A novel method for automatic face segmentation, facial feature extraction and tracking

Karin Sobottka Ioannis Pitas

Department of Informatics

University of Thessaloniki 540 06, Greece

E-mail: {sobottka, pitas}@zeus.csd.auth.gr

Abstract

The present paper describes a novel method for the segmentation of faces, extraction of facial features and tracking of the face contour and features over time. Robust segmentation of faces out of complex scenes is done based on color and shape information. Additionally, face candidates are verified by searching for facial features in the interior of the face. As interesting facial features we employ eyebrows, eyes, nostrils, mouth and chin. We consider incomplete feature constellations as well. If a face and its features are detected once reliably, we track the face contour and the features over time. Face contour tracking is done by using deformable models like snakes. Facial feature tracking is performed by block matching. The success of our approach was verified by evaluating 38 different color image sequences, containing features as beard, glasses and changing facial expressions.

1 Introduction

Up to now, increasing activity can be observed for the research topic of machine face recognition. It has a wide range of applications, e.g. model-based video coding, security systems, mug shot matching. However, due to variations in illumination, background, visual angle and facial expressions, robust face recognition is difficult. An excellent survey on human and machine

recognition of faces is given in [4]. Also [2] and [13] gives a very good overview about existing approaches and techniques.

In general, the procedure of machine face recognition can be described as follows (Figure 1): Still images or an image sequence and a database of faces are available as input. In a first step, facial regions are segmented. Then facial features are extracted. Afterwards an identification is done by matching the extracted features with features of the database. As result an identification of one or more persons is obtained. Obviously, the order of the first

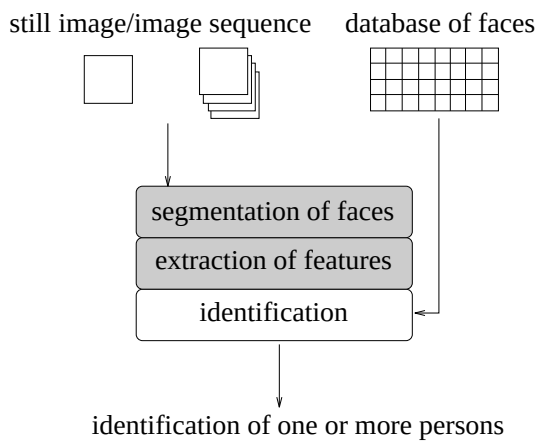


Figure 1: Procedure of machine face recognition

two steps can be interchanged. For example in [3], facial features are extracted first. Then constellations are formed from the pool of candidate feature locations and the most face-like constellation is determined. Also in [11] an approach is presented which derives locations of whole faces from facial parts.

The present paper deals with the first two steps of the face recognition problem. First, we segment face candidates by using color and shape information. Then we extract facial features by evaluating the topographic greylevel relief. Further, we propose an approach for tracking the face contour and facial features over time in image sequences. By using snakes a robust tracking of the face contour can be performed. Facial feature tracking is done by searching for corresponding feature blocks in consecutive frames.

As described in [1], this approach is integrated in a multi-modal system for person verification using speech and image information.

2 Face detection

Several approaches have been published so far for the detection of facial regions and facial features using texture, depth, shape, color information or combinations of them. For example, in [7], the extraction of facial regions from complex background is done based on color and texture information. The input images are first enhanced by using color information, then textural features are derived by SGLD matrices and facial parts are detected based on a textural model for faces. On the basis of facial depth information, primary facial features are localized in [10]. In the first step pairs of stereo images containing frontal views are sampled from the input video sequence. Then point correspondences over a large disparity range are determined using a multiresolution hierarchical matching algorithm. Finally nose, eyes and mouth are located based on depth information. In [8] face localization is done by using shape information. An ellipse is chosen as model for the facial shape and candidates for the head outline are determined based on edge information. In order to extract facial features, first the input image is segmented using color characteristics, then feature points are detected and in the last step the different facial features are approximated by polynomials [16].

Most of the published approaches to face localization and facial feature extraction suffer either from their highly computational expenses or seem to lack robustness. Often, edge information is used for facial feature extraction although facial features are not separated from the background by strong edges.

In this framework, we present an approach for face localization using color and shape information [22]. This combination of features allows a very robust face detection, because faces

can be characterized very well by their skin color and oval shape. Facial feature extraction is done using greylevel information inside the facial regions. Based on the observation that many important facial features differ from the rest of the face because of their low brightness, we first enhance dark regions by applying morphological operations. Then facial features are extracted by the analysis of minima and maxima.

2.1 Segmentation of face-like regions

A robust segmentation of face-like regions can be done by evaluating color and shape information. In a first step, skin-colored regions are detected and then face candidates are selected by verifying the shape information of the skin-colored regions. The effectiveness of using color information was also shown in [7], [25] and [12].

2.1.1 Color segmentation

In our approach we first locate skin-like regions by performing color segmentation. As interesting color space we consider the Hue-Saturation-Value (HSV) color space, because it is compatible to the human color perception. Alternatively, similar color spaces (e.g. HSI, HLS) can be used as well. The HSV color space has a hexcone shape as illustrated in Figure 2a. Hue (H) is represented as angle. The purity of colors is defined by the saturation (S), which varies from 0 to 1. The darkness of a color is specified by the value component (V), which varies also from 0 (root) to 1 (top level).

It is sufficient to consider hue and saturation as discriminating color information for the segmentation of skin-like regions. The hue and saturation domains, which describe the human skin color, can be defined or estimated a priori and used subsequently as reference for any skin color. After extensive experimentation in a large number of images, we have chosen these parameters as follows: $S_{min} = 0.23$, $S_{max} = 0.68$, $H_{min} = 0^0$ and $H_{max} = 50^0$. This is equivalent

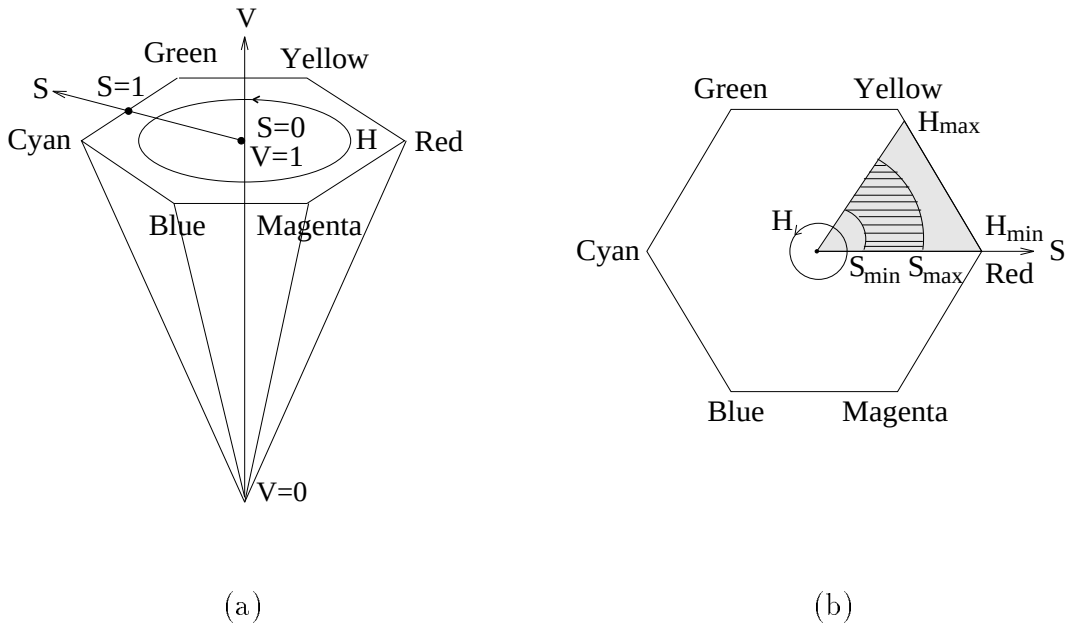


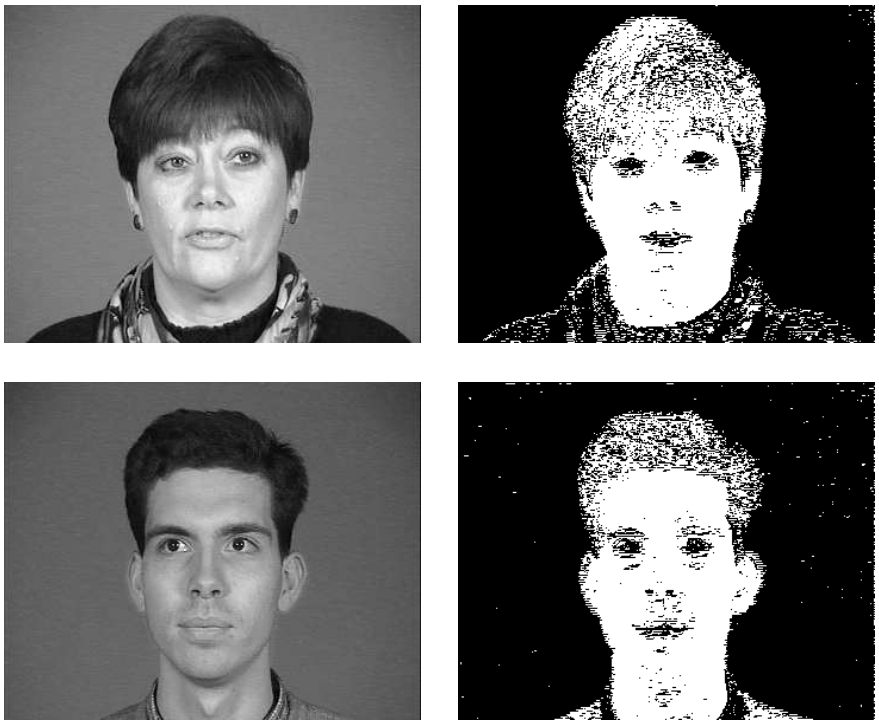
Figure 2: (a) Hue-Saturation-Value color space (HSV) and (b) skin color segmentation in HS space.

to a sector of the hexagon (as it is shown in Figure 2b). Tests on our image database show that these parameters are appropriate to segment the white skin as well as the yellow skin of human beings. How far a segmentation of black skin is possible with these parameters, isn't tested yet. Examples of such a color segmentation are shown in Figure 3a,b.

2.1.2 Evaluation of shape information

The oval shape of a face can be approximated by an ellipse. Therefore, face detection in an image can be performed by detecting objects with elliptical shape. This can be done based on edges [8] or, as we will show here, based on regions. The advantage of considering regions is that they are more robust against noise and changes in illumination. In a first step, we find the connected components. Then, we check for each connected component, whether its shape is nearly elliptical or not.

We find the connected components by applying a region growing algorithm at a coarse



(a) (b)

Figure 3: Color segmentation: (a) original and (b) segmented images.

resolution of the segmented image. For the images shown in Figure 3b, we obtain the results shown in Figure 5a. Then, for each connected component C with a given minimum size, the best-fit ellipse E is computed on the basis of moments [14]. An ellipse is exactly defined by its center $(\overline{x}, \overline{y})$, its orientation θ and the length a and b of its minor and major axis (Figure 4a). The center $(\overline{x}, \overline{y})$ of the ellipse is given by the center of gravity of the connected component. The orientation θ of the ellipse can be computed by determining the least moment of inertia:

$$\theta = \frac{1}{2} \cdot arctan\left(\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}\right) \tag{1}$$

where $\mu_{i,j}$ denotes the central moments of the connected component. The length of major and minor axis of the best-fit ellipse can also be computed by evaluating the moments of inertia. If I_{min}, I_{max} are the least and greatest moment of inertia, respectively, of an ellipse

with orientation θ:

$$I_{min} = \sum_{(x,y)\in C} [(x-\overline{x})cos\theta - (y-\overline{y})sin\theta]^2 \tag{2}$$

$$I_{max} = \sum_{(x,y)\in C} [(x-\overline{x})sin\theta - (y-\overline{y})cos\theta]^2 \tag{3}$$

the length a of the major axis and the length b of the minor axis are given by:

$$a = \left(\frac{4}{\pi}\right)^{1/4} \left[\frac{(I_{max})^3}{I_{min}}\right]^{1/8} \qquad b = \left(\frac{4}{\pi}\right)^{1/4} \left[\frac{(I_{min})^3}{I_{max}}\right]^{1/8} \tag{4}$$

On the basis of the already computed elliptical parameters, a reduction of the number of potential face candidates is possible. This is done by applying to each ellipse decision criteria concerning its orientation and the relationship between major and minor axis. For example we assume that the orientation has to be in the interval of $[-45^0, +45^0]$. For the remaining candidates we assess how well the connected component is approximated by its best-fit ellipse. For that purpose the following measure V is evaluated:

$$V = \frac{\sum_{(x,y)\in E}(1-b(x,y)) + \sum_{(x,y)\in C\setminus E} b(x,y)}{\sum_{(x,y)\in E} 1} \tag{5}$$

with

$$b(x,y) = \begin{cases} 1 & \text{if } (x,y)\in C \\ 0 & \text{otherwise} \end{cases} .$$

V determines the distance between the connected component and the best-fit ellipse by counting the "holes" inside of the ellipse and the points of the connected component that are outside the ellipse (Figure 4b).

The ratio of the number of false points to the number of points of the interior of the ellipse is calculated. Based on a threshold on this ratio, ellipses that are good approximations of connected components are selected and considered as face candidates. For the two images of Fig. 5a we obtain the results shown in Fig. 5b.

Subsequently the determined face candidates are verified by searching for facial features inside of the connected components.

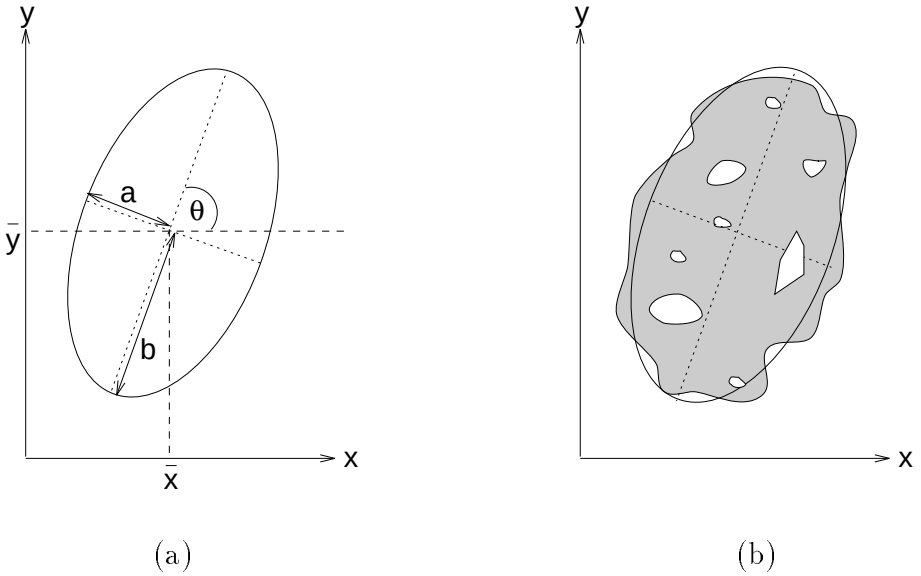


Figure 4: (a) Parameter of an ellipse and (b) approximation of a connected component.

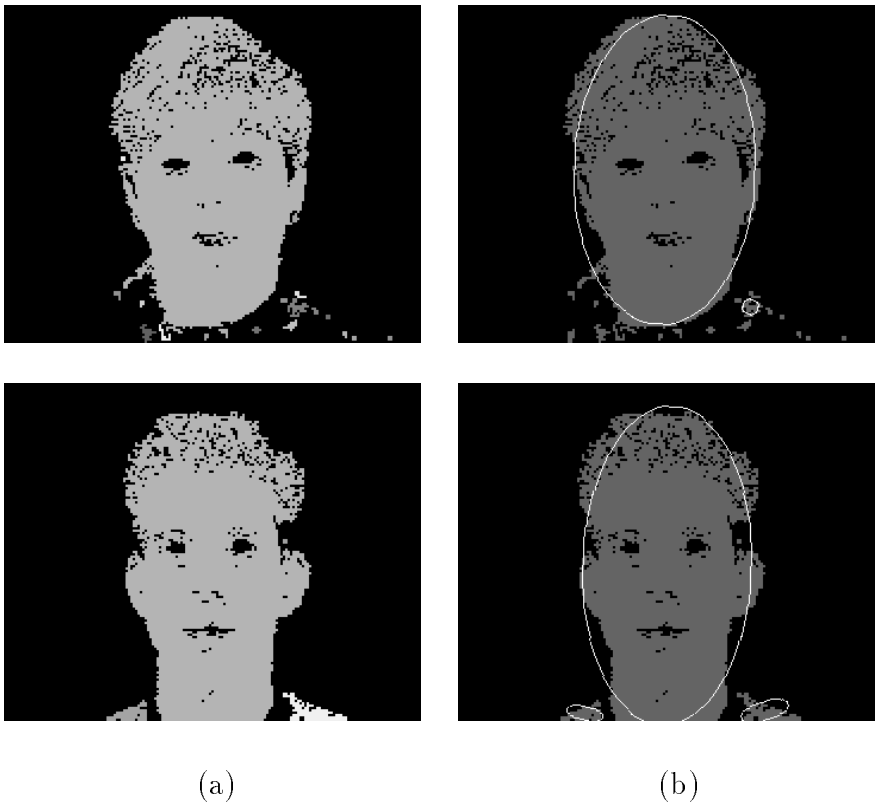


Figure 5: Evaluation of shape information: (a) connected components, (b) best-fit ellipses.

2.2 Facial feature extraction

Our approach to facial features extraction is based on the observation that, in intensity images, eyebrows, eyes, nostrils, mouth and chin differ from the rest of the face because of their low

brightness. For example, in the case of the eyes, this is due to the color of the pupils and the sunken eye-sockets. Even if the eyes are closed, the darkness of the eye sockets is sufficient for eye extraction. Therefore, in the following, we shall employ the intensity information in the interior of the connected components (Figure 6a).

2.2.1 Enhancement of facial features

In a preprocessing step, we enhance dark regions in the interior of the connected components by using morphological operations. First, we apply a greyscale erosion [20]. Then we improve the contrast of the connected component by the following extremum sharpening operation [17]:

$$g(x,y) = \begin{cases} min & \text{if } f(x,y) - min < max - f(x,y) \\ max & \text{otherwise} \end{cases} \tag{6}$$

Here min and max denotes the minimum and maximum value in the neighbourhood of $f(x,y)$. We choose the 5×3 rectangle as neighbourhood. Results of this preprocessing step are illustrated in Figure 6b.

All facial features and parts of the hair are emphasized.

2.2.2 Face model

In our approach, a face can be described by a constellation of minima and maxima of the topographic greylevel relief. Because of the similarity of some of the facial features in their appearance, we subdivide the facial features in three groups. The first group contains eyes and eyebrows, the second group describes the nostrils and the third group characterizes mouth and chin. For each group of facial features a description is defined based on minima and maxima and relative head position (Table 1). For example, candidates for group 1 consist of two significant minima, which are located in the upper or middle part of the head. Between the minima there is a significant maximum and the ratio of distance between the minima to head width is in

(a) (b)

Figure 6: Enhancement of facial features: (a) face candidates with intensity information and (b) enhanced face candidates.

a certain predefined range. Based on this rough description of facial feature groups, a face constellation can be defined as follows:

$$\text{Face constellation} = [cand_1]^*[cand_2][cand_3]^* \tag{7}$$

where $cand_i$ denotes candidates for group i, $i = 1, ..3$. Equation (7) defines that a face constellation may consist of a sequence of candidates for group 1, one candidate for group 2 and a sequence of candidates for group 3 (the star in (7) denotes a sequence of candidates). By considering each part of the face constellation as optional, the face model takes into consideration possible incomplete face constellations. The maximal number of candidates for group 1 and group 3 is limited to two, due to the fixed number of facial features inside of a human face.

group 1: eyebrows, eyes	group 2: nostrils	group 3: mouth, chin
two significant minima	two significant minima	two significant maxima
upper/middle part of head	middle part of head	middle/lower part of head
significant maximum between minima	significant maximum between minima	significant minimum between maxima
ratio of distance between minima to head width is in certain range	small distance between minima	ratio of distance between maxima to head width is in certain range
similar greylevels		

Table 1: Description of facial feature groups

2.2.3 Extraction of facial feature candidates

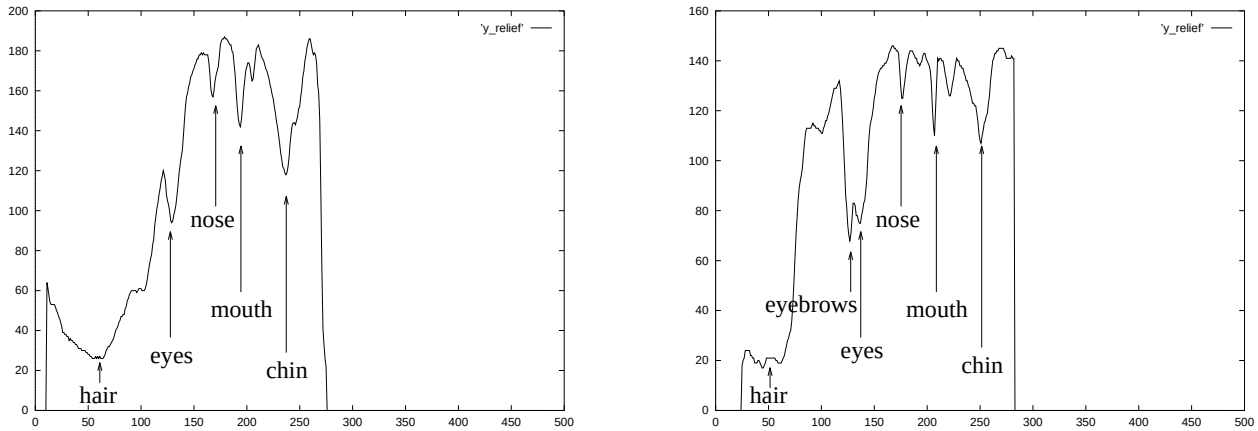
As described in the previous section, we consider three groups of candidates for facial features. Candidates for each of the groups are determined by searching for minima and maxima in the topographic greylevel relief. This can be done by using watersheds ([21],[23]), or as we will show here, by directly evaluating the x- and y-projections of the greylevel relief. Since all facial features are horizontally oriented, a normalization of the orientation of the face candidate is necessary. To obtain this normalization, a rotation of the interior of the connected component is done. The rotation angle is assumed to be equal to the orientation θ of the best-fit ellipse of the connected component. The coordinate transformation is defined by:

$$\begin{pmatrix} xr \\ yr \end{pmatrix} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{8}$$

Hereby xr and yr denote the rotated coordinates of x and y.

After normalization, we compute the y-projection of the topographic greylevel relief. This is

done by determining the mean greylevel of every row of the connected component. We smooth the y-relief by an average filter of width 3, in order to get rid of small variations in the y-relief. Typical examples of y-reliefs are illustrated in Figure 7. In Figure 7a significant minima can be seen for hair, eyes, nose, mouth and chin. The y-relief in Figure 7b contains minima for hair, eyebrows, eyes, nose, mouth and chin.



(a) (b)

Figure 7: Examples for y-reliefs

Significant minima are determined in the y-relief by checking the gradients of each minimum to its neighbour maxima. Each significant minima is considered as possible vertical position for facial features. Therefore, more detailled information is computed along the horizontal direction. Thus, beginning with the uppermost minima in y-direction (leftmost minima in Figure 7), we determine x-reliefs by averaging the greylevels of 3 neighbour rows of every column. Afterwards the resulting x-reliefs are smoothed in x-direction by an average filter of width 3 and minima and maxima are determined. As a result, we obtain for each face candidate one smoothed y-relief with an attached list of its minima and maxima and, for each significant minima of the y-relief, smoothed x-reliefs with attached lists of their minima and maxima.

By searching through the lists of minima and maxima, candidates of the three facial feature

groups are determined. A candidate for group 1 is found, if two significant minima in x-direction are detected that meet the requirements of group 1, concerning relative position inside of head, significance of maximum between minima, ratio of distance between minima to head width and similarity of greylevel values (see Table 1). An example of a x-relief containing a candidate for group 1 is shown in Figure 8a.

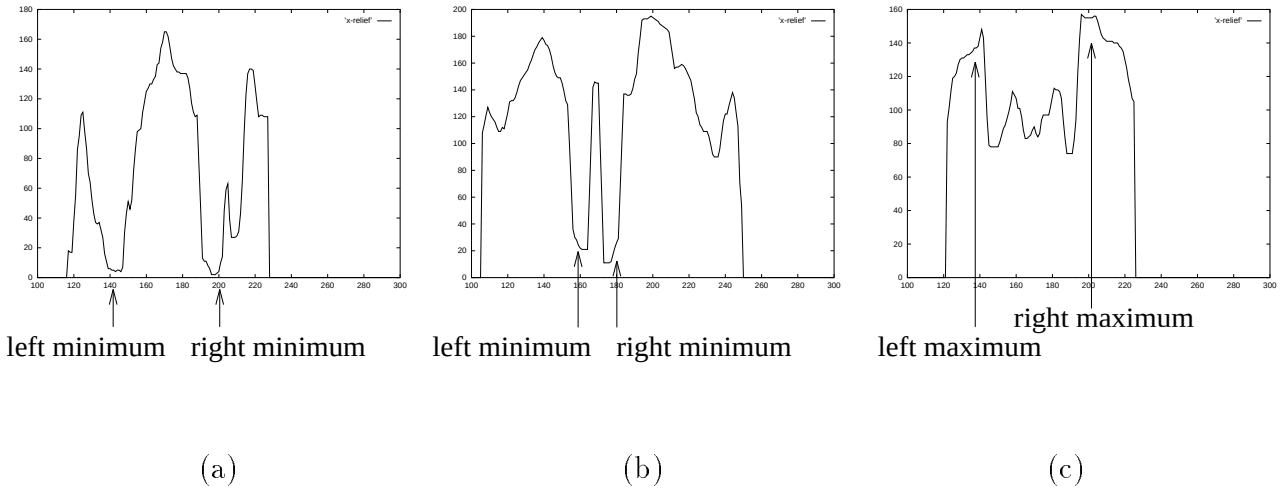


Figure 8: Examples for x-reliefs containing a (a) group 1, (b) group 2 and (c) group 3 candidate.

The assessment of how well a pair of minima meets the requirements, is done on the basis of fuzzy set theory. Thus, we define a membership function for each of the requirements. For example the membership function for the assessment of the ratio of distance between minima to head width is shown in Figure 9.

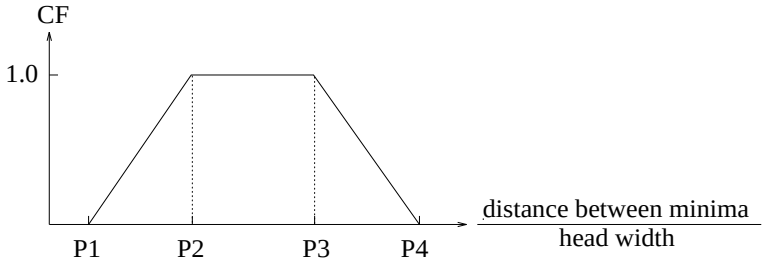


Figure 9: Assessment function for the ratio of distance between minima to head width.

The parameters $P1$, $P2$, $P3$ and $P4$ are defined in dependence on the width of the connected

component that is assumed to correspond to the width of the head. The certainty factor (CF) has the value 1.0, if the measured ratio r is in the range $P2 < r < P3$. If $P1 < r < P2$ or $P3 < r < P4$ holds, the requirement is only partially fulfilled and $0.0 < CF < 1.0$. Otherwise $CF = 0.0$. On the basis of these assessments, candidates for group 1 are selected. They have to meet a minimum assessment for each criterion as well as a minimum assessment for the weighted sum of all assessments.

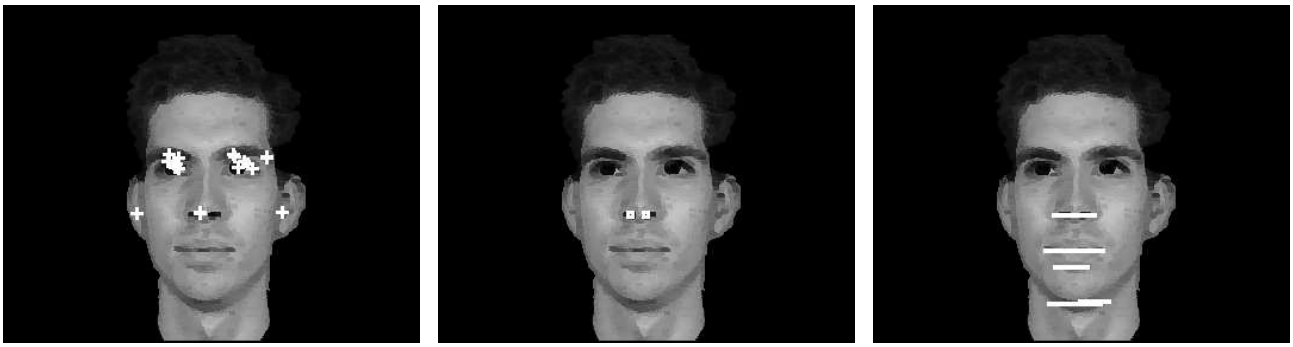
After candidates are determined, we cluster them according to their left and right minimum x-coordinates. This is done to group similar candidates into cluster. Outliers that meet the requirements of Table 1 are eliminated subsequently by considering the symmetry and distances between facial feature candidates.

Clustering is done using the unsupervised Min-Max-algorithm for clustering [9]. Starting with one pair of minima or maxima as first cluster center, distances to all other candidates are computed. The candidate with maximum distance is chosen as the second center. Then, the following iterative procedure starts: For each candidate all distances to all cluster centers are computed and an attachment to the cluster with minimum distance is done. The candidate with maximum distance to its attached cluster is chosen as new cluster center. The iterative process stops, when the maximum distance is smaller than half of the mean distance between cluster centers. As result we obtain a set of cluster centers. Each of them is a pair, consisting of the left and right center for minima or maxima.

By this step, we reduce the number of candidates significantly and obtain representative candidates for group 1. Examples of such candidates, represented by crosses, are illustrated in Figure 10a.

The search for candidates for group 2 also starts at the first minima of the y-relief. Candidates are found, if two minima are detected that fullfill the requirements of group 2, concerning the relative position inside of head, significance of maximum between them and distance be-

tween minima (see Table 1). An example of a x-relief containing a candidate for group 2 is illustrated in Figure 8b. In analogy to the search for candidates for group 1, every pair of minima is assessed and an unsupervised clustering algorithm is applied to reduce the number of candidates. Examples of selected candidates for group 2, represented by small squares, are shown in Figure 10b.



(a) (b) (c)

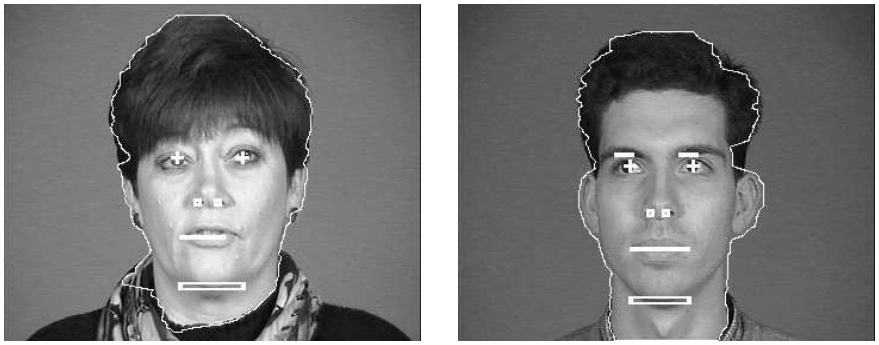
Figure 10: Extracted candidates for (a) group 1, (b) group 2 and (c) group 3

In order to determine candidates for group 3, a full search through all determined x-reliefs is done as well. Group 3 candidates are found by looking for two significant maxima that form the borders of the mouth or chin and meet the requirements of group 3, concerning relative position inside of head, significance of minimum between them and ratio of distance between maxima to head width (see Table 1). An example of a x-relief containing a candidate for group 3 is shown in Figure 8c. In analogy to the search for group 1 candidates, every pair of maxima is assessed and an unsupervised cluster algorithm is applied to select representative candidates. Examples of such candidates, represented by horizontal line segments, are shown in Figure 10c.

2.2.4 Selection of the best constellation

After candidates for group 1, group 2 and group 3 are detected, the best constellation of facial features is determined. For that we build all possible face constellations and assess each of them based on the vertical symmetry of the constellation, the distances between facial features and the assessment of each facial feature candidate. Incomplete face constellations are taken into consideration by allowing void facial feature positions. In order to enforce that more complete face constellations are preferred against less complete face constellations, the assessment of a face constellation is multiplied by a weighting factor, which depends on the number of facial features belonging to the face constellation. According to the assessment of the constellations, the constellations are ranked and the best constellation is chosen.

Results of the detection of facial features for the two example scenes are shown in Fig. 11. Eyebrows are represented by two short horizontal line segment, eyes by crosses, nostrils by small squares, the mouth by a horizontal line segment and the chin by an elongated rectangle.



(a) (b)

Figure 11: Results of facial feature extraction

The facial features are well localized in both examples. In Figure 11a no eyebrows are detected because hair covers the forehead.

2.3 Evaluation of results

To assess the robustness of our approach, we applied our method to the European ACTS project M2VTS database. The database includes 37 different faces containing features like beard, glasses and different facial expressions.

For the segmentation of faces we obtain the results shown in Table 2. The segmentation fails only in one case, in which the correct face candidate is detected. However, due to hair segmentation, it doesn't fullfill the elliptical shape criterion and thus, it is rejected. In the case that a face candidate is detected, it is in all cases detected correctly.

	detected [%]	correctly [%]	falsely [%]
face	97	100	0

Table 2: Detection rates for the segmentation of faces

The results of evaluating the facial feature detection are illustrated in Table 3. Eyebrows are detected in 81% of all cases. Out of this 81%, they are detected correctly in 83%. The eyes are extracted in 94% of the test images and in 74% of them correctly. Nostrils are detected in 94% of the test images and in 97% of them correctly. The mouth is detected in 92% of the frames and in 97% correctly. The chin is extracted in 86% of all cases and in 97% correctly.

These results are rather satisfactory. Detection errors for eyebrows and eyes tend to correlate with each other.

Results for face segmentation and facial feature extraction are shown in Figure 12 and 13. The faces illustrated in Figure 12 are segmented correctly and facial features are located correctly.

Examples for partially false extracted face constellations are shown in Figure 13. For ex-

features	detected [%]	correctly [%]	falsely [%]
eyebrows	81	83	17
eyes	94	74	26
nostrils	94	97	3
mouth	92	97	3
chin	86	97	3

Table 3: Detection rates for facial feature extraction

ample in Figure 13a the eyebrows are detected at the position of the eyes. This problem arises particulary in cases, when only eyes or eyebrows are extracted. In those cases the contextual information doesn't help in distinguishing eyebrows from eyes. Figure 13b shows an example, in which the right eye is located wrongly. This could occur, if feature candidates spread widely around the correct position and, by determining the cluster center, an inexact position is chosen. Further problems in facial feature detection arise because of the hair style or in the case of full beards. Because of the hair style of the person in Figure 13c, a fault in the detection of the eyebrows occurs. The reason for that is the incorrectly measured head width that causes assessment rules, which are defined in dependence on the head width, to fail. Such problems can be solved by a preprocessing step, in which the hair region is segmented out. Problems in chin detection arise, if a person has a full beard (Figure 13d). Such cases have to be handled separately for facial feature detection.

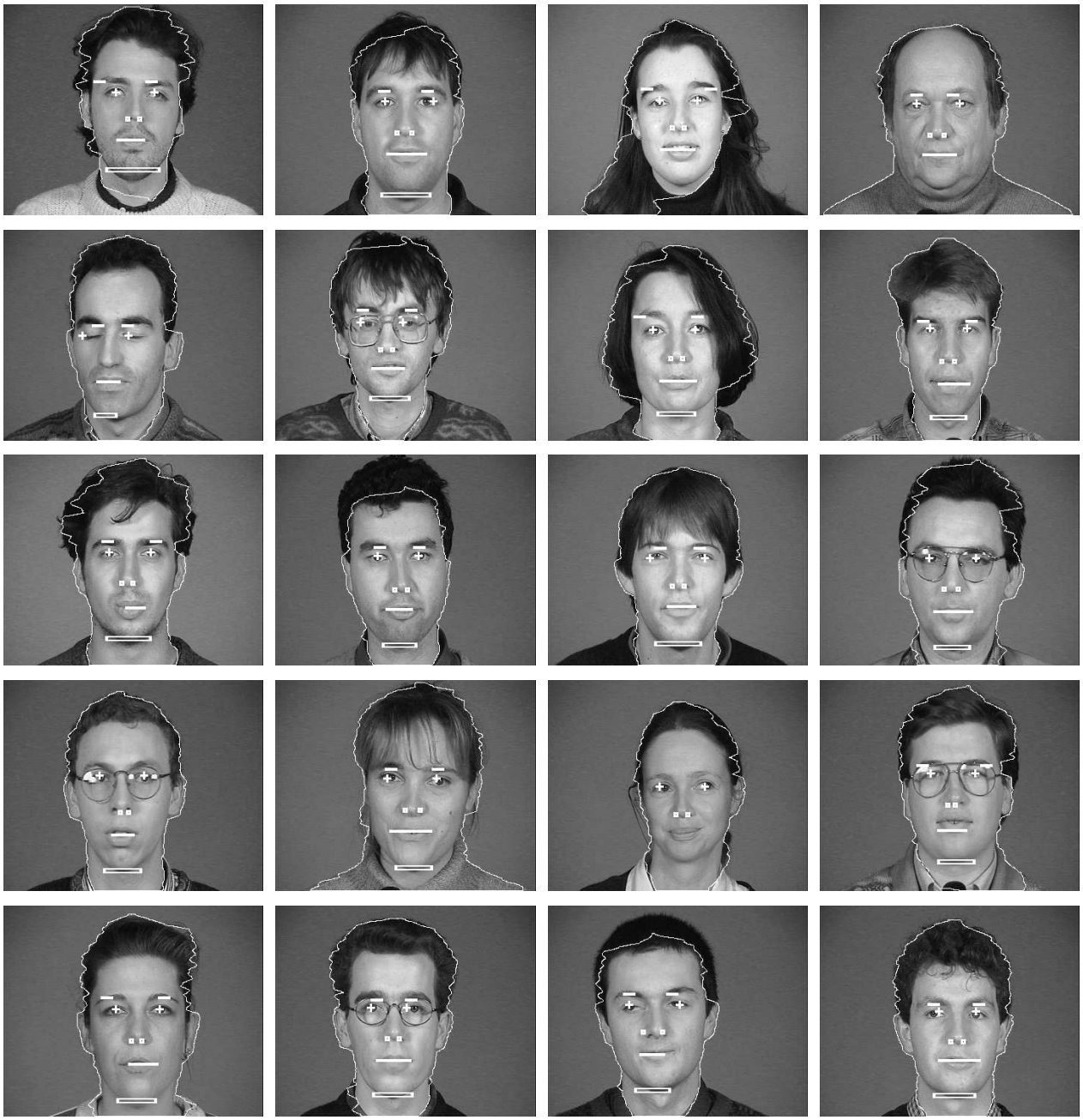


Figure 12: Results of face segmentation and facial feature detection

3 Face tracking

If a face is segmented and facial features are reliably detected, the face contour and facial features can be tracked over time in an image sequence. We perform face contour tracking by using a deformable curve. The exterior forces that influence the curve are defined based on color characteristics. The tracking of facial features is performed based on block matching. By

(a) (b) (c) (d)

Figure 13: Problems in facial feature detection

verifying the content of the block, we ensure the correctness of the block content.

3.1 Face contour tracking

A very efficient method for tracking face contours are active contours, also commonly known as snakes. An active contour is a deformable curve or contour, which is influenced by its interior and exterior forces. The interior forces impose smoothness constraints on the contour and the exterior forces attract the contour to edges, subjective contours or other significant image features. In a first step the snake has to be initialized and then the contour of the object is tracked by placing the snake of image f_t on image f_{t+1}. By minimizing the energy of the snake the best position of the snake in image f_{t+1} is determined.

3.1.1 Snake initilization

As described in section 2.1, we extract the face contour automatically. Thus also our snake initilization is done without interaction. Snake initialization is performed by sampling the face contour at time t into M nodes $v_i = (x_i, y_i)$, $i = 0, .., M-1$, also called snaxels (snake elements). There are different alternatives for the choice of the initial position of snaxels. In our case we have chosen constant Euclidean distance between successive snaxels (Figure 14).

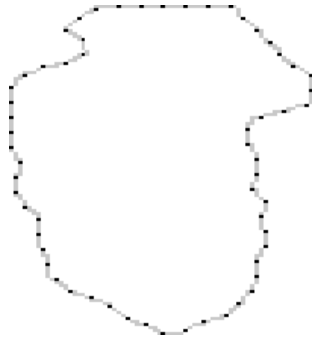


Figure 14: Snake initilization using constant Euclidean distance between snaxels.

3.1.2 Snake energy

An active contour is an energy minimizing curve. In the discrete case, the energy of a snake is defined by

$$E_{snake} = \sum_{i=0}^{M-1} E_{int}(v_i) + E_{ext}(v_i) \tag{9}$$

where E_{int} and E_{ext} denotes the interior and exterior energy terms of the snaxels v_i. The dynamic behaviour of the snake depends on the definition of these energy terms. It is necessary to choose these terms carefully to obtain a stable snake behaviour. By minimizing the energy of the snake, its optimal position is determined.

3.1.3 Interior energy

The interior energy makes the snake resistant to stretching and bending and ensures its smooth behaviour. In the discrete case, it is defined as follows:

$$E_{int}(v_i) = w_1 \cdot \left|\frac{dv_i}{ds}\right|^2 + w_2 \cdot \left|\frac{d^2 v_i}{ds^2}\right|^2 \tag{10}$$

with $v_i = (x_i, y_i)$, $i = 0, .., M-1$. The first-order term $w_1 \cdot \left|\frac{dv_i}{ds}\right|^2$ makes the snake behave like a string. A behaviour like a rod is achieved by the second-order term $w_2 \cdot \left|\frac{d^2 v_i}{ds^2}\right|^2$. The first-order and second-order derivatives can be approximated by finite differences. Thus we obtain

$$\left|\frac{dv_i}{ds}\right|^2 \approx 0.5 \cdot \left(|v_i - v_{i-1}|^2 + |v_i - v_{i+1}|^2\right) \tag{11}$$

$$\left|\frac{d^2 v_i}{ds^2}\right|^2 \approx |v_{i-1} - 2 \cdot v_i + v_{i+1}|^2 \tag{12}$$

The weights w_1 and w_2 regulate the tension and rigidity of the snake. The choice of these weights is critical and it is discussed in detail in [15]. In order to mimic their physical significance, w_1 should be defined as a function of the distance between snaxels and w_2 as a function of the local curvature of a snaxel. However, it is expensive and not trivial to compute and approximate the curvature in the discrete case. Thus Leymarie and Levine recommend to fix w_2 to a small positive constant. In our approach we have decided to define these weights as functions in dependence on predefined values for a natural distance and natural curvature between snaxels. In the case of w_1, we choose the initial Euclidean distance d_{init} as natural distance. The resulting function of w_1 is illustrated in Figure 15a.

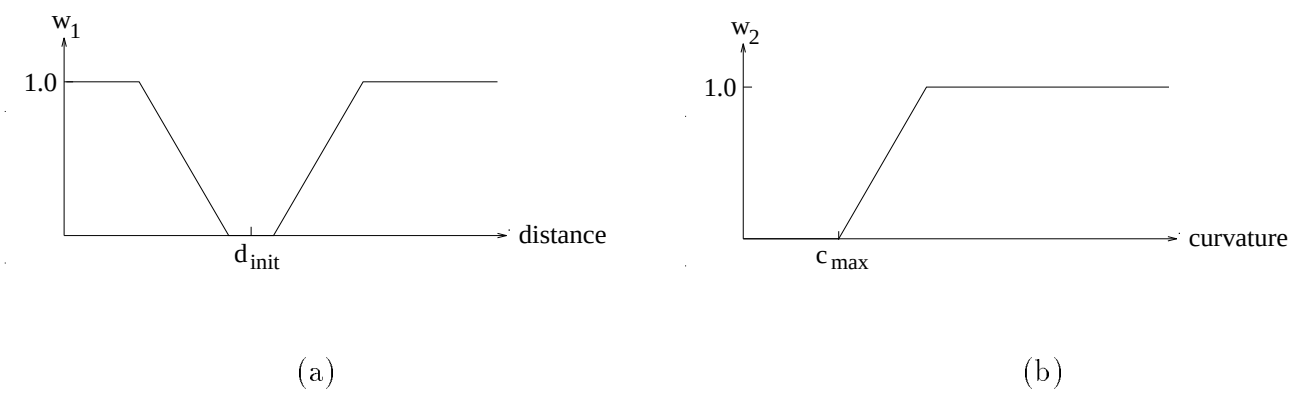


Figure 15: Weighting functions (a) w_1 and (b) w_2.

The weighting factor w_1 has the value 0, if the distance between snaxels is similar to the initial Euclidean distance. Thus, the first-order term of the interior energy becomes 0 and no stretching costs arise. In case that the distance between snaxels is very large or very small, w_1 has the value 1.0 and the full stretching costs occur.

Because the local curvature of an ellipse varies between 0 and a maximal value c_{max}, we define the natural curvature in the interval $[0, c_{max}]$. In general, c_{max} depends on the ratio of minor and major axis of an ellipse. However, we choose this value to be a constant that is appropriate

for the elliptical shape of the face. The resulting function of w_2 is shown in Figure 15b.

3.1.4 Exterior energy

The exterior energy arises from the force that pulls the snake to the significant image features. The features of interest depend on the application. Often edge information is used to adapt the snake to the object contour. In our case, we consider color features that push or pull snaxels perpendicular to the snake contour.

Because the face is more or less a connected component containing skin color, we check for every snaxel if the pixels in its neighbourhood have skin color or not. Pixels with skin color that are outside of the snake pull snaxels outside, pixels with no skin color that are inside of the snake push snaxels inside (Figure 16a). Thus, we define the exterior energy term of a snaxel as follows:

$$E_{ext}(v_i) = -\sum_{(x,y)\in N_{int}(v_i)} 1 - s(x,y) + \sum_{(x,y)\in N_{ext}(v_i)} s(x,y) \tag{13}$$

Hereby $s(x,y)$ denotes the indicator function of skin color which is defined on the basis of the color attributes hue and saturation as described in section 2.1.1. N_{int} and N_{ext} are the interior and exterior neighbourhood of a snaxel (Figure 16b). The directions of these neighbourhoods are computed perpendicular to the contour direction at the snaxel position. The size of N_{int}, respectively N_{ext}, is fixed to be 5×8 pixels.

The behaviour of these exterior forces is similar to the balloon forces that are decribed in [5].

3.1.5 Energy minimization

The interior and exterior forces of the snake are in balance, if the energy of the snake E_{snake} is minimal. In this case, a stable energy state is reached. The minimization of E_{snake} can be done by using the Euler-Lagrange equations of motion. A solution is obtained by using finite

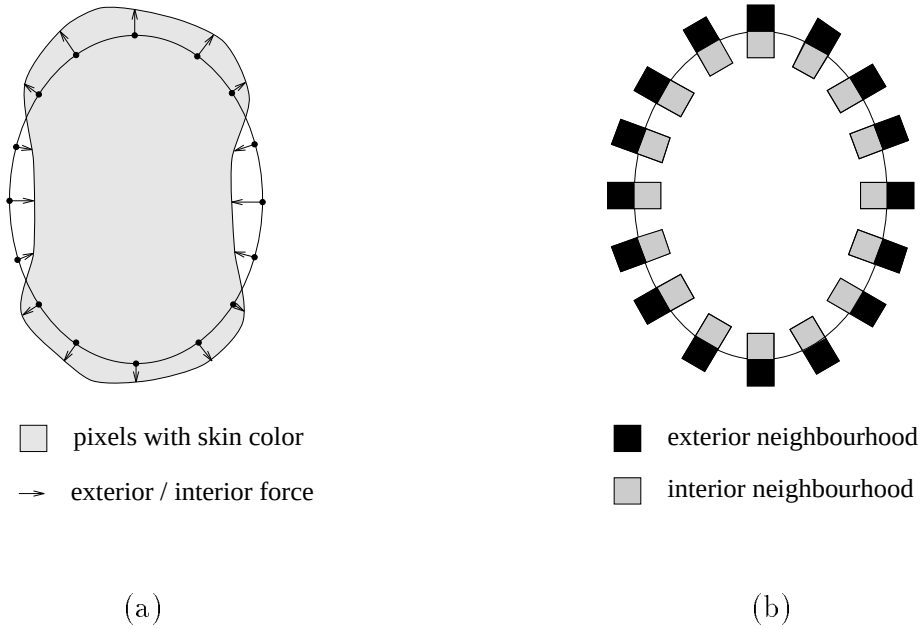


Figure 16: (a) Exterior forces on snaxels and (b) neighbourhood of snaxels.

differences (e.g. [15]) or finite elements (e.g. [5], [19]). An alternative method for determining the energy minimum of a snake is the Greedy algorithm ([18], [24]). We use it in our approach, because of its low computational costs. It uses the fact that the energy of the snake decreases, if the energy of the snaxels decreases. Thus, the minimization problem is solved locally. Each snaxel is moved in its 8-neighbourhood and for each position the sum of interior and exterior energy is determined. The new position of the snaxel is determined based on the local energy minimum. The interior energy term of a snaxel ensures that the snake is not stretched or bended too much. This process is iterated and, after each iteration, E_{snake} is computed. The process stops, when E_{snake} converges.

The main advantage of the Greedy algorithm is its low computational costs. However, by solving the minimization problem locally, the risk arises that snaxels get stuck in local minima. In our case this risk is very small, because we evaluate for each snaxel the color characteristics in a large neighbourhood and thus the exterior forces pull or push the snaxel to a significant minimum.

3.1.6 Addition and deletion of snaxels

In order to obtain a high variability of the snake, we allow the addition and deletion of snaxels. For example in case that a rigid object moves away from the camera, its size decreases and less snaxels are necessary for tracking. On the other hand, if a rigid object moves towards the camera, its size increases and more snaxels are necessary for robust tracking. Obviously, also in the case of tracking non-rigid objects a variable number of snaxels is of advantage. The decision, whether snaxels should be added or deleted, is taken in dependence on the inter-snaxel distance. In the case that the measured Euclidean distance between two snaxels is lower than a minimal threshold, one of them is deleted. If the inter-snaxel distance exceeds a maximal threshold, a new snaxel is added.

3.1.7 Results of face contour tracking

The presented tracking algorithm was tested successfully on a number of image sequences. An example on an image sequence having 150 frames is shown in Figure 17. Although the person moves its head very much, we succeed to track it over the entire sequence. Results for every tenth frame are shown in Figure 17.

3.2 Tracking of facial features

When facial features are reliably detected, they can be tracked over time. Up to now facial feature tracking is only realized for eyes and mouth. But in analogy to the proposed method also tracking of eyebrows, nostrils and chin can be performed.

Robust tracking is possible using block matching. First, initial blocks have to be extracted. Then they can be tracked over time by searching for corresponding blocks in consecutive frames. To ensure that the reference blocks always contain facial features, we verify the block contents

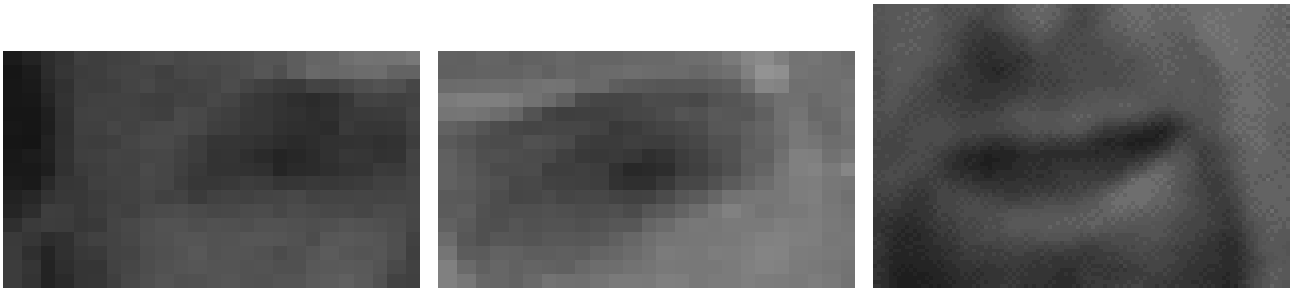
Figure 17: Face contour tracking

by minima analysis. Block matching for facial feature tracking is also used in [6].

3.2.1 Initialization of facial feature blocks

The initial block size is defined in case of the eyes dependent on the eye distance and, in case of the mouth, dependent on the mouth width. Let us denote the block width and block height of a feature block B by b_x and b_y. Then the initial block for a feature is defined by placing the center of the block frame with dimension b_x and b_y on the detected feature position (x, y) at time t. Examples of initial feature blocks for left eye, right eye and mouth are illustrated in

Figure 18.



(a) (b) (c)

Figure 18: Initial feature blocks for (a) left eye, (b) right eye and (c) mouth

3.2.2 Matching of facial feature blocks

When initial feature blocks at time t are extracted, the position of features at time $t+1$ can be determined by matching the blocks of time t to the frame of time $t+1$ (Figure 19). Because the displacement between two consecutive frames is restricted due to video rate, the search for corresponding blocks can be restricted to a search region of dimension s_x and s_y centered at the feature position (x, y) at time $t+1$ (Figure 19b).

For each position of the search region a similarity measure between reference block and test block is evaluated. In our case we use the sum of absolute greylevel differences as block difference BD. For a displacement (u, v) between reference block and test block, the block difference at position (x, y) results in:

$$BD_{x,y}(u, v) = \sum_{i,j \in B_{x,y}} |f(i, j, t) - f(i+u, j+v, t+1)| \tag{14}$$

where $f(i, j, t)$ denotes the greylevel value at position (i, j) at time instance t. The best match position is found, if the block difference is minimal.

Matching results are illustrated in Figure 20. The search regions for corresponding blocks

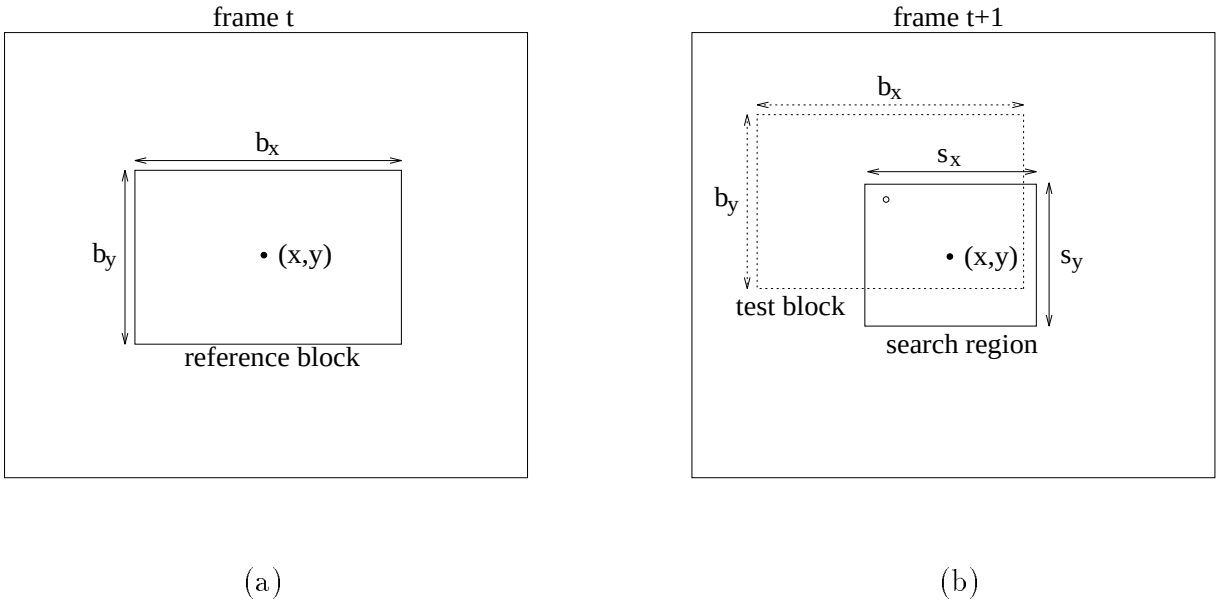


Figure 19: Block matching (a) frame t with reference block and (b) frame $t+1$ with search region and one of the test blocks

are shown for eyes and mouth. The greylevel values inside of the search regions represent the quality of block matching: the higher the greylevel value the better the matching.

Figure 20: Results of block matching

3.2.3 Refinement of best-match position

After performing block matching we refine the best-match position by minima analysis. In analogy to the extraction of facial features, we determine minima in the y-relief of the best-

match block and for each minima, we determine minima in the related x-reliefs. After minima localization we select the minimum that is closest to the best match position and consider its positon as refined feature position. By this refinement step we ensure that the best-match block at time t still contains a facial feature as content.

3.2.4 Updating of facial feature blocks

Facial feature block updating is critical for tracking. Once a fault occurs, the whole tracking process collapses and a new face recognition is necessary. Thus, it has to be ensured that block updating is done correctly. In our approach, we ensure this by evaluating the best-match position and the refined feature position. If the distance between best-match position and refined position is small, we define a new reference block at the refined feature position. Otherwise, we consider the match as uncertain and the reference block remains constant.

3.2.5 Results for facial feature tracking

Results of facial feature tracking for an image sequence consisting of 150 frames are shown in Figure 21. For every tenth frame the search regions for facial features and the results of block matching are shown.

The reference blocks extracted for the left eye, right eye and mouth are illustrated in Figure 22. Though eyes and mouth close and open during tracking, reference blocks contain always the right content.

4 Conclusion

In this framework we have presented a fully automatic approach for face segmentation, facial feature extraction and tracking of the face contour and facial features over time.

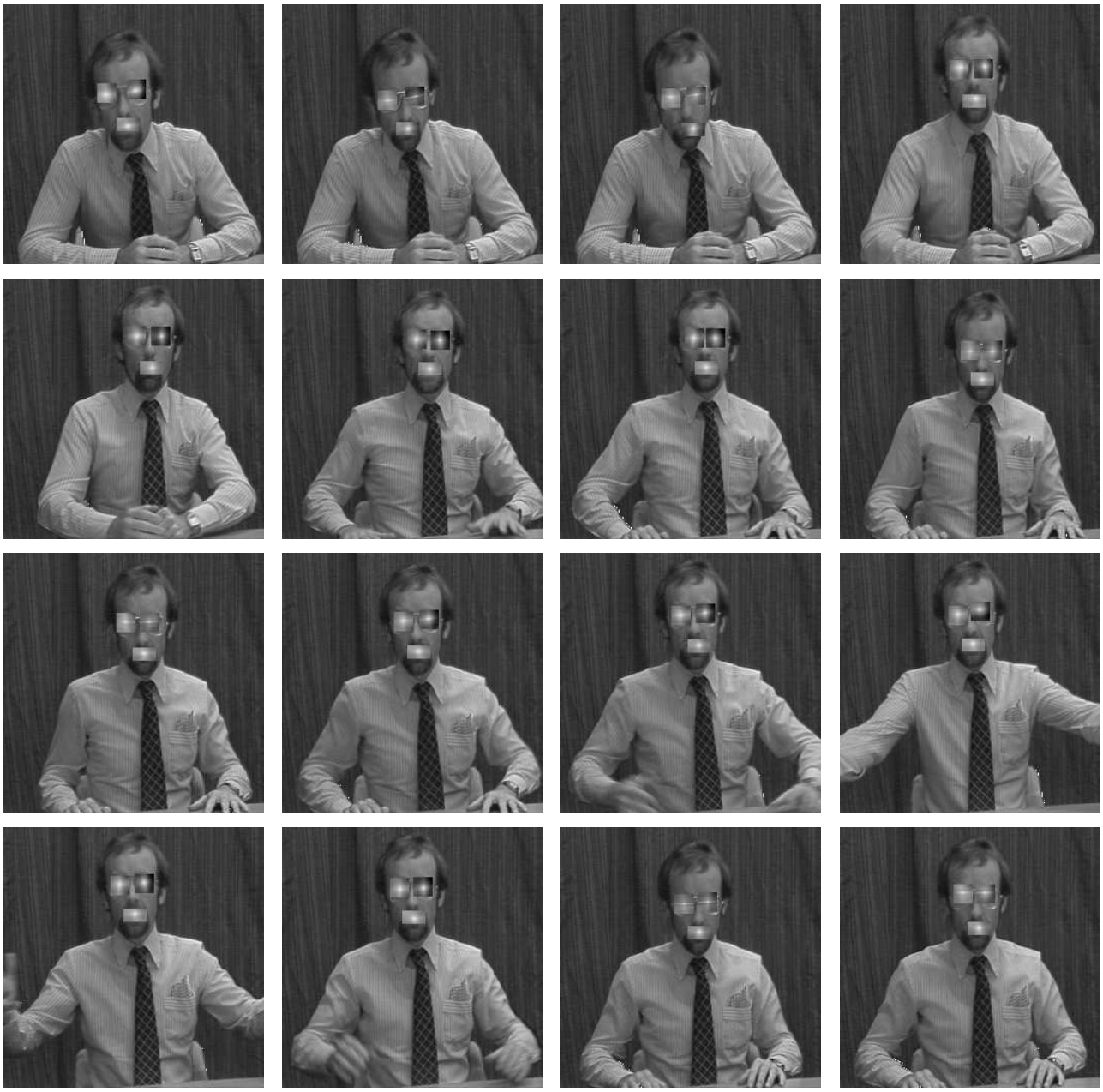


Figure 21: Results of facial feature tracking: Search regions for corresponding feature blocks and quality of block matching represented as greylevel intensity inside the search regions (the smaller the block difference the higher the greylevel value).

Face segmentation is done by using color (HSV) and shape information. First skin-like regions are segmented based on hue and saturation information and then we check for each of the regions, if they have elliptical shape or not. Skin like regions with elliptical shape are considered as face candidates and are verified by searching for facial features in their interior.

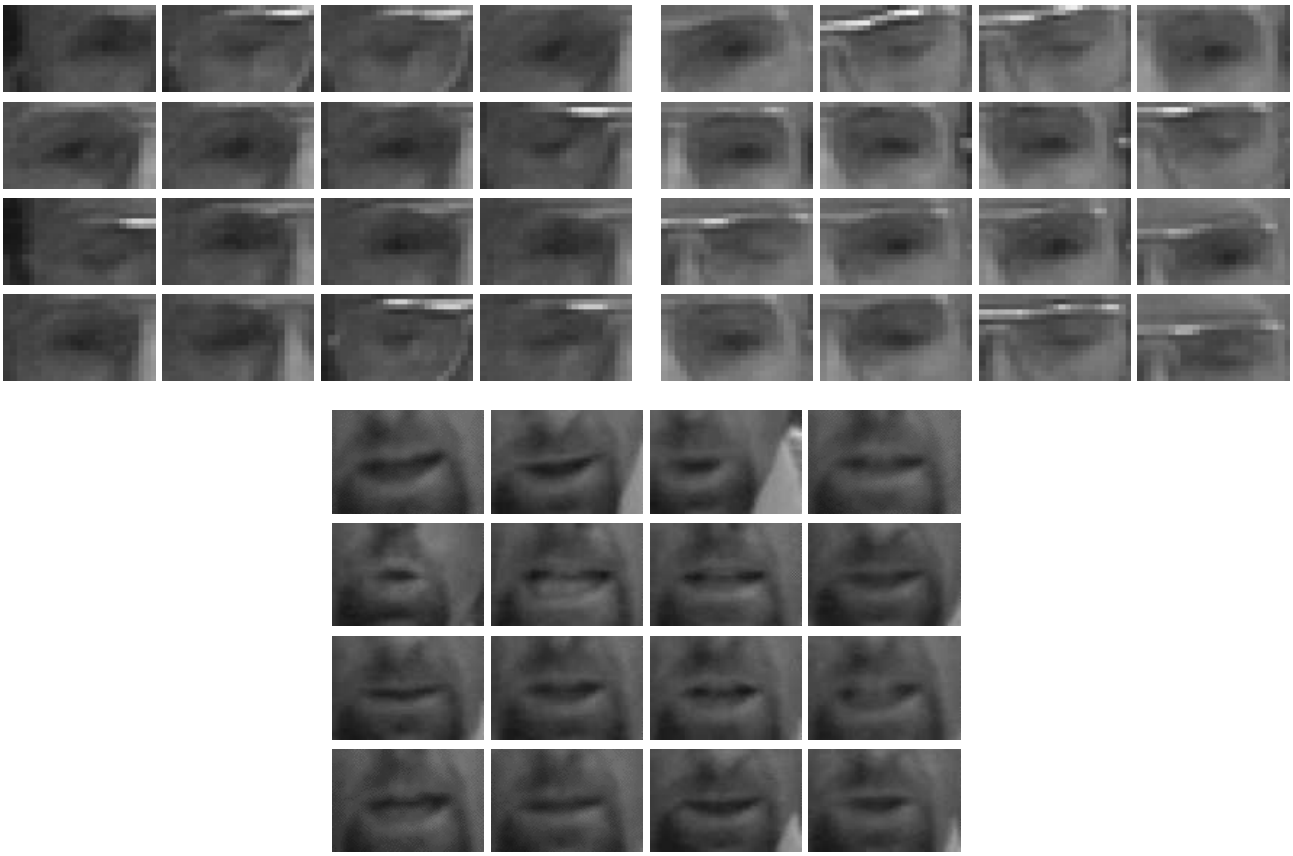


Figure 22: Reference blocks for left eye, right eye and mouth during tracking.

As interesting facial features we consider eyebrows, eyes, nostrils, mouth and chin. Facial feature extraction is based on the observation that facial features differ from the rest of the face because of their low brightness. Thus we extract facial feature candidates by evaluating the topographic greylevel relief of the face candidates. The best face constellation is chosen based on vertical symmetry, distances between facial features and assessment of each facial feature. Incomplete face constellations are considered as well.

Once faces are segmented and facial features are detected, they can be tracked over time. Tracking of the face contour is performed by using a deformable curve. The exterior forces on the curve are defined based on color characteristics. Facial feature tracking is done by block matching. If the best-match position is found, we check if there is a minimum close to the best-match position. If this is the case, we refine the facial feature position and define a new

reference block. Otherwise, we consider the match as uncertain and the reference block for the facial feature remains unchanged.

The robustness of our approach was tested on 38 different color image sequences containing faces. The results of this evaluation are very satisfactory.

References

[1] M. Acheroy, C. Beumier, J. Bigün, G. Chollet, B. Duc, S. Fischer, D. Genoud, P. Lockwood, G. Maitre, S. Pigeon, I. Pitas, K. Sobottka, and L. Vandendorpe. Multi-modal person verification tools using speech and images. In *European Conference on Multimedia Applications, Services and Techniques*, pages 747–761, Louvain-La-Neuve, Belgium, May 28-30 1996.

[2] M. Bichsel, editor. *International Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 26-28 1995. IEEE Computer Society, Swiss Informaticians Society et al., MultiMedia Laboratory, Department of Computer Science, University of Zurich.

[3] M.C. Burl, T.K. Leung, and P. Perona. Face localization via shape statistics. In *International Workshop on Automatic Face and Gesture Recognition*, pages 154–159, Zurich, Switzerland, June 26-28 1995.

[4] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, May 1995.

[5] Laurent D. Cohen and Isaac Cohen. Finite-element methods for active contour models and balloons for 2-D and 3-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1131–1147, November 1993.

[6] J. Coutaz, F. Bérard, and J.L. Crowley. Coordination of perceptual processes for computed mediated communication. In *International Conference on Automatic Face and Gesture Recognition*, pages 106–111, Killington, Vermont, USA, October 14-16 1996.

[7] Y. Dai and Y. Nakano. Extraction of facial images from complex background using color information and SGLD matrices. In *International Workshop on Automatic Face and Gesture Recognition*, pages 238–242, Zurich, Switzerland, June 26-28 1995.

[8] A. Eleftheriadis and A. Jacquin. Automatic face location, detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates. *Signal Processing: Image Communication*, 7(3):231–248, Jul 1995.

[9] H. Ernst. *Introduction into digital image processing (in German)*. Franzis, 1991.

[10] G. Galicia and A. Zakhor. Depth recovery of human facial features from video sequences. In *IEEE International Conference on Image Processing*, pages 603–606, Washington D.C., USA, October 23-26 1995. IEEE Computer Society Press, Los Alamitos, California.

[11] H.P. Graf, T. Chen, E. Petajan, and E. Cosatto. Locating faces and facial parts. In *International Workshop on Automatic Face and Gesture Recognition*, pages 41–46, Zurich, Switzerland, June 26-28 1995.

[12] H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan. Multi-modal system for locating heads and face. In *International Conference on Automatic Face and Gesture Recognition*, pages 88–93, Killington, Vermont, USA, October 14-16 1996.

[13] IEEE Computer Society Press, Los Alamitos, California. *International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, USA, October 14-16 1996.

[14] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.

[15] Frédéric Leymarie and Martin D. Levine. Tracking deformable objects in the plane using an active contour model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):617–634, June 1993.

[16] Y. Li and H. Kobatake. Extraction of facial sketch images and expression transformation based on FACS. In *IEEE International Conference on Image Processing*, pages 520–523, Washington D.C., USA, October 23-26 1995. IEEE Computer Society Press, Los Alamitos, California.

[17] H. Niemann. *Pattern Analysis and Understanding*. Springer Verlag, 1990.

[18] Urthe Pautz. Active contours for 3D-segmentation of magnetic resonance images (in German). Diploma Thesis, Bavarian Research Center for Knowledge Based Systems, Erlangen, Germany, July 1995.

[19] Alex Pentland and Stan Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):715–729, July 1991.

[20] I. Pitas and A.N. Venetsanopoulos. *Nonlinear Digital Filters: Principles and Applications*. Kluwer Academic Publishers, 1990.

[21] K. Sobottka and I. Pitas. Localization of facial regions and features in color images. In *4th Open Russian-German Workshop: Pattern Recognition and Image Analysis*, pages 134–138, Valday, The Russian Federation, March 3-9 1996.

[22] K. Sobottka and I. Pitas. Segmentation and tracking of faces in color images. In *International Conference on Automatic Face and Gesture Recognition*, pages 236–241, Killington, Vermont, USA, October 14-16 1996.

[23] K. Sobottka and I. Pitas. Looking for faces and facial features in color images. *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Russian Academy of Sciences*, 7(1), 1997.

[24] D.J. Williams and M. Schah. A fast algorithm for active contours and curvature estimation. *CVGIP: Image Understanding*, Bd. 55(1):14–26, Jan 1992.

[25] H. Wu, Q. Chen, and M. Yachida. An application of fuzzy theory: Face detection. In *International Workshop on Automatic Face and Gesture Recognition*, pages 314–319, Zurich, Switzerland, June 26-28 1995.