Attentive Face Detection and Recognition *

Volker Krüger, Udo Mahlmeister, and Gerald Sommer

University of Kiel, Germany,
Preußerstr. 1-9, 24105 Kiel, Germany
Tel: ++49-431-560496
FAX: ++49-431-560481
vok@informatik.uni-kiel.de,
WWW home page: http://www.informatik.uni-kiel.de/~vok

Abstract. In this paper we will present an approach for the attentive detection and recognition of faces in gray-value images. The approach is biologically motivated. The attentive face system, as we call it, shows great robustness with respect to scale, rotation, viewing orientation, changes in illumination, facial expressions, partial occlusions and other distortions caused, e.g., by glasses or a beard. The system has knowledge of several templates of different persons as well as of their exact relative positions. In a first low-level step the system detects relevant image features by evaluating a similarity measurement between local image features and known facial templates. In a second high-level step the system verified the consistency of these features by using the knowledge of the exact relative positions of the templates and reports whether a face was recognized, detected or whether no face was present.

Keywords: face detection, face recognition

1 Introduction

A face detection and recognition system that is supposed to work under the variety of different real world situations has to be robust and discriminative at the same time. It has to be robust in order to compensate changes in lightning and contrast in the image, changes in different expressions and orientations of the faces that is to be detected and changes in scale. Furthermore it has to be able to detect all different kinds of faces, including bearded faces, faces with glasses, faces with long hair, etc. Still, the system has to be discriminative for recognition purposes. In [9] it is argued that, in contrast to [1], neither a template based approach nor a feature based approach per se is able to accomplish such a task. According to them, rather a "hybrid" approach is able to meet the real world requirements. In their approach, they represent faces with graphs labeled with the responses of locally applied gabor wavelets, called jets. The graphs were matched using a graph similarity function [9]. In our paper we want to promote an attentive approach for a hybrid type of recognition and detection system.

* This work is partially supported by the DFG grand So 320/1-2

Given a gray-level image, our attentive face system (AFS) is able to detect or even recognize faces in a very robust manner by first

How the attention is directed and how the evaluation is done we will describe in detail in the next sections 2, 3 and 4. In section 5 we will show our experimental results while we will finish our contribution in section 6 with concluding remarks.

2 Methodology

In this section we will delve into details about the methodology and the mechanisms employed by our AFS.

The AFS uses as model knowledge facial templates and their exact relative positions. Facial templates can be derived from different persons and may contain various facial features. We assume for the sake of simplicity, that always the same number J of templates T_{ij} with the same relative positions are taken from each sample face so that T_{ij} represents the jth template at position $\mathbf{p}_j$ in the ith face.

Given a gray-level image the AFS uses attentional mechanisms in order to be attracted by certain for the recognition and detection process relevant image features while, at the same time, arbitrary image features do not have an attractive effect. This is done by defining the relevance by a similarity measurement between an image region and a template: For each template a similarity measurement, the *attention image* is calculated. Similar to the HVS the attentional mechanism of the AFS is very robust with respect to several situations and image formation parameters.

As the success of the AFS is highly depending on a good trade off between robustness and discriminability, the AFS verifies if the image regions that were detected in the previous step are facial features. This is done by using the knowledge about the exact relative positions of the templates. If the overall similarity of templates derived from one single person and positioned at exact relative positions is larger than a given threshold a_{rec}, then a face has been recognized. If, on the other hand, the overall similarity for templates that are derived possibly from different persons and that are positioned at exact relative positions is larger that a certain threshold a_{det}, then a face has been detected. An exact definition of the thresholds will be given in section 5.

The AFS contains therefore two stages that can be summarized as follows:

1. The *attention stage*: The attention stage is a low level stage that classifies the visual information within an image. Using known facial templates, local image features are rated with respect to their importance and relevance for the recognition and detection of a face. The attention stage assures the robustness of the AFS with respect
   - different viewing angles,
   - different colors and intensities of lightning,
   - facial expressions,
   - partial occlusion,
   - size and
   - pan, tilt and curl.

2. The *evaluation stage*: This stage is the high level stage that is responsible for the evaluation of the local image features that were classified in the preceding stage as "important". It discriminates between object features and arbitrary image features. Depending on the local features that were identified as facial features, this stage decides whether a face has been identified, detected or whether no face is present.

In the following two sections we will delve into details about the two stages. In section 3 the details on how images features and facial templates are represented is explained. In section 4 we will discuss the evaluation stage in detail.

3 Attention Stage

In this section we will explain, how local image features and facial templates are represented so that a high robustness of the AFS is garanteed. The attention stage consists of two steps: In the first step the representations of local image features and facial templates are calculated. Facial templates consist of local facial image features such as the eyes, the cheeks, the mouth or nose of possibly different persons. Their exact relative positions are known to the system. Also it is known from which person they are derived. Facial templates and local image features are represented in the same manner. Then, in the second step, the representations of the facial templates are each compared with the representations of the local image features. A similarity measurement of the facial templates with local image features is used as the measurement for their relevance in the image. The result of the attention stage is an *attention image*. In figure 1 an example can be seen. The very left image is given as a test image. In this image, the right eye of that person (figure 1, center) has to be detected. The application of the attention system results then in an attention image (figure 1, right) that shows a maximum of attention (black) at the position of the right eye. The grade of blackness in the attention image represents the grade of relevance or similarity for the test image.

It has to be pointed out that the representations of the local facial features have to be calculated just once. Therefore only for the given image, the corresponding representation needs to be calculated.

3.1 Representation of Local Image Features

As the basic image representation we use *local orientations* of gray-value patterns within an image. As inspired by Granlund[2] we assign to each point within the image a local amplitude a and a local orientation θ. With this, a correlates with local contrast while θ correlates with the angle of the predominant orientation. Each value is given as the sum of four orientation selective filters h_i, $i = 0 \ldots 3$:

$$a = \left| \sum_{i=0}^{3} |h_i| \, e^{-j2\Theta_i} \right| \qquad \theta = \arg \sum_{i=0}^{3} h_i e^{-j\Theta_i} \tag{1}$$

where Θ_i denotes the orientation of the ith filter. It has to be pointed out that, in opposition to Granlund, equation (1) defines orientations ranging form $0° \ldots 360°$, whereas Granlund defined orientations ranging form $0° \ldots 180°$. In

Fig. 1. The very left image shows a test image while the center image shows the right eye of the person in the image. For both images the representations of local image features are computed and compared. The attention image to the right shows a maximum of relevance (black) at the position of the right eye. White regions in the attention image represent a minimum of relevance.

Fig. 2. From left to right: points (black) without orientation, points with one predominant orientation, point with multiple orientations.

figure 2 the left image of figure 1 has been used as an example: The very left image shows points without orientation, the center image shows points with one predominant orientation while the right image shows image points with multiple orientations.

Local orientations have been embedded into the steerable pyramids[6,7] as introduced by [3]. This concept we call *local orientation pyramid* (LOP) which is the means we use for representation of facial templates and local image features. The LOP provides joint steerability in position as well as in orientation. Fig. 3 shows the block diagram of the steerable pyramid architecture. By replacing the filled dot with the dashed box a new level of recursion is entered. Four pyramid levels were used for our experiments. The orientations h_i (see eqn. (1)) are calculated with four band passes $\mathcal{B}_i,\ i = 0 \ldots 3$. Their responses supply a 4D-vector $(h_0\ h_1\ h_2\ h_3)^T$ for each point within the pyramid. The band passes $\mathcal{B}_i,\ i = 0 \ldots 3$ are scale dependent and are designed as rotated copies at orientations $\theta_i = \frac{1}{4} i\pi$.

Steering position enables the interpolation of missing samples which allows a conversion of the image pyramid into an image heap that contains images of even size. Steerability in orientation is used for calculating local amplitude and orientation as denoted in equation (1) while using the responses $h_i,\ i = 0 \ldots 3$ as basis functions.

3.2 Similarity Measurement of Local Image Features

In this subsection we will discuss how the similarity of facial templates with local image features is calculated. It has been shown in [5] that the use of color

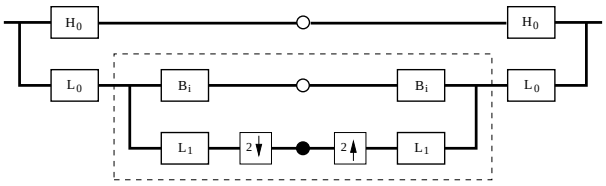

Fig. 3. Steerable pyramid architecture[7]: initial high-pass $\mathcal{H}_0$, low-pass $\mathcal{L}_0$, recursive subsystem (dashed) with low-pass $\mathcal{L}_1$, sub-sampling, and orientation selective band-passes $\mathcal{B}_i$.

histogram matching provides a fast and robust technique to determine the position of a specified object within a scene. As described in [3], we adapted the *intersection*-technique [5, 8] for multiple-scale local orientation. To be precise, the upper three pyramid layers of the LOP are transformed into a heap of even sized layers by enlarging the first and second layer to the size of the third pyramid layer. This is achieved by steering in position which allows the interpolation of the missing samples. With this, a $N \times N$ region in each heap layer can be associated with each pixel $\mathbf{x}$ of the top LOP layer ($N \times N$ is the size of the region of support of pixel $\mathbf{x}$).

The histograms are then calculated by first extracting 3D vectors $z^T = (z_0 \;\; z_1 \;\; z_2)^T$ from the upper three heap layers where the ith vector component z_i is derived from the ith heap layer. Then, the vector components are quantized to five different values: the "null-orientation" if the contrast of the corresponding pattern is smaller than a certain threshold $|z_i| < a_{\mathrm{thr}_i}$ and the four values 0°, 90°, 180°, 270°. The resulting histograms have therefore each a size of $L = 5^3 = 125$ bins. In the left image of fig. 1 regions can be seen that show patterns with very low contrast. To these regions the "null-orientation" has been assigned as shown in the left image of fig. 2. The threshold a_{thr_i} is assigned to a fixed fraction of the maximum amplitude occurring in each pyramid layer i. For each top layer pixel, the histogram is calculated over the three associate heap regions.

For example, the support of one pixel in the top layer of the LOP, in case of the four-layer pyramid of our AFS, is $N = 16$ image pixels. We calculate for each LOP top layer pixel a histogram. Each histogram then contains $3 \cdot 16^2 = 768$ values. For an image I of size 640×480 we get $640/2^4 \cdot 480/2^4 = 40 \cdot 30$ histograms $I^H(\mathbf{x})$ where $\mathbf{x} = (0 \ldots 39, 0 \ldots 29)$. A 32×32 template T is therefore described by $2 \cdot 2$ histograms $T^H(\mathbf{x})$ with $\mathbf{x} = (0 \ldots 1, 0 \ldots 1)$, respectively.

Given a template histogram T and an image histogram H the intersection is defined as [8]:

$$\cap(T, H) = \sum_i \min(T_i, H_i) \tag{2}$$

The attention map $A(I, T)$ for an image I and a template T is then given by evaluating at each position of the corresponding histogram image I^H the intersection with the histogram image T^H of the template T:

$$A(I, T)(\mathbf{x}) = \sum_{\mathbf{k}} \cap \left(T^H(\mathbf{k}), I^H(\mathbf{x} + \mathbf{k}) \right) \quad . \tag{3}$$

In [5] several different histogram techniques have been tested. Furthermore it is reported that the intersection technique needs sparse histograms. However, we have found that for our purposes this method shows the best results. For extensive experiments and more details see [4].

Insensitivity of the AFS with respect to changes in illumination color and intensity is achieved because the LOP basis filter responses are quite insensitive to illumination changes. Further insensitivity is achieved by considering primarily local orientation information during histogram evaluation while local amplitude, which carries contrast information, is discarded. In this respect our method differs from those primarily ignoring phase[10] or those using the filter responses directly[5]. Robustness to geometric distortions implies a mechanism which considers a large local context. This is realized by the sliding window for histogram calculation. The size of the sliding window determines to what extend neighboring local orientations are considered for the similarity measurement between the image and the template. Robustness to changes in orientation, changes in scale of the face, partial occlusion of changes of the viewing position are gracefully inherited from the similarity measurement[8].

4 Evaluation Stage

In this section we will delve into details about the evaluation stage. This stage is responsible for the evaluation of the local image features that were classified in the preceding stage as "relevant". It discriminates between object features and arbitrary image features. Depending on the local features that were identified as facial features, this stage decides whether a face has been identified, detected or whether no face is present.

As explained above J different facial templates of each face as well as their exact relative positions are known to the AFS. For simplicity, let T_{ij} be the jth template at position $\mathbf{p}_j$ of the ith face. Given an image I, attention maps $A(I, T_{ij})$ for each template T_{ij} were calculated. The AFS reports a face as recognized if the overall similarity of the templates T_{ij} that are derived from the ith person and that are positioned at the positions $\mathbf{p}_j$ is larger than a given threshold a_{rec}:

$$\sum_j A(I, T_{ij})\,(\mathbf{x} - \mathbf{p}_j) > a_{\text{rec}} \quad . \tag{4}$$

The AFS reports a face as detected if there exists a set of templates $S = \{T_{i_j j} | \text{ for each } j\}$ containing a template for each $j = 0 \ldots J-1$ of the i_jth face so that the overall similarity of the templates of set S with the image I is larger than a give threshold a_{det}:

$$\sum_S A(I, T_{i_j j})\,(\mathbf{x} - \mathbf{p}_j) > a_{\text{det}} \quad . \tag{5}$$

In practice the similarities are normalized with the maximal similarities of the templates with their original faces. With this, the threshold were empirically chosen.

5 Experiments

For our experiments presented here we used nine different templates of just one single person. The templates are positioned on a regular 3×3 grid within

Fig. 4. The above images show sample results of the AFS: Top row, from left to right: rotation, distortion, glasses. Bottom row: large, small, background without face.

the eye-mouth region of a sample face. As sample face the left image in fig. 1 was used. The templates were taken of constant size 32×32. The eye in the center image of fig. 1 is one of these templates. The size of the sliding window is adapted to the size of the template. Fig. 4 shows some sample results of the AFS. The top row shows (from left to right) a rotated face, a distorted face, and a face with glasses. The bottom row shows faces of different sizes. The table to the left shows the overall similarities of the example images. The similarities have been normalized to the maximal overall similarity of the templates with the original image. Thresholds were choosen empirically and were set in our experiments to $a_{\text{rec}} = 0.85$ and $a_{\text{det}} = 0.7$. As it can be seen the first two images were reported as recognized. By choosing templates of just one single example face allows to demonstrate the great robustness/discriminability tradeoff of the AFS.

image	rel. ov. sim.
rotation	0.9
distortion	0.89
glasses	0.81
large	0.71
small	0.72
background	0.65

It can be seen in fig. 4 that the box that indicates the facial position never encloses the inner face precisely. The reason for this is the large support for each histogram (32×32) and therefore for each pixel within the attention image. This does not allow a precise positioning. It needs to be mentioned that the large neighborhoods considered here for the histogram calculations are, in fact, the major drawback of the system: The size of an image region considered for a single histogram is 16×16 image pixels. Therefore, the size of a template is always a multiple of 16×16.

6 Conclusions

In this paper we have presented an approach for the attentive detection and recognition of faces in gray-value images. The approach is biologically motivated and showed in our experiments a very good robustness including robustness to scale, rotation, facial expressions and illumination changes. Robustness to illumination was realized by discarding local amplitude information while using local orientation pyramids instead. Robustness to geometric distortions, on the other hand, was achieved by using histogram techniques and by considering large local neighborhoods. It is well known that the local amplitude/phase representation decouples detection and classification of oriented patterns. Whereas a significant amplitude value indicates the presence (visibility) of a pattern the phase at that point yields a classification of its symmetry. With this local amplitude and phase have been accepted as two real feature dimensions. We believe that the role of local phase has to be re-estimated in the context of natural image statistics. In fact, we have kept the evaluation stage simple in order to not covering up the potency of the orientation approach.

The approach of our AFS has proven to be very potential for recognition and detection and will be adapted for a general object recognition system, which is future work.

References

1. Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, Oct. 1993.
2. G. H. Granlund and G. D. Knutsson. *Signal Processing For Computer Vision*. Kluwer Academic Publisher, 1995.
3. U. Mahlmeister, M. Sandgaard, and G. Sommer. Sample-guided progressive image coding. http://www.informatik.uni-kiel.de/~uhm/research/icpr1.ps.gz, submitted to ICPR'98, Brisbane Australia, 1997.
4. M. Sandgaard. Attentionsgesteuerte progressive Übertragung von Bildern. Master's thesis, Institute of Computer Science and Applied Mathematics, University of Kiel, Germany, 1997. in German.
5. B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *Proc. Fourth European Conference on Computer Vision*, Cambridge, UK, April 15-18, 1996.
6. E. P. Simoncelli and W. T. Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. Technical report, GRAPS Laboritory, Philadelphia, 1995.
7. E. P. Simoncelli, W. T. Freeman, E. A. Adelson, and D. J. Heger. Shiftable multiscale transforms. *IEEE Trans. Information Theory*, 38(2):587–607, 1992.
8. M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.
9. L. Wiskott, J. M. Fellous, N. Krüger, and C. v. d. Malsburg. Face recognition and gender determination. In *International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, Switzerland, June 26-28, 1995.
10. L. Wiskott, J. M. Fellous, N. Krüger, and C. v. d. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.