Synthesis of Facial Expressions for Semantic Coding of Videophone Sequences

Markus Kampmann, Bruno Nagel
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover, Appelstraße 9A, 30167 Hannover, F.R.Germany
Email: {kampmann,nagel}@tnt.uni–hannover.de
WWW: http://www.tnt.uni–hannover.de/{~kampmann,~nagel}

Abstract

In this contribution, a method for the synthesis of facial expressions for semantic coding of videophone sequences is presented. Firstly, a generic 3D face model is automatically adapted to the eyes, mouth, eyebrows, nose and chin and cheek contours of the individual face in the sequence. Secondly, a synthesis of facial expressions of this individual face is carried out using a model of the human facial muscles. Applying the described algorithms to the videophone sequences Akiyo *and* Miss America *(CIF, 10Hz), the face model is successfully adapted at the beginning of the videophone sequence and facial expressions of the individual faces are synthesized.*

1. Introduction

In typical videophone sequences, head and shoulder of human persons appear in the scene. For coding of such kind of sequences at very low bit rates, semantic coding (also called model–based coding) has been proposed [1][2]. In this coding scheme, facial expressions of a person are described by mimic parameters e.g. action units [3] or facial muscle parameters [4]. These parameters are estimated in an image analysis at the sender, then coded and transmitted to the receiver. At the receiver, facial expressions are synthesized using the decoded parameters.

In a semantic coder, the human face is represented by a generic 3D face model consisting of a mesh of triangles. This generic face model has the proportions of an average human face. At the sender, the generic face model is adapted automatically to the individual face of the person in the sequence at the beginning of the videophone sequence. It is adapted to the indivual shape of the person's face, e.g. to the face outline and to the nose. Furthermore, the face model is adapted to the actual mimic of the person at the beginning of the sequence, e.g. to the mimic of eyes, mouth and eyebrows. After adaptation of the face model at the sender, the adapted face model is coded and transmitted to the receiver.

In a semantic coder, adaptation and transmission of the face model is carried out only once at the beginning of the videophone sequence. For the following frames, only mimic parameters are estimated at the sender and then coded and transmitted to the receiver. Using the decoded mimic parameters, a synthesis of facial expressions is carried out at the receiver by a deformation of the mesh of triangles of the adapted face model.

In order to realize a complete semantic coder, algorithms for the analysis and synthesis of facial expressions as well as for the adaptation of the generic face model are necessary. This contribution will deal with the problems of the synthesis of facial expressions and the adaptation of the generic face model. The analysis of facial expressions is disregarded in this contribution.

In the literature, some methods are proposed for synthesis of facial expressions in a semantic or model–based coder. In [5], the face model CANDIDE [6] and action units are used. Since CANDIDE consists only of a small number of triangles, the image quality of synthesized facial expressions is not satisfactory. In [7], a more complex face model combined with action units is used. Instead of using action units, the approach in [4] uses the facial muscle parameters from [8] for the synthesis of facial expressions. In [9], the muscle model from [8] is extended to a physics–based facial tissue model.

Mostly, the adaptation of the generic face model to the individual face in the sequence is done manually [4][5][7]. Recently, an algorithm has been proposed allowing an automatic adaptation of a generic face model to the individual face in the videophone sequence [10].

In this contribution, a method for the synthesis of facial expressions of the individual face in a videophone sequence is proposed. Using the algorithm in [10], a 3D face model (Fig. 1) is adapted automatically to the individual face in the videophone sequence. Afterwards, facial expressions are synthesized applying the muscle model from [8]. This muscle model is chosen because it allows a realistic synthesis of facial expressions with low computational cost compared with the tissue model from [9].

Section 2 describes the automatic adaptation of the face model, Section 3 the synthesis of facial expressions. Experimental results are given in Section 4.

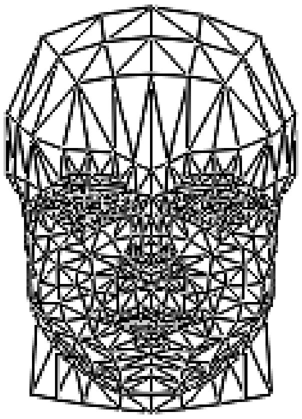


Fig. 1: 3D face model.

2. Automatic face model adaptation

For automatic adaptation of the face model, facial features are estimated first. In this approach, the eyes, the mouth, the chin and cheek contours, the eyebrows and the nose are estimated from an image [11][12][13][14][15]. Using these facial features, the face model is adapted in two steps [10]. In the first step, a global adaptation is carried out. Here, the size of the face model and its orientation in 3D space is adapted. In the second step, a local adaptation adapts the face model to the estimated facial features.

2.1 Global adaptation

By global adaptation, the size of the face model and its orientation in 3D space is adapted. Here, only the eyes and mouth center positions as well as the face contour are evaluated. The global adaptation is carried out in four steps. In the first step, the face model is rotated around the y–head axis (see Fig. 2). By a rotation around the y–head axis, a deviation between $\boldsymbol{p_3}$, the midpoint between the two eyes centers $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$, and $\boldsymbol{p_6}$, the midpoint between the face contour points $\boldsymbol{p_4}$ and $\boldsymbol{p_5}$, occurs (see Fig. 3). If the face is looking straight into the camera, $\boldsymbol{p_3}$ corresponds with $\boldsymbol{p_6}$. Therefore, a measure for the amount of rotation

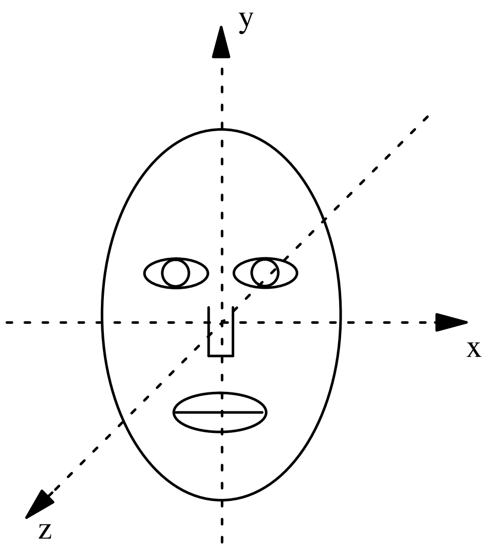


Fig. 2: Head rotation axes.

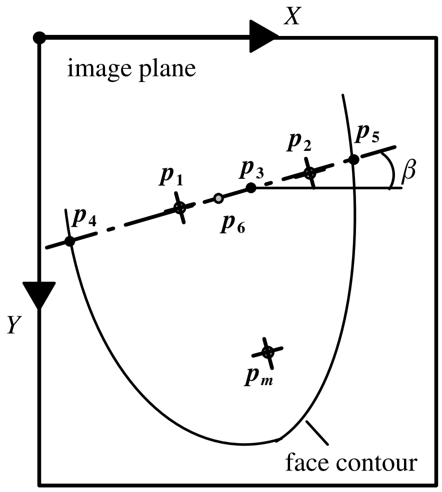


Fig. 3: Facial features in the image plane:
$\boldsymbol{p_1}, \boldsymbol{p_2}$: estimated eyes center positions,
$\boldsymbol{p_3}$: middle between $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$,
$\boldsymbol{p_4}, \boldsymbol{p_5}$: estimated face contour points,
$\boldsymbol{p_6}$: middle between $\boldsymbol{p_4}$ and $\boldsymbol{p_5}$,
$\boldsymbol{p_m}$: estimated mouth center position.

$$\alpha_{real} = \frac{|\boldsymbol{p}_3 - \boldsymbol{p}_6|}{|\boldsymbol{p}_4 - \boldsymbol{p}_5|} \quad (2.1)$$

of the real face is introduced. By projection of the face model onto the image plane, a similar measure for the amount of rotation

$$\alpha_{model} = \frac{|\boldsymbol{p}'_3 - \boldsymbol{p}'_6|}{|\boldsymbol{p}'_4 - \boldsymbol{p}'_5|} \quad (2.2)$$

of the face model is calculated (see Fig. 4). Now, the face model is iteratively rotated around the y–axis until

$$|\alpha_{model} - \alpha_{real}| < th_\alpha \quad (2.3)$$

with a given threshold th_α. In the second step, the face model is rotated around the z–head axis with a rotation angle β such that the lines through the estimated eyes center positions and through the projected positions of the eyes of the face model are parallel (see Fig. 3). In the third step, the face model is scaled by evaluating the distance between the eyes centers and the distance between the eyes and the mouth center. In the fourth step, the face model is shifted in the 3D space in such a way that the projection of the eyes and mouth centers of the face model corresponds with the estimated eyes and mouth center positions in the image plane.

2.2 Local adaptation

After global adaptation, the local adaptation adapts the face model to the estimated facial features. First, according to the estimated mouth's mimic in the image the mouth of the face model is opened by a movement of the face model's jaw. Next, the face model is adapted to the estimated chin and cheek contours. Therefore, the control points of the chin and the cheek of the face model are shifted towards the estimated chin and cheek contours. Neighbouring control points of the face model are also shifted in order to keep the proportions of the human face. The control points of the eyes and mouth centers of the face model are not shifted, because they have been already adapted by global adaptation. In the next step, the eye lids of the face model are moved towards the estimated eye lid contours. Then, the control points of the lips are shifted towards the estimated mouth lip contours. Finally, the face model is adapted to the estimated sides of the nose and to the estimated eyebrows borders by moving the corresponding control points.

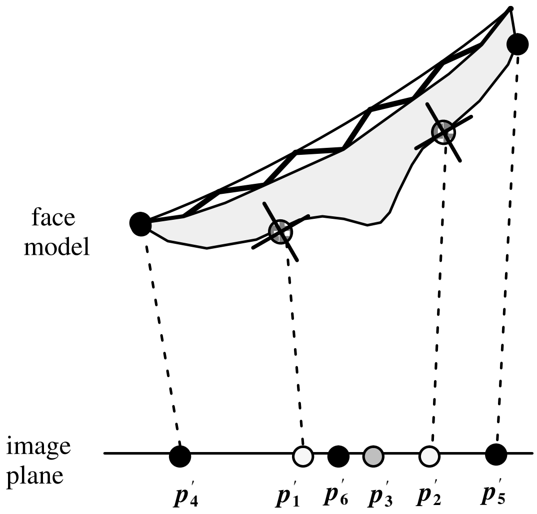


Fig. 4: Projection of the face model onto the image plane:
$\boldsymbol{p}'_1, \boldsymbol{p}'_2$: eyes center positions,
$\boldsymbol{p}'_3$: middle between $\boldsymbol{p}'_1$ and $\boldsymbol{p}'_2$,
$\boldsymbol{p}'_4, \boldsymbol{p}'_5$: face contour points,
$\boldsymbol{p}'_6$: middle between $\boldsymbol{p}'_4$ and $\boldsymbol{p}'_5$.

3. Synthesis of facial expressions

After adaptation, the face model is textured by projecting the corresponding image area of the person's face onto the face model's surface. A synthesis of facial expressions is carried out by a deformation of the mesh of triangles of the face model. For deformation, the vertices of the mesh of triangles are shifted. The shifting is controlled by a model of the human facial muscles. In this approach, the muscle model from [8] is used.

In [8], three different facial muscles types are modelled. The first muscle type, the linear muscle has parallel muscle fibers with one end fixed on the bone and the other end connected to the tissue. An example for this muscle type is the lateral corrugator which pulls the eyebrows down and towards each other (see Fig. 5). The second muscle type is the sphincter muscle with an oval configuration of muscle fibers. An example is the orbicularis oris which purses the mouth (see Fig. 5). The third type, the sheet muscle, represents a large area of parallel muscle fibers. An example of a sheet muscle is the frontalis major muscle of the forehead, which is involved in raising the eyebrows (see Fig. 5). Fig. 5 shows a map of all modelled facial muscles in this contribution.

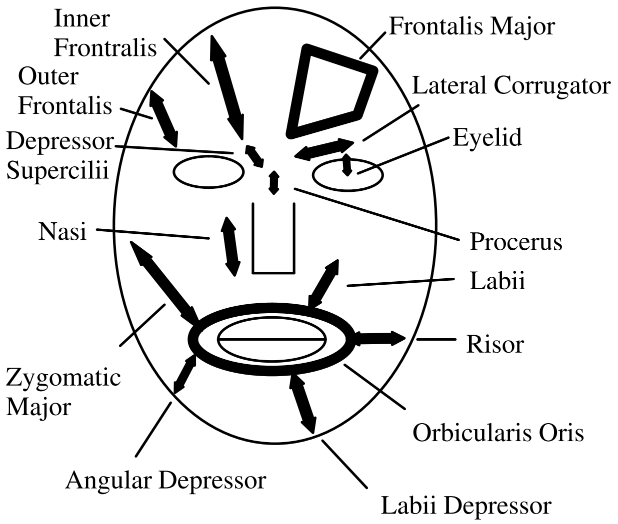


Fig. 5: Map of human facial muscles simulated in the muscle model.

4. Experimental results

The described methods have been applied to the videophone sequences *Akiyo* and *Miss America* with a spatial resolution corresponding to CIF and a frame rate of 10Hz.

Fig. 6 and Fig. 7 show the results for the automatic face model adaptation. For the videophone sequence *Miss America*, Fig. 6 (a)(b) show the estimated facial features at the beginning of the image sequence. Using these facial features, a generic face model (Fig. 1) is adapted (Fig. 6 (c)(d)(e)). Control points of the face model match the corresponding facial features in the image with subjectively high accuracy (see Fig. 6 (d)(e)). Since the borders of the person's neck in the image have not been estimated, the neck of the face model is not adapted (see Fig. 6 (c)). For the videophone sequence *Akiyo*, Fig. 7 shows the estimated facial features and the adapted face model.

After adaptation, texture from the original images (Fig. 8 (a), 9 (a)) is projected onto the face models. For the synthesis of facial expressions, an animation of the human facial muscles is carried out using the muscle model described in Section 3. Fig. 8 (b)(c)(d) and Fig. 9 (b)(c)(d) show examples of synthesized facial expressions. Since an algorithm for the analysis of facial expressions is not implemented yet, the synthesized facial expressions are generated by a manual selection of facial muscle parameters.

5. Conclusions

A method for the synthesis of facial expressions of the individual face of a person in a videophone sequence has been presented. First, a generic 3D face model is automatically adapted to the eyes, mouth, eyebrows, nose and chin and cheek contours of the individual face in the sequence. Afterwards, a synthesis of facial expressions of this individual face is carried out using a model of the human facial muscles. This facial muscle model contains three different muscle types, the linear muscle, the sphincter muscle and the sheet muscle. Applying the described algorithms to the videophone sequences *Akiyo* and *Miss America* (CIF,10Hz), the face model is adapted at the beginning of the videophone sequence and facial expressions of the individual faces are synthesized. In the future, an algorithm for the analysis of facial expressions will be developed in order to complete a semantic coder.

6. References

[1] H. Musmann, "A layered coding system for very low bit rate video coding", *Signal Processing: Image Communication*, Vol. 7, Nos. 4–6, pp. 267–278, November 1995.

[2] D. Pearson, "Developments in Model–Based Video Coding", *Proceedings of the IEEE*, Vol. 83, No. 6, pp. 892–906, June 1995.

[3] P. Ekman, W. Friesen, *Facial Action Coding System*, Consulting Psychologists Press, Palo Alto CA., 1977.

[4] J. Fischl, B. Miller, J. Robinson, "Parameter tracking in a muscle–based analysis/synthesis coding system", *Picture Coding Symposium* (PCS '93), Lausanne, Switzerland, No. 2.3, March 1993.

[5] H. Li, P. Roivainen, R. Forchheimer, "3D Motion Estimation in Model–Based Facial Image Coding", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 6, pp. 545–556, June 1993.

[6] R. Rydfalk, "CANDIDE, A parameterised face", Internal Report Lith–ISY–I–0866, Linköping University, Linköping, Sweden, 1987.

[7] C. S. Choi, K. Aizawa, H. Harashima, T. Takebe, "Analysis and synthesis of facial image sequences in model–based image coding", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 4, pp. 257–275, June 1994.

[8] K. Waters, "A muscle model for animating three–dimensional facial expressions", *Computer Graphics (SIGGRAPH '87)*, Vol. 21, No. 4, pp. 17–24, July 1987.

[9] D. Terzopoulus, K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 6, pp. 569–579, June 1993.

[10] M. Kampmann, R. Farhoud, "Precise face model adaptation for semantic coding of videophone sequences", *Picture Coding Symposium* (PCS '97), Berlin, Germany, pp. 675–680, September 1997.

[11] M. Kampmann, J. Ostermann, "Automatic Adaptation of a Face Model in a Layered Coder with an Object–

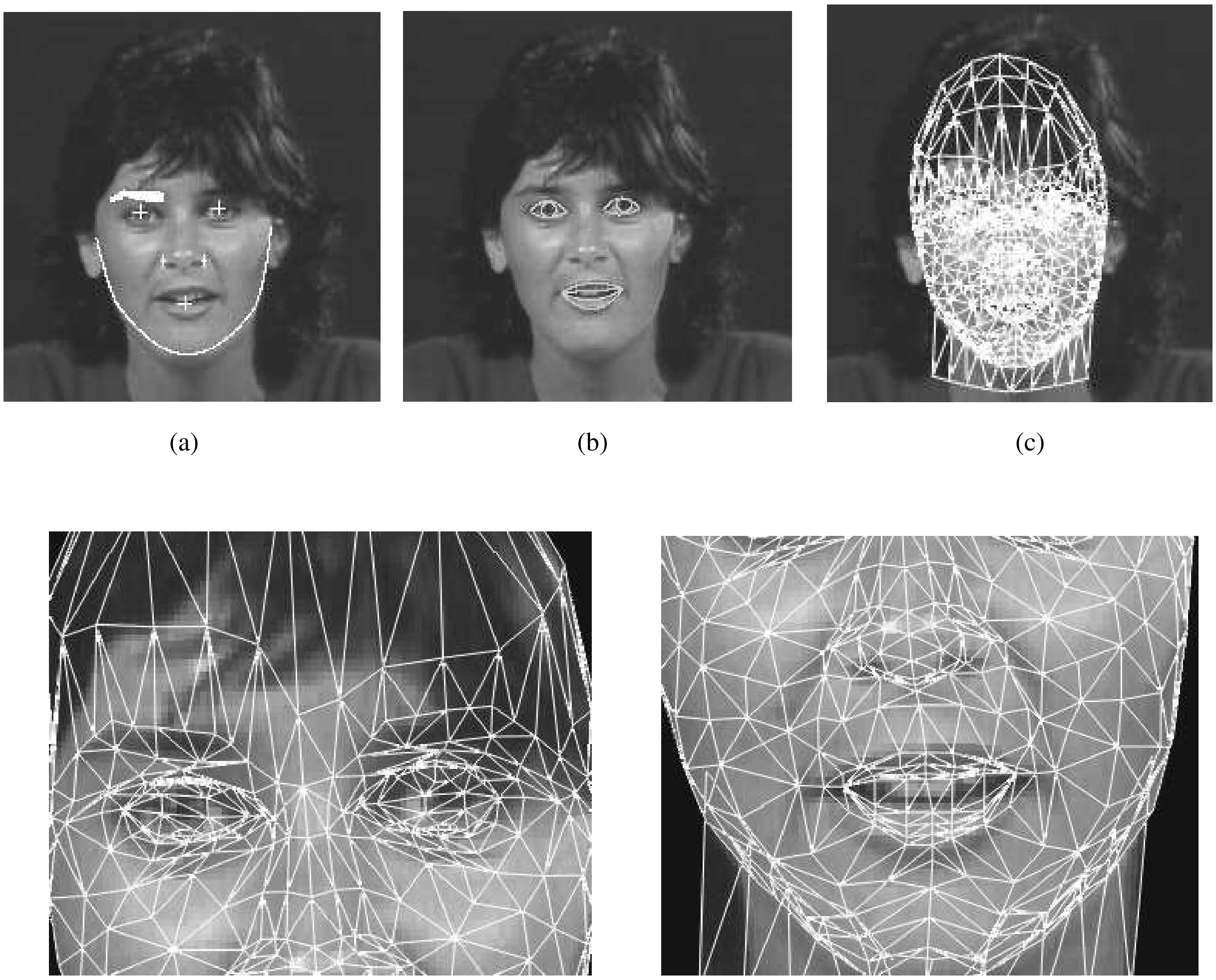
based Analysis–Synthesis Layer and a Knowledge–based Layer", *Signal Processing: Image Communications*, Vol. 9, No. 3, pp. 201–220, March 1997.

[12] L. Zhang, "Estimation of the mouth features using deformable templates", *International Conference on Image Processing (ICIP–97)*, Santa Barbara, CA, USA, pp. 328–331, October 1997.

[13] M. Kampmann, "Estimation of the chin and cheek contours for precise face model adaptation ", *International Conference on Image Processing (ICIP97)*, Santa Barbara, CA, USA, pp. 300–303, Oct. 1997.

[14] M. Kampmann, L. Zhang, "Estimation of eye, eyebrow and nose features in videphone sequences", submitted to VLBV 98, Urbana, IL, USA, October 1998.

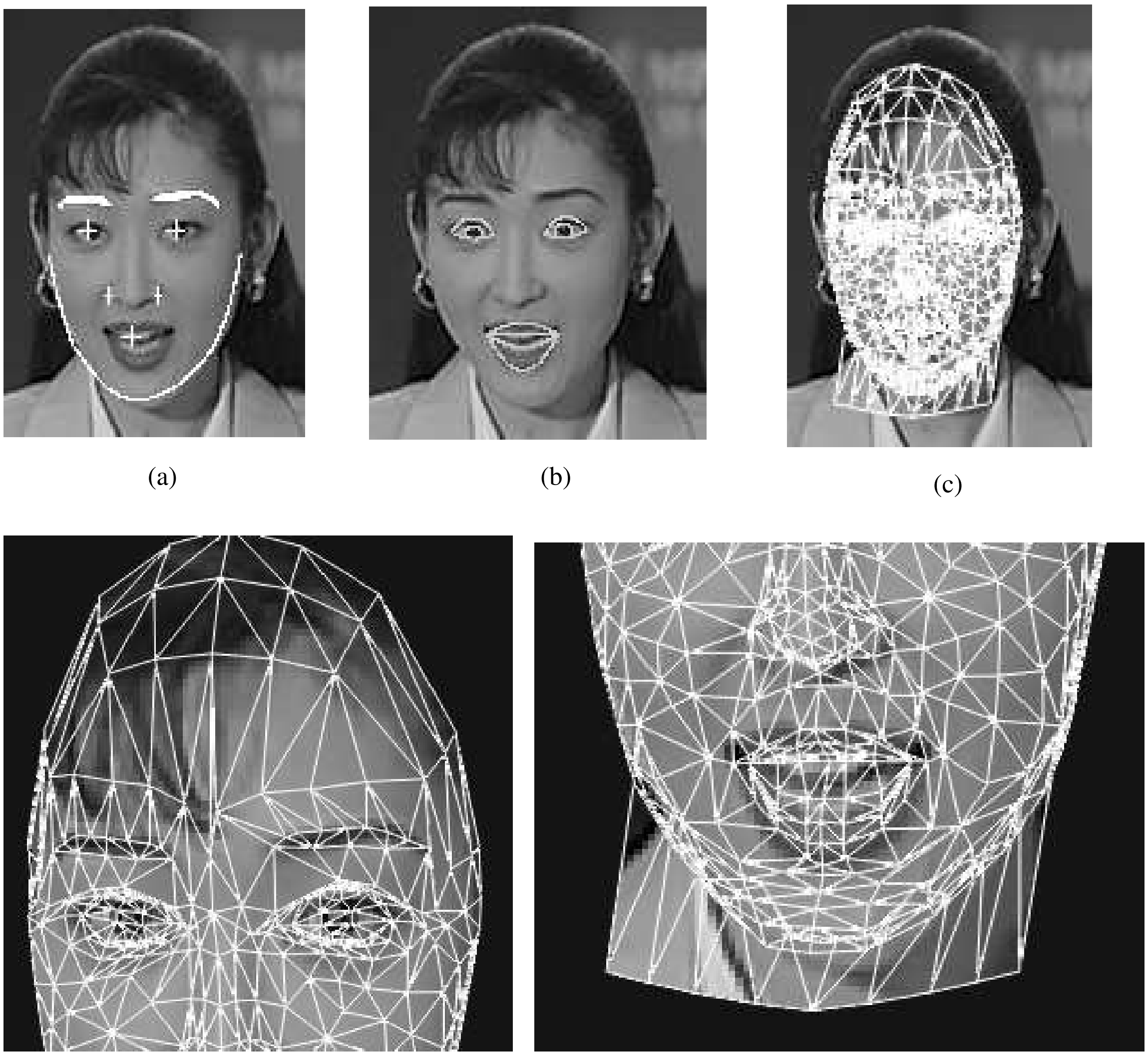
[15] V. Chanawatr, "Modellierung und Segmentierung von Nase und Augenbrauen in Bildtelefonsequenzen ", Studienarbeit, University of Hannover, June 1997.



(a) (b) (c)

(d) (e)

Fig. 6: Test sequence *Miss America* (CIF, 10Hz): (a) estimated eyes and mouth centers, chin and cheek contours, eyebrows and sides of the nose, (b) estimated eye lids and mouth lips, (c) adapted face model, (d) top part of the adapted face model, (e) bottom part of the adapted face model.



(a) (b) (c)

(d) (e)

Fig. 7: Test sequence *Akiyo* (CIF, 10Hz): (a) estimated eyes and mouth centers, chin and cheek contours, eyebrows and sides of the nose, (b) estimated eye lids and mouth lips, (c) adapted face model, (d) top part of the adapted face model, (e) bottom part of the adapted face model.

(a)

(b)

(c)

(d)

Fig. 8: Synthesis of facial expressions (*Miss America*): (a) original image; (b)(c)(d) images with various synthesized facial expressions.

(a)

(b)

(c)

(d)

Fig. 9: Synthesis of facial expressions (*Akiyo*): (a) original image; (b)(c)(d) images with various synthesized facial expressions.