ELSEVIER

Signal Processing: *Image Communication* 7 (1995) 57–74

SIGNAL PROCESSING: IMAGE COMMUNICATION

# Facial feature localization and adaptation of a generic face model for model-based coding

M.J.T. Reinders[a,*], P.J.L. van Beek[a], B. Sankur[b], J.C.A. van der Lubbe[a]

[a] *Department of Electrical Engineering, Information Theory Group, Delft University of Technology, Delft, The Netherlands*
[b] *Department of Electrical–Electronic Engineering, Boğaziçi University, Bebek, Istanbul, Turkey*

Received 26 January 1994

**Abstract**

A method for the adaptation of a generic 3-D face model to an actual face in a head-and-shoulders scene is discussed, with application to video-telephony. The adaptation is carried out both on a global scale to reposition and resize the wire-frame, as well as on a local scale to mimic individual physiognomy. To this effect a hierarchical scheme is developed to extract the semantic features in the head-and-shoulders scene, such as silhouette, face, eyes and mouth, using a knowledge-based selection mechanism. These algorithms, which are to be an integral part of a general model-based image coder, are tested on typical videophone sequences.

*Keywords*: Model-based image coding; Knowledge-based segmentation; Facial model adaptation

## 1. Introduction

Model-based coding is a newly emerging image sequence compression technique [1, 7, 17]. In contrast to conventional image compression schemes that exploit pixel to pixel correlations, model-based coding takes a more global view of objects and their specific 3-D representations, and it relies on a priori knowledge about the scene. The basic assumption in this coding technique is that the expected scenes are known and these are constrained to a few world objects, like a speaker's head.

Model-based coding adapts generic models of objects to actual objects encountered in the scene, and then tracks their evolution and changes throughout the sequence. In this way objects in an image sequence can be parsimoniously coded, as the information to be transmitted to graphically re-enact the sequence consists simply of the changes in the model parameters.

The block diagram of such a model-based coding scheme is illustrated in Fig. 1 for the specific case of head-and-shoulder scenes. As shown in Fig. 1 a generic face model (e.g. a wire-frame model) is present at both the receiving and transmitting sides. This generic face model is adapted during the initial frames of a sequence to the physiognomy, that is to the actual features of the face in the scene. The adaptation is based on the information extracted

* Corresponding author. Tel: 31 15 783084. Fax: 31 15 781843. E-mail: marcel@it.et.tudelft.nl.


*SSDI* 0923-5965(94)00042-5

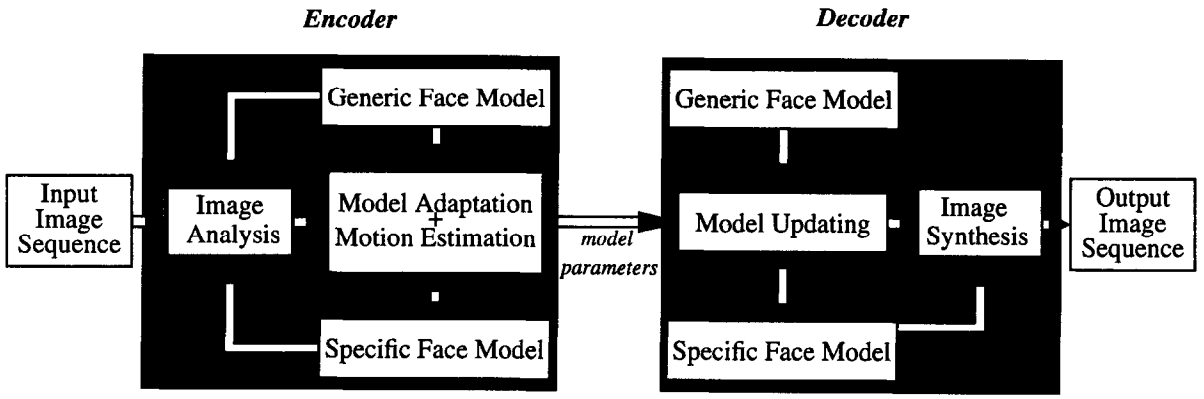



Fig. 1. Model-based image coding scheme.

from the scene frames by using image analysis techniques (see Fig. 1). The model becomes a specific model when it is adapted to the face in the scene. As the videophone conversation proceeds, the scene object is tracked, so that both the global motion parameters of the face as well as the local deformations corresponding to facial expressions are mimicked. The estimates of motion and expression parameters are transmitted to the receiving side. At the receiver side the model is continuously adapted and animated with the incoming updates. In addition to the motion and expression data, additional data for texture and illumination changes may also be needed. The final rendition of the specific model after the model updates is implemented in the image synthesis box.

The fundamental assumption in these techniques is that the scenes are constrained to a few objects, like a speaker's head, for which a priori models can be developed. The flexibility and accuracy of the models is crucial to obtain high compression ratios with realistic reproductions of the scenes. If a scene change occurs which is outside the range of model objects, one can always fall back to conventional coding techniques, e.g. CCITT's H.261 [5]. However, problems caused by scene changes fall outside the scope of this paper. The implications of the model-based coding schemes will be, however, beyond video-telephony. In fact these techniques may turn out to be more relevant in the context of automatic answering machines, graphic animation, archival search for human faces, etc.

Comparing a model-based coder to current statistical coders, like the CCITT H.261, one notices that while the compression ratio increases [1, 7], the complexity of the image processing tasks also increases. The main tasks of a model-based coder consist of: (i) detection and localization of semantic features to supply initial information for the remaining tasks of the coder; (ii) adaptation of the 3-D generic model to the actual object, i.e. scaling and posing of the generic model [1, 6], and local adaptations to reflect the individual physiognomy [1, 3, 17]; and (iii) tracking of motion parameters [1, 6, 7] as well as expression parameters [8, 10, 11, 15].

Evidently the first task is crucial for a model-based coder because all other tasks depend on it. Pioneering work on the localization of facial feature points has been done by Kanade [9]. Recently the automatic extraction of these features has received more attention as part of face recognition systems, e.g. in [14] a number of these techniques are listed, However, all these methods at some point assume either a fixed setting of the background or they are scale/rotation dependent. The limitations of these methods indicate that extracting the shape of the facial features from the image directly – without any prior knowledge – is not practical. In any case it is imperative to obtain an accurate estimate of feature positions before their shape can be extracted.

We have focussed on typical videophone scenes, i.e. on the head and shoulders of one speaker in

front of a still camera. The main goal of our research is the automatic adaptation of a generic face model to the face in the scene. In particular we are looking at the automatic detection of facial features. The contributions of our work against the background of previous research results can be summarized as follows: (i) A robust method for the automatic localization of semantic features has been developed using a knowledge-based selection mechanism. This scheme does not presuppose a fixed setting, and imposes restrictions only on extreme rotations that preclude visibility of certain facial features. (ii) A method to adapt a generic face model to the extracted facial contours. This method first transforms the generic model globally to account for scaling and posing and then refines the mismatch between the generic model and the facial contours to account for individual physiognomy by applying a local adaptation based on a graph matching algorithm [4].

This paper is organized as follows. Section 2 addresses the problem of detection and localization of features of interest on human faces, such as occluding silhouette, face, mouth and eyes. Given these localized features, the global adaptation of the model as well as a more refined local fitting are described in Section 3. Experimental results and performance figures obtained from typical videophone test sequences are discussed in Section 4, and concluding remarks are given in Section 5. In the following, we have used lowercase boldface italic characters to denote vector variables, and the symbol '$\sim$' stands for 'is directly proportional to'.

## 2. Localization of facial features in head-and-shoulder scenes

The localization of facial features is based on the propagation of knowledge about these features. This concept, called the knowledge-based selection mechanism, leads to a hierarchical localization scheme in which each feature is pinpointed sequentially. The first feature which is localized is the one which can be localized reliably based on a priori knowledge solely.

For each of these features, first, candidate regions are generated by a segmentation of the image.

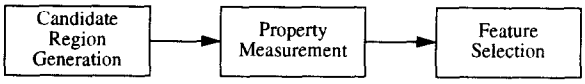



Fig. 2. Consequent steps when localizing a facial feature.

From this set of candidate regions, the region which best matches the feature searched for is selected based on the knowledge-based selection mechanism (Appendix A), illustrated in Fig. 2. According to this mechanism one can predict the values of certain properties of a feature based on a priori knowledge about that feature and previously localized features. For example, eyes are searched for within the previously located face region, and the selection mechanism uses a priori knowledge about face sizes and eye-to-eye distances. Thus, to select the most likely goal region from a set of candidates, one selects that region whose current property values match closest the predicted ones. The properties used should have high discriminatory power, and should be invariant to the allowed rotations, translations and zoomings.

To be able to perform these tasks, the following basic assumptions about head-and-shoulder sequences are made: (i) There exists only one moving object in the foreground consisting of the head and shoulders of a talking person, while the background is stationary. (ii) The motion of the speaker is moderate with respect to the frame rate. (iii) Certain facial features are always visible, in other words, head rotations and tilts that impede the visibility of eyes and mouth are precluded; similarly the face is not occluded by other objects such as gesticulating hands. (iv) The head inclination (rotation around the $z$-axis) should be less than 45°. (v) The human face has an approximate vertical symmetry.

In the following sections, the localization of each semantic feature of interest in the head-and-shoulder scene is described in more detail. The hierarchical approach starts with localizing the silhouette in the image, followed by the head, the face and finally the eyes and mouth. Some of the basic ideas and preliminary results were also presented in [2]. Here an extended and more refined system will be presented, which has an improved performance while localizing more facial features.

### 2.1. Localization of the silhouette

Based on the above assumptions, the occluding silhouette of the head and the shoulders can be extracted quite easily from thresholded frame differences. The motion of the speaker induces frame differences, which are smoothed to obtain connected regions by exploiting their spatial correlations. Often, however, evidence about moving objects in the scene gathered from a single frame difference may not suffice to portray a speaker in its entirety. It is then necessary to recover fully the speakers silhouette by observing a number of successive change detection masks. Because the motion of the speaker is generally slow with respect to the frame rate, the sequence of change detection masks are temporally correlated, and this can be used to improve the silhouette.

*Candidate region generation*

A block diagram describing the segmentation of silhouette regions for a moving speaker is illustrated in Fig. 3, and it consists of the following steps:

(i) The frame differences are spatially low-pass filtered with a uniform filter and downsampled by a factor of 4.

(ii) A threshold is calculated from the histogram of absolute frame differences. Most techniques try to find the best threshold value assuming a bimodal histogram. However, histograms of frame differences are typically unimodal, with a peak close to zero. Zack et al. [19] developed a technique which assumes unimodal histograms. This thresholding method resulted in good and sufficiently consistent silhouette regions from the frame difference images.

(iii) The sequence of smoothed and thresholded change detection masks, $b_t(i,j)$, are processed with a spatiotemporal filter ($t$ stands for time). The function of the spatiotemporal filter is to fuse a number of consecutive masks. The temporal fusing between previous change decisions and new change evidences is implemented as

$$c_t(i,j) = \max\{Nb_t(i,j), f_{t-1}(i,j) - 1\}, \quad (1)$$

where $b_t(i,j) \in \{0,1\}$, $f_t(i,j) \in \{0,1,2,\ldots,N-1,N\}$ and $f_0(i,j) = 0$. Also, these evidences undergo smoothing via median filtering:

$$f_t(i,j) = \text{Median}\{c_t(i,j)\}. \quad (2)$$

The scheme to determine $c_t(i,j)$ can be thought of as a counter, which starts at $N$ when the most recent change evidence $b_t(i,j)$ equals 1. Otherwise, if there is no change at pixel location $(i,j)$ at instance $t$, previous change decisions, $f_{t-1}(i,j)$, are not discarded immediately, but the counter is decremented by one. An illustration of fused masks obtained from the Miss America sequence is given in Fig. 4(a) and (b).

The parameter $N$ now controls how long frame information is retained when it is not refreshed. For example, keeping $N$ large will tend to create connected and smooth foreground blobs, at the expense of smearing the silhouette, especially if the object motion is excessive. In our experiments, $N = 10$ was found to be adequate, although this parameter may also be determined adaptively, e.g. considering some norm of motion vectors.

*Feature selection*

Although this algorithm always yields a large connected region approximately portraying the silhouette of the speaker, it is possible that other smaller regions are found as well, due to noise or fragmentation of the silhouette region (e.g. regions

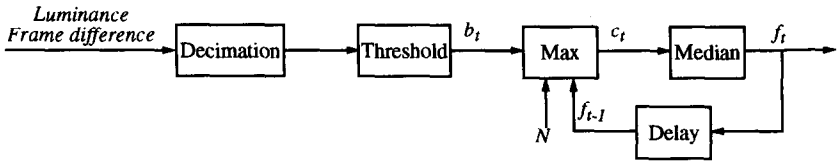



Fig. 3. Block diagram describing the generation of candidate silhouette regions.

(a) (b)

Fig. 4. Dynamic segmentation images obtained using past and present change detection masks: (a) sequentially fused images of Miss America sequence of frames 70 to 90; (b) the resulting candidate regions for the silhouette of Miss America in frame 90.

in the hair). Therefore, the region with the *largest area* is always taken to be the silhouette.

In principle a more accurate silhouette can be obtained with the application of motion compensation on the change masks. However, the accuracy of the silhouette algorithm without motion compensation was found satisfactory for the subsequent steps of the algorithm, hence this simpler version was preferred.

### 2.2. Localization of the head

The head region can be separated from the shoulder region by noticing that the silhouette contour has always a pair of concavities at the neck. The position of these concavities can be identified simply by first finding the convex hull of the earlier localized silhouette region, and then marking the place where the silhouette and hull contours are most distant from each other. The search along the contour should start from the top of the silhouette region, and proceed downwards on both sides, keeping track of the minimum of the hull-to-silhouette distances. The coordinates of the maximum of these distances are identified as the neck points, $\boldsymbol{n}_r$ and $\boldsymbol{n}_l$, as illustrated in Fig. 5(a). Notice that one should travel down from the top not more than half the contour lengths, in order not to get confused by the ambiguous and fragmented region of the shoulders. Finally, the chin contour is roughly approximated by a circle with its midpoint at the center of the two neck points $(\boldsymbol{n}_r, \boldsymbol{n}_l)$, and the diameter equal to the distance between the neck points. Fig. 5(b) shows the thus extracted head from the silhouette found in Fig. 4(b).

### 2.3. Localization of the face

Analyzing typical head-and-shoulder images, one finds that the facial region, especially at low resolution, exhibits a uniform color. Hence, the facial region can be extracted by segmenting the image by using a region growing technique based on the color components, i.e. luminance ($y$) and chrominance ($u, v$) components.

*Candidate region generation*

A block diagram describing the generation of candidate regions for the face is illustrated in Fig. 6, and consists of the following steps.

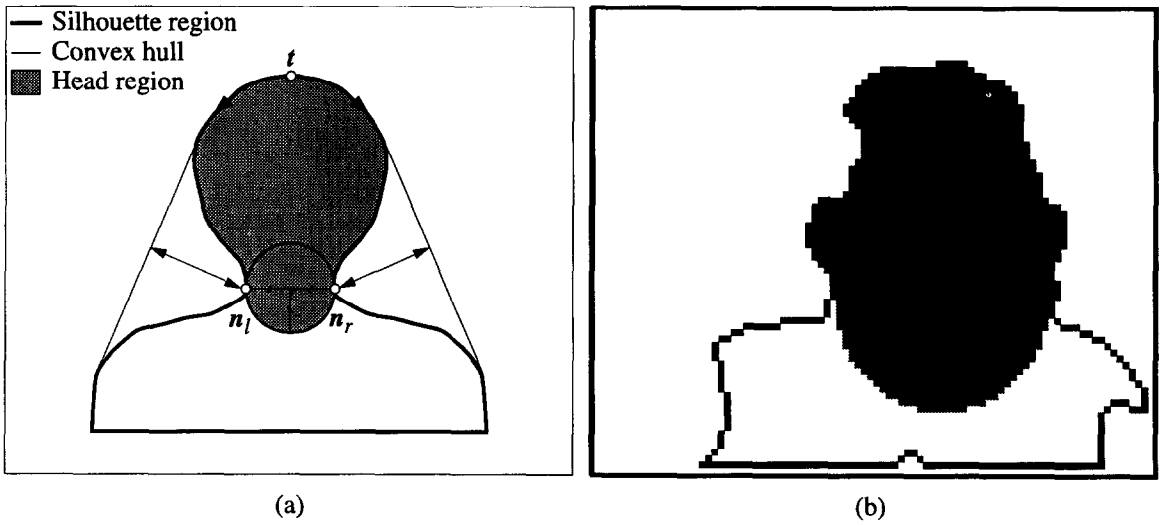



Fig. 5. (a) The extraction of the head region from the silhouette region. (b) The head region extracted from the silhouette region in Fig. 4(b).

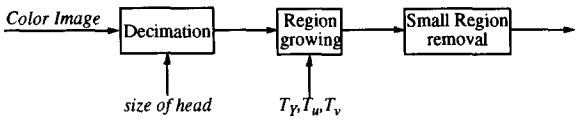



Fig. 6. Block diagram describing the generation of candidate face regions.

(i) The image is low-pass filtered with a uniform filter and downsampled repetitively until the size of the head region is reduced down to an image approximately $32 \times 32$ pixels in size.

(ii) The downsampled image is segmented with a region growing algorithm based on the three color features and the chessboard distance. At present we use a serial scheme, whereby every region is initiated by the first 'non-region' pixel, that is the first pixel that does not belong to any previously found region, and it is then grown to its completion. After experimenting with different color spaces, e.g. YUV, RGB, HSV, the YUV color space gave the best results for our test images. Also a weighted Euclidean distance in feature space was considered but the chessboard distance produced better results. The luminance is taken into consideration to aid in the segmentation of those areas where the chrominance is not well defined, like for example the hair. Note that larger variations are allowed in the luminance component of a region than in the chrominance component. The thresholds for each component are kept fixed, in our experiments we have used the following thresholds: $(T_y, T_u, T_v) = (40, 10, 10)$.

(iii) Finally, regions smaller than 10 pixels are not taken into consideration and removed.

*Feature selection*

Frame number 79 of the Miss America sequence thus segmented is shown in Fig. 7(a), which exemplifies that in general several candidate regions may emerge. Therefore, from this map, the region which resembles the face most should be selected. In order to accomplish this task the knowledge-based selection mechanism as detailed in Appendix A is used. The selection proceeds by computing a score for each region, which in turn can be interpreted as the likelihood that the region corresponds to the face. The score calculation uses a set of region properties chosen to best differentiate the true facial region from the other regions. The properties that proved to discriminate well are: the *distance between the head and face centroids* and the *face area size*. The combination of these two properties is denoted by

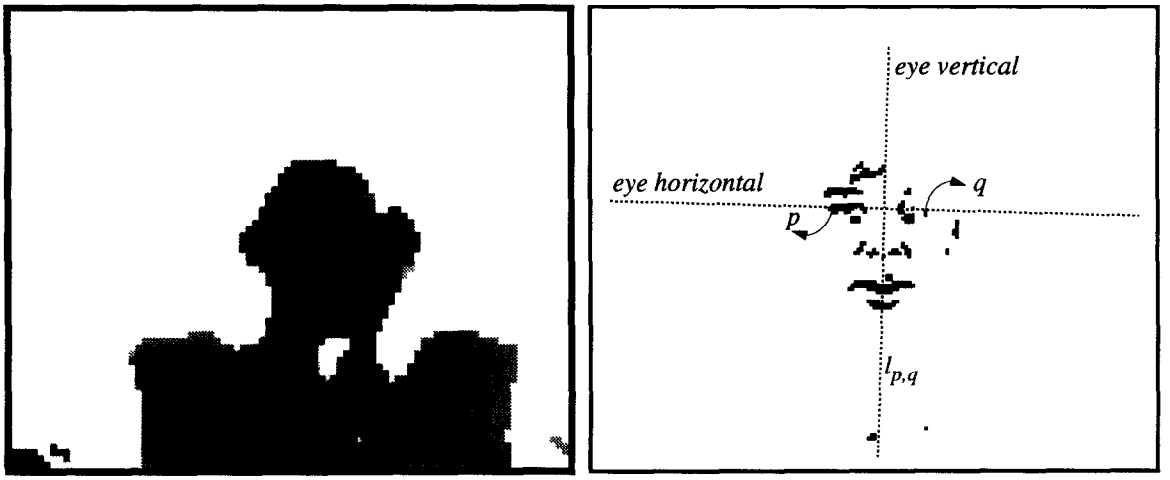



Fig. 7. (a) The generated candidate facial regions for frame 79 of Miss America. (b) The generated candidate eyes and mouth regions for the same frame of Miss America. Also, as an example, the eye vertical ($l_{p,q}$) and eye horizontal given two candidate eye pair regions $r_p$ and $r_q$ are shown.

the vector $\boldsymbol{m}_j$ for the region $r_j$. The face score, $s_j$, of the region $r_j$ can then be calculated as (see Appendix A)

$$s_j = P(\text{Face}|\boldsymbol{m}_j) = \frac{P(m_{j,\text{dist}}|\text{Face})}{P(m_{j,\text{dist}})}\frac{P(m_{j,\text{area}}|\text{Face})}{P(m_{j,\text{area}})}$$

$$\sim \exp\left[-\frac{\| C_j - C_{\text{head}} \|^2}{2(0.18\,\text{Width}_{\text{head}})^2}\right]$$

$$\times \exp\left[-\frac{(\text{Area}_j - \text{Area}_{\text{head}})^2}{2(0.35\,\text{Area}_{\text{head}})^2}\right]. \quad (3)$$

In other words, the centroid of the facial region is expected to coincide with that of the head region with an allowed deviation of 18% of the head width, and the area size is expected to be proportional to the area size of the head region. Allowing deviations from the expected values accounts for variations in physiognomy and minor segmentation errors due to noisy data. Note that, by relating these expectations to characteristics of the head region, relative scale independence is maintained.

Finally, a heuristic rule is needed since the region corresponding to the hair can sometimes have approximately the same score as the face region. According to this rule the face region is always the lower one if there are two regions that closely match the expected property values.

### 2.4. Localization of the eyes and mouth

One can observe that the eyes and mouth appear in intensity images as small dark areas surrounded by brighter areas. If distinguishable, then the pupils and eyelashes are always darker than the surrounding areas. The eye sockets strengthen this effect due to their shadowing, in fact even with closed eyelids the low intensity zones of the eye sockets are sufficient to delimit the eyes. Finally, eyebrows, nostrils and lips and/or a mouth ajar create also darker regions. These observations hold true over a wide range of lighting conditions and head orientations with respect to the camera.

*Candidate region generation*

We have used these considerations to localize the eyes and mouth in head-and-shoulder images. A block diagram describing the generation of candidate regions for eyes and mouth is shown in Fig. 8 and consists of the following steps:

(i) The eyes and mouth are searched only within the convex hull of the localized face region. The

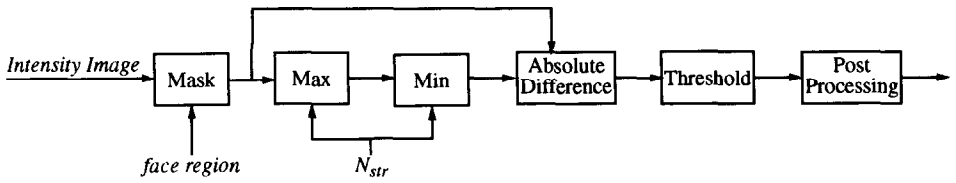



Fig. 8. Block diagram describing the generation of candidate eyes and mouth regions.

convex hull is taken because the localized facial region can contain 'fjords' due to locks of hair occluding parts of the front and the eyes.

(ii) The small dark areas in a brighter surrounding are enhanced via gray scale morphological operators, e.g. max–min filters. The max–min filter is implemented with a circular structuring element of size $N_{str}$, which is estimated from the size of the localized facial region. For example, in the Miss America sequence $N_{str}$ always equals 7. The max–min filters calculate the upper envelope of the gray scale landscape. The local minima can now be enhanced by taking the absolute difference between the upper envelope and the original image.

(iii) The image containing the enhanced local minima is thresholded using the well-known 'isodata' thresholding algorithm yielding a binary map which reveals the local minima.

(iv) Finally, the binary map is post-processed to remove erroneously candidate regions. To remove small regions due to noise, regions smaller than 0.075% of the face area are removed (for the Claire image sequence this is about 3 pixels). Further, to remove regions which are falsely generated at the contour of the facial region (because it is dark at one side and bright at the other), regions which are closer than 2.5% of the face width to this contour are also removed.

*Feature selection*

In Fig. 7(b) an example of the resulting candidate regions is shown. Again many regions other than the eyes and mouth are present (e.g. the nostril regions). From these regions, the eyes and mouth have to be selected on the basis of their properties. The set of properties which are used to distinguish the eye regions from the other candidate regions are the *eye-to-eye distance* and the *symmetry*. The eye-to-eye distance is expected to be equal to 45% of the width of the localized face, and the allowed standard deviation from this expectation is 7.1% of the face width. The symmetry property exploits the assumption that human faces are vertically symmetric. The vertical symmetry axis of concern on the human face, denoted by the *eye vertical*, is the line passing through the center point *between* the eyes and is perpendicular to the line passing through both eye-centers, denoted as the *eye horizontal* (see also Fig. 7(b)).

Let us denote the binary image, as in Fig. 7(b), containing the eye candidate regions as $e$, with the extracted regions labeled as 1 and the background labeled as $-1$. For each pair of candidate regions $(r_p, r_q)$, hypothesized to correspond to a left and right eye, a symmetry axis $l_{p,q}$ can be constructed. The symmetry score for the above pair can then be calculated easily as an inner product between the image $e$, and its mirror reflection with respect to the $l_{p,q}$ axis, as follows:

$$\text{symmetry} = \frac{1}{|D|} \sum_{x, R_{p,q}x \in D} e(x)e(R_{p,q}x), \tag{4}$$

where $D = \{x = (i\ j\ 1)^T \mid 0 \leqslant i \leqslant W, 0 \leqslant j \leqslant H\}$ in which $W$ and $H$ are the width and height of the image $e$. $R_{p,q}$ denotes the reflection operator in the line $l_{p,q}$. For the true eye pairs, this symmetry score is expected to be high, ideally 1. In the selection mechanism the standard deviation is chosen as 0.42.

To find the true eye pair, each candidate pair of regions should be tested on the eye-to-eye distance and derived symmetry value. Especially the last property is computationally expensive. Therefore, before these actual measurements are made, the following tests (illustrated in Fig. 9) are applied on

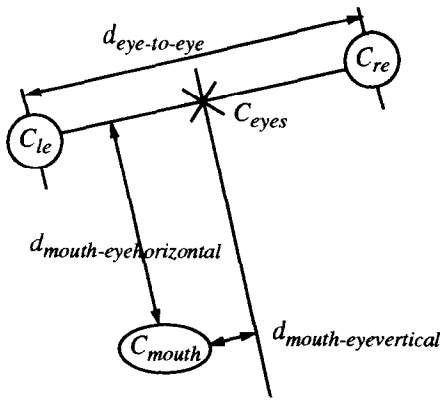



Fig. 9. The geometry of the eye pair regions and the mouth region, which is used in the reduction of the number of candidate eye pair regions. $C_{le}$, $C_{re}$ and $C_{mouth}$ are respectively the center point of the left eye, right eye and the mouth region. $C_{eyes}$ is the center point of the left and right eye regions. The line connecting the eyes is called the eye horizontal and the vertical symmetry axis is called the eye vertical.

each candidate pair of regions to constrain the search space:

(i) The head inclination, measured as the angle of the eye horizontal, is constrained to be within $\pm 45°$, so that region pairs with higher inclinations are disregarded.

(ii) The eye-to-eye distance, $d_{eye\text{-}to\text{-}eye}$, (Fig. 9), is constrained to be between 25% and 70% of the width of the localized face.

(iii) There should be at least one region from the map of candidate *mouth* regions which fulfills the following conditions:

- The center ($C_{mouth}$) is below the center of the eyes ($C_{eyes}$).
- The distance of $C_{mouth}$ (Fig. 9) to the eye horizontal is restricted to be within 100% and 175% of the eye-to-eye distance.
- The distance of $C_{mouth}$ (Fig. 9) to the eye vertical should be less than 25% of the eye-to-eye distance.

By applying these tests the number of region pairs for which the symmetry value should be determined was reduced considerably. For frame 30 of the Talking sequence this number was reduced to 16 as compared to 435 measurements originally. The actual number of symmetry measurements is significantly less as compared to symmetry measurements one has to do with a relatively unconstrained method as in [12].

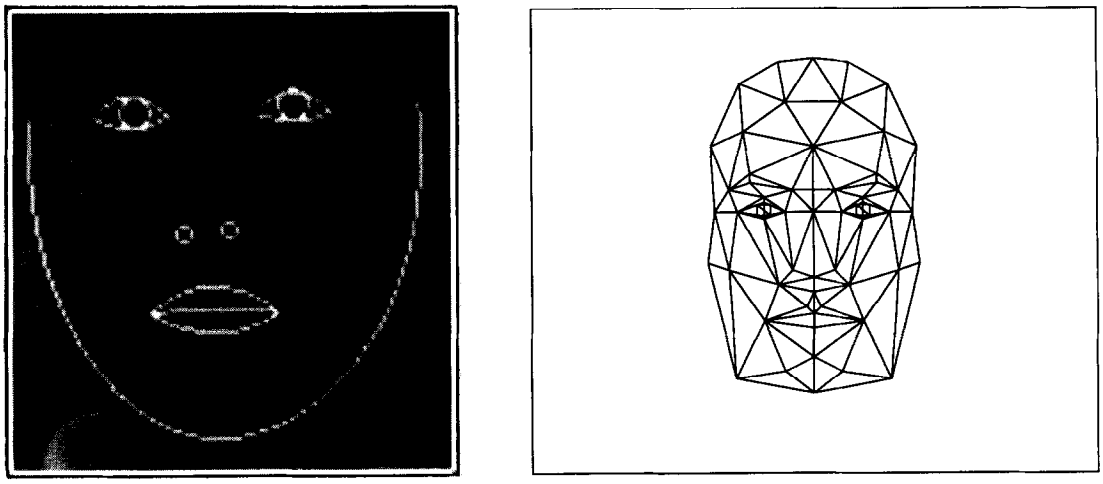
In conclusion, the eyes and mouth regions are detected as a threesome ensemble. First, an eye pair is selected. A pair of regions is only considered as a true eye pair candidate if it passes the above-mentioned tests. Then, the score for the candidate eye pair is calculated on the basis of the *eye distance* and the *symmetry* value. The candidate eye pair with the best score is selected to represent the eye pair. However, if there is a region pair (besides the best scoring pair) which has a competing score (i.e. the score of the competing pair is not less than 65% of the best score), then one always selects the lower pair. At the same time, the upper pair of regions is hypothesized to belong to the eyebrows. This heuristic rule is applied because the eyebrows can have property values closely matching the ones of the eyes.

In a second step the mouth region is selected among the candidates satisfying the above constraints (iii). The *distance of the center of the mouth to resp. the eye vertical and the eye horizontal* are used as properties in the selection mechanism. The expected values and their allowed deviations for the distance to the eye vertical are resp. 0% and 18% of the eye-to-eye distance, and for the distance to the eye horizontal resp. 125% and 18% of the eye-to-eye distance.

The feature localization technique described above has worked very satisfactory in typical videophone sequences, as will be shown in Section 4. The information thus obtained is used to adapt and guide the wire-frame model so that individual somatic traits, motion, and expressions can be reproduced at the receiving site.

## 3. Adaptation of the generic face model

In the foregoing, a method has been described to locate the semantic features in a head-and-shoulder scene. These features can be used for the adaptation of a generic face model to the face in the scene in order to acquire an accurate description. However, to acquire an accurate adaptation, not only the locations of these features need to be known but also accurate contour information of the facial



(a) (b)

Fig. 10. (a) Automatically extracted contours of the chin outline, eyes, nostrils and mouth for frame 10 of the Talking sequence. (b) Front view of the CANDIDE 3-D wire-frame model description of a general face.

features, such as eyes, mouth and the face itself, is required. In [18] a method has been described for contour extraction of facial features using deformable templates, and in [11] this approach has been integrated with the feature localization described in the previous section. We refer to [11] for further details of the contour extraction algorithm, which will not be discussed here. Fig. 10(a) illustrates a set of contours obtained by this algorithm as applied to a frame of the Talking sequence.

The adaptation of the generic model on the basis of these shapes takes place in two successive steps: (i) the global transformation, and (ii) local transformations. The global transformation accounts for the resizing of the wire frame as well as repositioning to give it the initial pose of the speaker. However, after the global transformation has been applied, there remain residual differences between the model and the scene facial outlines as well as mouth and eye contours. The local transformations deal with differences in such facial geometries, e.g. correcting for slight asymmetries, and repositioning of the model eyes and mouth. The generic wire-frame model of the face used in our experiments is shown in Fig. 10(b).

### 3.1. Global adaptation

The global adaptation of the generic face model consists of 3-D rotation, translation and scaling operations of the face. Thus, in principle, nine parameters must be estimated, three for scaling $(s_x, s_y, s_z)$, three for translation, and three rotation angles $(r_x, r_y, r_z)$. As illustrated in Fig. 11(a), in our notation, these three angles correspond to the head tilt $r_x$ (around the $x$-axis), the head rotation $r_y$ (around the $y$-axis), and head inclination $r_z$ (around the $z$-axis). In reality it proves difficult to extract depth information at this stage from planar contours, hence they are derived indirectly from other parameters which are readily estimated.

The parameters are estimated using six points $(\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4, \boldsymbol{p}_5, \boldsymbol{p}_6)$, obtained from measurements on the extracted facial contours, which are defined as follows: $\boldsymbol{p}_1$ is the left corner point of the left eye contour, $\boldsymbol{p}_2$ is the right corner point of the right eye contour, $\boldsymbol{p}_3$ is the midpoint of $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, $\boldsymbol{p}_4$ is the topmost point of the mouth contour, $\boldsymbol{p}_5$ and $\boldsymbol{p}_6$ are projections of the face contour on the eye horizontal (the line through $\boldsymbol{p}_1$–$\boldsymbol{p}_2$) that are maximally distant from the point $\boldsymbol{p}_3$. Using these references,

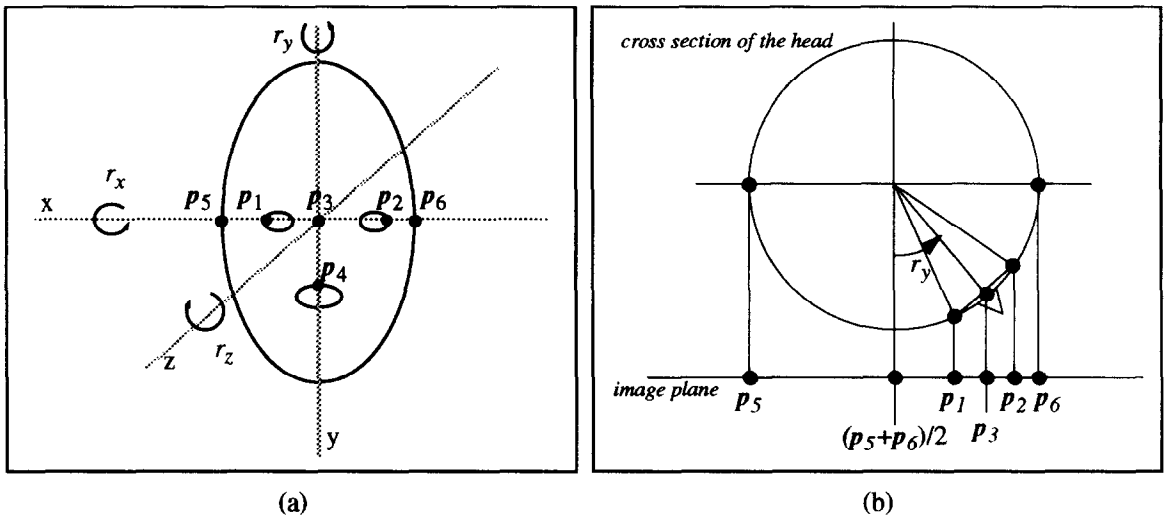



Fig. 11. (a) Global adaptation parameters and control points. (b) Cross-section of the head at the height of the eyes, showing the orthogonal projection of the control points on the image plane and their relation to the head rotation.

the estimation of the rotation and scaling parameters proceed as follows.

*Pose angles*: The inclination of the head ($r_z$) is simply estimated as the angle of the eye horizontal. The head rotation, on the other hand, is found by considering the drift of the eye center from that of the head center on the same latitude. Assuming a circular cross-section of the head at the eye height, one compares the midcenter of the eyes, $\boldsymbol{p}_3$, to that of the line through $\boldsymbol{p}_5$ and $\boldsymbol{p}_6$, in the formula below (Fig. 11(b) illustrates this geometry):

$$\sin(r_y) = 2\frac{\langle(\boldsymbol{p}_3 - (\boldsymbol{p}_5 + \boldsymbol{p}_6)/2), (\boldsymbol{p}_6 - \boldsymbol{p}_5)\rangle}{\|(\boldsymbol{p}_6 - \boldsymbol{p}_5)\|^2}, \qquad (5)$$

where $< \cdot, \cdot >$ is the inner product. The head tilt, $r_x$, is assumed to be negligible or known at this stage. This assumption had to be made due to such difficulties as the vertical length of the head not being constant due to jaw movement, and the top of the head often being ambiguously delineated due to hair cover. Furthermore the distance between $\boldsymbol{p}_3$ and $\boldsymbol{p}_4$ is affected in a coupled way both by the actual scale of the subject and the head tilt.

*Scales*: The scaling factor $s_x$ can be estimated from the ratio of lengths $\|(\boldsymbol{p}_6 - \boldsymbol{p}_5)\|$ as measured both in the model and in the actual image. For a known head tilt, the factor $s_y$ can be similarly estimated from the ratio of the $\|(\boldsymbol{p}_4 - \boldsymbol{p}_3)\|$ lengths. Since range data are not readily available, the depth scaling is taken at this stage as the average of the $s_x$ and $s_y$ factors.

*Position*: The position of the face in the projection plane can be simply estimated by matching the coordinates of the control points with the corresponding model coordinates in the $x$–$y$ plane (assuming an orthographic projection).

### 3.2. Local adaptation

The goal of the local adaptation, which takes place after the affine transformation has been executed in the global adaptation step, is to provide a more refined fit to the somatic contours of the actual face. The model wire frame is thus subjected to various local deformations. These local adaptation steps are significant (to differentiate between different people), although usually small in magnitude as compared to the initial global adaptation.

The implementation of the local adaptation is based on a method first described by Burr [4]. In this technique both the wire frame and the goal contours are treated as planar graphs; therefore the curvilinear contours must first be polygonized, e.g.

using the Wall–Danielson technique [16]. As the matching of the two graphs proceeds, the vertices of the model are driven towards those of the scene contours to coincide. However, to maintain the naturalness of the face and to avoid disproportionate triangulation, each contour translation is propagated smoothly to the remaining vertices of the face model. The displacement vectors for each vertex are derived from local mismatches between the positions of the vertices of the start contour (certain parts of the wire frame contour corresponding to the contours of the facial features) and the goal contour (actual contours of the facial features). Note also that the upper contour of the actual face will be almost always missing due to hair cover. To recuperate for the missing portions of the facial contour, we make use of the knowledge that the shape of the head is symmetric around the center point of the eye horizontal axis (see Fig. 11(a)), whereby the missing contour portions are found by reflecting the visible symmetric counterparts across the center point of $\boldsymbol{p}_5$ and $\boldsymbol{p}_6$.

The contour mismatch (between model and scene), as shown in Fig. 12(a), creates a displacement field which can be interpreted as a force field acting on the model, which will reshape it towards the face contour. For each vertex $\boldsymbol{s}_i$ of the start contour (wire-frame contour) a displacement vector $\boldsymbol{d}_{s,i}$ to the goal contour (actual facial contour) is calculated, as illustrated in Fig. 12(a). Similarly, for each vertex $\boldsymbol{g}_j$ of the goal contour, a displacement vector $\boldsymbol{d}_{g,j}$ towards the start contour is calculated.

The graph (wire-frame model) can now be deformed in a controllable manner by defining a smoothed displacement vector, $\boldsymbol{d}_x$, for any grid position, $\boldsymbol{x}$, illustrated in Fig. 12(b), as a weighted average of the calculated displacement vectors, using Eq. (6). Now, every vertex can be visited in the wireframe and the displacement $\boldsymbol{d}_x$ at that vertex is calculated as in Eq. (6), which computes the accumulated effect of contour displacements of the $N_s + N_g$ contour vertices. Fig. 12(b) illustrates the computed weighted displacement vectors, $\boldsymbol{d}_x$, for a number of grid points $\boldsymbol{x}$. As is apparent from the formula, $\boldsymbol{d}_x$ is a weighted mean of all $\boldsymbol{d}_{s,i}$ (displacements of $N_s$ vertices of the start contour) and all $\boldsymbol{d}_{g,j}$ (displacements of $N_g$ vertices of the goal contour), where the exponential terms are the weights. The weights with which the displacement vectors $\boldsymbol{d}_{s,i}$ and $\boldsymbol{d}_{g,j}$ influence any vertex of the graph depend upon the Euclidean distance of this vertex (at position $\boldsymbol{x}$) to all $\boldsymbol{s}_i$ and $\boldsymbol{g}_j$ vertices. The smoothed displacement vector then becomes

$$\boldsymbol{d}_x = \frac{1}{\gamma}\left\{\frac{\sum_{i=1}^{N_s} \boldsymbol{d}_{s,i}\exp(-\|\boldsymbol{x}-\boldsymbol{s}_i\|^2/\sigma^2)}{\sum_{i=1}^{N_s}\exp(-\|\boldsymbol{x}-\boldsymbol{s}_i\|^2/\sigma^2)} - \frac{\sum_{j=1}^{N_g}\boldsymbol{d}_{g,j}\exp(-\|\boldsymbol{x}-(\boldsymbol{g}_j+\boldsymbol{d}_{g,j})\|^2/\sigma^2)}{\sum_{j=1}^{N_g}\exp(-\|\boldsymbol{x}-(\boldsymbol{g}_j+\boldsymbol{d}_{g,j})\|^2/\sigma^2)}\right\}. \tag{6}$$

In this formula a damping factor, $\gamma$, controls the overshoots/undershoots of the iterations within each frame, whereas $\sigma$ plays the role of a stiffness parameter. In fact, large values of $\sigma$ correspond to a rigid graph, not allowing much local deformation. On the other hand, small values of $\sigma$ signify that displacement effects remain very local, such that every displacement is limited to a few vertices nearby, which effectively corresponds to a very elastic graph. Initially it is assumed that the discrepancy between the two contour graphs is large, hence iterations start with a high stiffness parameter. This stiffness parameter is gradually decreased so that neighborhoods of interacting vectors become smaller, and hence matching in finer detail can be realized. The iterations terminate when the mean value of the displacement vectors $\boldsymbol{d}_{s,i}$ falls below a threshold. This mean value can be interpreted as a measure of residual mismatch between the two contours. We have used the setting suggested by Burr for the stiffness parameter:

$$\sigma_k = \sigma_0/f^k, \tag{7}$$

where $k$ is the iteration number, and $f$ is some constant between 1 and 2. The results shown in Fig. 12(c) were produced in ten iterations with $\gamma = 1.3$, $\sigma_0 = 200$ and $f = 1.2$. Fig. 12(d) shows the adapted wire-frame overlaid on the actual Miss America image it was adapted to.

We have implemented the local adaptation based on Eq. (6) only in the initial frames. But in fact, this adaptation scheme can be invoked at any instant to

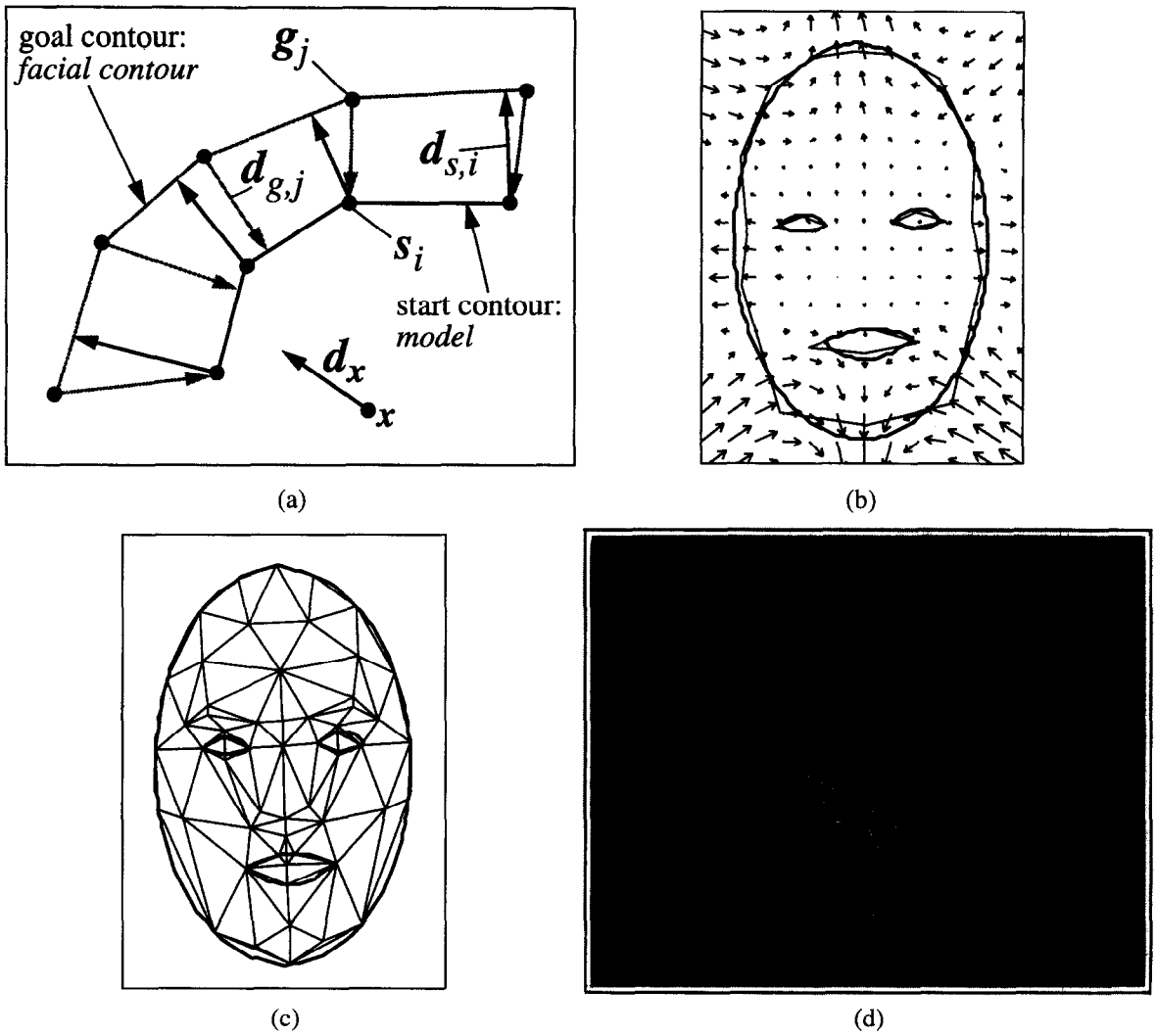



Fig. 12. (a) Illustration of feature displacement vectors: $s_i$ the start contour; $g_j$ the goal contour; $d_{s,i}$ the pushing feature displacement vectors, $d_{g,i}$ the pulling feature displacement vectors; $d_x$ the smoothed displacement vector at position $x$ in the image. (b) The start contour (thin lines), the goal contour (thick lines), and the smoothed displacement vectors for a set of grid points. (c) The adapted model (thin lines) and the extracted feature contours (thick lines). (d) Original gray value image with the adapted model overlaid.

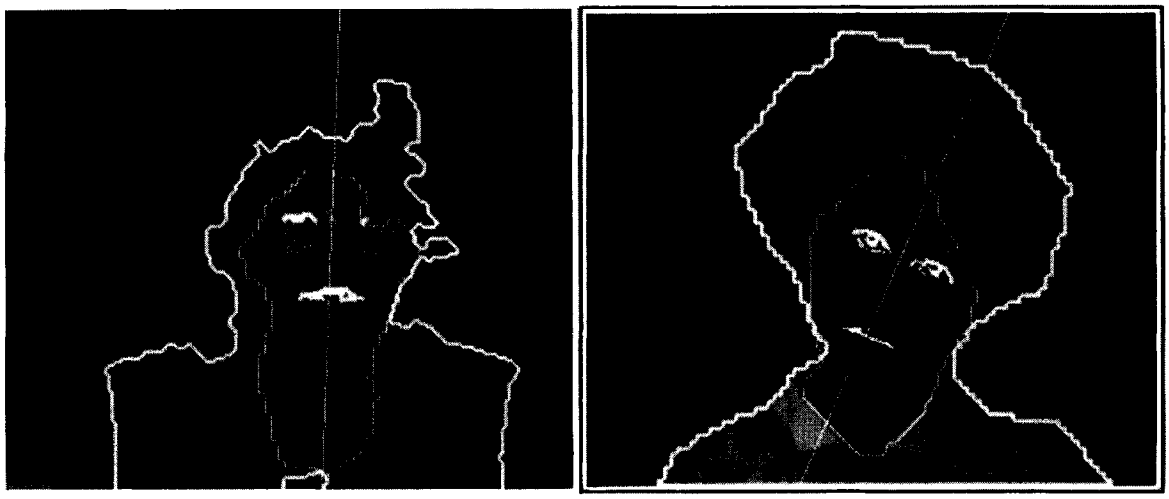
compensate for facial deformations arising from expressions.

## 4. Experimental results

Experimental results and performance figures obtained from typical head-and-shoulder test images are discussed in the sequel.

### 4.1. Evaluation of the localization of facial features

In this section we report the performance of the above algorithms in locating each feature in the head-and-shoulder scene. The reported performances are the worst-case results, in that the algorithms are tested on individual frames, that is without making use of their sequence properties. In practice the localization algorithms would be



(a) (b)

Fig. 13. Selected regions representing the features of interest (silhouette, face, eyes and mouth) and the corresponding eye vertical for (a) frame 79 of the Miss America sequence, and (b) frame 205 of the Talking sequence.

tracking the regions of interest making explicit use of information accumulated in previous frames. We have used the following sequences in our experiments: Miss America, Claire and Talking. The statistics of property values (as explained in Section 2 and Appendix A) are collected from measurements on 25 frames in the Claire and 25 frames in the Miss America sequences. Hence, these frames should be considered as our training set, while the remaining frames of these sequences and the frames in the Talking sequence are the test set. The head in the Talking sequence shows large rotations and thus can be a proof of the proposed scheme. In Fig. 13 typical localization results are shown for frame 79 of the Miss America sequence and for frame 205 of the Talking sequence.

*Silhouette region*: The silhouette detection algorithm is tested on 271 different frames, selected from the three different sequences. A silhouette region was detected in all cases. However in some frames the silhouette was somewhat fragmented, especially on the torso region, which either lacked sufficient motion or gray level details for a good segmentation.

*Head region*: In all cases the silhouette encompassed a connected head region, which was extracted successfully.

*Facial region*: Since the facial region was segmented on the basis of color attributes of the skin, the neck was also occasionally included. In all tested cases the correct facial region was selected. The heuristic rule to separate the hair blob from the facial blob proved indeed necessary, especially in scenes where the speaker was bending forward.

*Eyes and mouth*: The localization performance for the eye- and mouth-features are shown in Table 1. An eye or mouth is said to be identified correctly if the center of gravity of the corresponding region falls close enough to the correct points manually determined in each frame. In Table 1 we indicate for each sequence: the number of frames tested, the number of frames in which both the eyes and the mouth are located correctly, the mean and variance of the distance of these features to their actual positions (as determined by a human operator), as well as the mean eye-to-eye distance to give an impression of the scale of the face. The percentage of correct frames were 100%, 100% and 98.4%, respectively, for Miss America, Claire and Talking, or, in other words, an overall performance of 99.6% across all sequences. Note that for the eyes and mouth to be located correctly all other features should have been located correctly. Hence, these performance scores give an

Table 1
Percentage of correctly processed frames (column 3). The mean and standard deviation of the distances in pixels of the various facial features from their correct position are also shown in relation to the actual eye-to-eye distance

| Sequence title | Number of tested frames | Number of correct frames | Facial feature | Absolute mean deviation | Standard deviation | $d_{\text{eye-to-eye}}$ |
|---|---|---|---|---|---|---|
| Miss America | 140 | 140 (100%) | $C_{\text{le}}$ | 2.1 | 2.2 | 41.0 |
| | | | $C_{\text{re}}$ | 2.6 | 1.9 | |
| | | | $C_{\text{mouth}}$ | 8.0 | 2.7 | |
| Claire | 69 | 69 (100%) | $C_{\text{le}}$ | 1.6 | 1.8 | 28.5 |
| | | | $C_{\text{re}}$ | 1.9 | 0.7 | |
| | | | $C_{\text{mouth}}$ | 2.2 | 1.4 | |
| Talking | 62 | 61 (98.4%) | $C_{\text{le}}$ | 3.7 | 2.6 | 43.7 |
| | | | $C_{\text{re}}$ | 4.2 | 1.4 | |
| | | | $C_{\text{mouth}}$ | 2.5 | 4.9 | |

indication on the capability of the system to localize *all* features.

Only in one case, frame 21 of the Talking sequence, the right eye was not found (although the other features were found), because no candidate right eye region was generated due to severe motion blur in the region of the right eye (combination of head movement and blinking). In this case the eye was no longer a clearly distinctive dark blob in a bright surrounding, thus violating the underlying assumption of the candidate eye region generation.

### 4.2. Extension of the localization to tracking

A preliminary experiment has been performed to study the ability to detect the eyes and mouth consecutively in a sequence of frames. In these experiments, the features are localized for the first frame in initialization mode, i.e. according to the description in Section 2; for all following frames they are localized in tracking mode. In the tracking mode, the search space of the localization scheme can be constrained significantly since the information about the features from previous frames can be used. According to this information most candidate regions in the segmentation map can be ruled out using regions of interest based on the positions of the feature in the previous frames. Of course this assumption holds only if the speed of motion of these features with respect to the frame rate is moderate. The search space can be restricted further by also basing the prediction of the property values on knowledge about features found in the previous frames. Thus, the knowledge-based prediction is no longer based on static (intraframe) information, but now can also use dynamic (interframe) information. Hence, to localize the eyes it is no longer necessary to localize the facial region first. Among the set of properties used for the eyes and mouth regions in the tracking mode were two intra-properties: (i) the *symmetry* score, (ii) the *distances between the eyes*; and two were new inter-properties: (i) the *distance between the positions of the left eye in two consecutive frames*, (ii) the *distance between the positions of the right eye in two consecutive frames*. Besides these intra- and inter-properties, also the tests as discussed in Section 2.4 are applied on each candidate pair of regions.

The performance of the region localization algorithms in the tracking mode remains the same as compared to the performance in the initialization mode. The computational load of the search, however, was much reduced due to the restricted number of candidate regions.

### 4.3. Evaluation of the model adaptation

In general, it is very difficult to find objective measures to assess the performance and quality of

the model adaptation. Part of the difficulty resides in the fact that the wire-frame model itself is rather coarse. One measure of goodness of fit could have been a norm of the residual displacement vectors. However, this measure is already incorporated in the stopping criterion of the iterations. Hence at this stage the best judgment of the model adaptation algorithms is subjective assessment.

The subjective assessment results obtained from the three test sequences were overall satisfactory. In all cases the model and facial contours are well aligned during the global adaptation, so that during the local adaptation the contours of the model could be transformed correctly to the facial contours in the scene. Also, the propagation of the contour displacement vectors did not yield any visible problems.

## 5. Conclusion

A method for the adaptation of a generic 3-D face model using 2-D projections data of a head-and-shoulders scene has been presented. More specifically the contributions of this investigation as presented are: (i) a robust method for the automatic localization of semantic facial features, using a knowledge-based selection mechanism, and (ii) a method to adapt a generic face model to facial contours. Both methods, as tested on typical videophone sequences, perform satisfactory. The localization algorithm is also shown to work well in the tracking mode. The limitations of the algorithms are presently that excessive head rotations are to be avoided. Other occlusions, for instance, due to hands or beards are also precluded.

Localization of the facial features in situations which violate the assumptions, such as occlusion, as well as the local adaptation in the depth direction are being investigated.

## Appendix A. Knowledge-based selection mechanism

The knowledge-based selection mechanism [13] purports to selecting the most probable region from a set of candidate regions for a particular object class $C$. We are interested in the probability that region $r_i$ represents this class given the set of measurements ($\boldsymbol{m}_j$) made on all of the regions:

$$P\left(\text{region } r_i \text{ represents } C \,\middle|\, \bigcup_{j=1}^{N_r} \boldsymbol{m}_j\right), \tag{A.1}$$

where $N_r$ is the number of candidate regions and $\boldsymbol{m}_j$ is the set of measurements made on region $r_j$.

The region which is supported most by the evidence, given by all measurements $\boldsymbol{m}_j$, resembles the class $C$ most. Thus, the region which maximizes Eq. (A.1) is selected to represent class $C$. The next step would be to break this expression up into simpler terms, involving only probabilities conditioned on an individual set of measurements $\boldsymbol{m}_i$, $P(C|\boldsymbol{m}_i)$, because these are the only distributions which can be determined a priori. However, factorizing Eq. (A.1) into such terms requires independence of the measurements $\boldsymbol{m}_i$. This cannot be guaranteed because certain regions may belong to the same object or to objects which have properties in common, e.g. the color of the face and hands is very likely to be the same. Therefore, we are forced to adopt a suboptimal approach: we proceed by selecting the region having a maximum probability that it belongs to the object class $C$, *based on its individual measurements only*. As the performance of our system is still very good, this simplification does not seem to have any noticeable effects.

The selection process then reduces to the maximization of $P(C|\boldsymbol{m}_i)$ over all candidate regions $r_i$, or

$$\max_{i=1,\ldots,N_r} (P(C|\boldsymbol{m}_i)). \tag{A.2}$$

Applying Bayes rule to Eq. (A.2) gives

$$\max_{i=1,\ldots,N_r} \left(\frac{P(\boldsymbol{m}_i|C)P(C)}{P(\boldsymbol{m}_i)}\right). \tag{A.3}$$

We would like to point out that the maximization takes place over all measurement vectors, rather than over the different object classes as is the case in classical pattern recognition, where one tries to classify a region into one of the possible classes, given its measured property vector. Consequently, $P(C)$ is now constant over the maximization while

$P(\boldsymbol{m}_i)$ would have been held constant in a pattern recognition problem. The optimal decision rule, under the assumption that different measurements *within each measurement set* are independent, becomes

$$\max_{i=1,\ldots,N_r}\left(\frac{P(\boldsymbol{m}_i|C)}{P(\boldsymbol{m}_i)}\right)=\max_{i=1,\ldots,N_r}\left(\frac{P(m_{i,1}|C)}{P(m_{i,1})}\times\cdots\frac{P(m_{i,l}|C)}{P(m_{i,l})}\cdots\frac{P(m_{i,N_m}|C)}{P(m_{i,N_m})}\right), \tag{A.4}$$

where the set of measurements $\boldsymbol{m}_i$ is split into $N_m$ different measurements, which are elements of the set.

Each ratio $P(m_{i,l}|C)/P(m_{i,l})$ in Eq. (A.4) can be considered as a score representing the goodness of fit to the object class $C$. In other words, the selection process pinpoints that region whose set of property values $\boldsymbol{m}_i$ are closest to the expected values based on the *prior* knowledge about the object class, i.e. regions which have high scores.

Although there is no fixed rule for property selection it is desirable that they have high discriminative power, whereby the ratio $P(m_{i,l}|C)/P(m_{i,l})$ peaks for the sought class region. Furthermore, to apply Eq. (A.4), these properties should be independent. In our experiments the set of properties are chosen heuristically, and, in the absence of further knowledge, we assumed that the measured property values have uniform prior distributions. Further, their posterior distributions were assumed to be independent and Gaussian with means and variances estimated from measurements on training images. More specifically, the mean and variance of the normal distribution for each property are derived from a set of training images (25 frames of the Miss America sequence and 25 frames of the Claire sequence). Although the selection mechanism worked very satisfactorily under these assumptions, further work remains to be done for their justification.

To give a concrete example, consider an object class 'eye pair'. The set of property values which are used to select the pair of regions corresponding to the eye pair from the other region pairs are the eye-to-eye distance ($m_{i,\mathrm{dist}}$) and the symmetry score ($m_{i,\mathrm{sym}}$), see also Section 2.4. Then the conditional probability becomes

$$P(\mathrm{Eyepair}|\{m_{i,\mathrm{dist}},m_{i,\mathrm{sym}}\})\sim\exp\left[-\frac{\|m_{i,\mathrm{dist}}-0.45\,\mathrm{Width}_{\mathrm{face}}\|^2}{2(0.071\,\mathrm{Width}_{\mathrm{face}})^2}\right]\times\exp\left[-\frac{(m_{i,\mathrm{sym}}-1)^2}{2(0.42)^2}\right]. \tag{A.5}$$

This score (as in Eq. (A.5)) is calculated for each pair of candidate regions, and the region which has the highest score is then selected to represent the object class 'eyepair'.

## References


[1] K. Aizawa, H. Harashima and T. Saito, "Model-based analysis synthesis image coding (MBASIC) system for a person's face", *Signal Processing: Image Communication*, Vol. 1, No. 2, October 1989, pp. 139–152.

[2] P.J.L. van Beek, M.J.T. Reinders, B. Sankur and J.C.A. van der Lubbe, "Semantic segmentation of videophone image sequences", *Proc. SPIE Visual Communications and Image Processing '92*, Vol. 1818, Boston, MA, 15–20 November 1992, pp. 1182–1193.

[3] G. Bozdağı, A.M. Tekalp and L. Onural, "3-D motion and structure estimation including photometric effects with application to model-based coding of facial image sequences", *Proc. 8th Workshop on Image and Multidimensional Signal Processing*, Cannes, France, 8–10 September 1993, pp. 114–115.

[4] D.J. Burr, "Elastic matching of line drawings", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 3, No. 6, November 1981, pp. 708–713.

[5] CCITT SGXV WP XV/1 – Specialist Group on Coding for Visual Telephony, Draft revision of recommendation H.261, Doc. no. 584, Tokyo, November 1989.

[6] A.F. Clark and M. Köküer, *Proc. 11th IAPR Internat. Conf. on Pattern Recognition*, Vol. 3, The Hague, The Netherlands, 30 August–3 September 1992, pp. 79–82.

[7] R. Forchheimer and T. Kronander, "Image coding – From waveforms to animation", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 37, No. 12, 1989, pp. 2008–2023.

[8] T.S. Huang, S.C. Reddy and K. Aizawa, "Human facial motion modeling, analysis and synthesis for video compression", *Proc. SPIE Visual Communications and Image Processing: Visual Communication*, Vol. 1605, Boston, MA, November 1991, pp. 234–241.

[9] T. Kanade, *Computer Recognition of Human Faces*, Birkhäuser, Basel, 1977.

[10] H. Li, P. Roivainen and R. Forchheimer, "Recursive estimation of facial expression and movement", *Proc. ICASSP'92 Internat. Conf. Acoust. Speech Signal Process.*, Vol. 3, San Francisco, CA, 23–26 March 1992, pp. 593–596.

[11] M.J.T. Reinders, F.A. Odijk, J.C.A. van der Lubbe and J.J. Gerbrands, "Tracking of global motion and facial expressions of a human face in image sequences", *Proc. SPIE Visual Communications and Image Processing '93*, Vol. 2094, Cambridge, MA, 8–11 November 1993, pp. 1516–1527.

[12] D. Reisfeld and Y. Yeshurun, "Robust detection of facial features by generalized symmetry", *Proc. 11th IAPR Internat. Conf. on Pattern Recognition*, Vol. 1, The Hague, The Netherlands, 30 August–3 September 1992, pp. 117–120.

[13] E.M. Riseman and A.R. Hanson, "A methodology for the development of general knowledge-based vision systems", in: M.A. Arbib and A.R. Hanson, eds., *Vision, Brain and Cooperative Computation*, MIT Press, Cambridge, MA, 1987, pp. 285–328.

[14] A. Samal and P.A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey", *Pattern Recognition*, Vol. 25, No. 1, 1992, pp. 65–77.

[15] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 15, No. 6, June 1993, pp. 569–579.

[16] K. Wall and P.E. Danielsson, "A fast sequential method for polygonal approximation of digitized curves", *Comput. Vision Graph. Image Process.*, Vol. 28, 1984, pp. 220–227.

[17] W.J. Welsh, S. Searby and J.B. Waite, "Model-based image coding", *British Telecom Technol. J.*, Vol. 8, No. 3, July 1990, pp. 94–106.

[18] A.L. Yuille, P.W. Hallinan and D.S. Cohen, "Feature extraction from faces using deformable templates", *Internat. J. Comput. Vision*, Vol. 8, No. 2, 1992, pp. 99–111.

[19] G.W. Zack, W.E. Rogers and S.A. Latt, "Automatic measurement of sister chromatid exchange frequency", *J. Histochem. Cytochem.*, Vol. 25, No. 7, 1977, pp. 741–753.