in vision*, CVGIP, vol.31, 156-177, 1985.
10. J.K. Tsotsos, *Localizing stimuli in a sensory field using an inhibitory attentional beam* Tech.-Rep. M5s 1A4, Univ. of Toronto, Dep. of Comp. Sci., Canada (1991)
11. A.L. Yarbus, *Eye Movements and Vision*, New York: Plenum Press, 1967.