AUTOMATIC ADAPTATION OF A HUMAN FACE MODEL FOR MODEL-BASED CODING

Mauricio Hess[1] *Geovanni Martinez*[2]

Image Processing and Computer Vision Research Lab (IPCV-LAB)
Escuela de Ingenieria Electrica, Universidad de Costa Rica
Codigo postal 2060, San Jose, Costa Rica
[1]mauhf@costarricense.cr, [2]gmartin@pacuare.eie.ucr.ac.cr

ABSTRACT

For coding of videophone sequences at very low bit rates, model-based coding is investigated. In a model-based coder, the human face in the videophone sequence is described by a three-dimensional (3D) face model. At the beginning of the videophone sequence, the model has to be adapted automatically to the shape, position and orientation of the real face present in the scene. In this paper, a new face model adaptation algorithm is presented. Facial features such as eye corners, mouth corners, chin and cheek borders, and nose corners are supposed to be known or previously estimated. First, the orientation, shape and position of the model are adapted using the positions of the mouth corners, outer corners of the eyes and top pixels of the cheek borders. Then, the shape of the model is locally adapted using the chin and cheek borders as well as the nose corners. The algorithm was implemented and applied to the first 40 frames of the videophone sequence *Claire* (CIF, 10Hz) with very encouraging results, even in cases when the person's eyes are closed or the person is facing downward.

1. INTRODUCTION

For coding of videophone sequences at very low bit rates, model-based coding is investigated [1, 2, 3, 4, 5, 6]. In a model-based coder, the human face in the videophone sequence is described by a three-dimensional (3D) face model. The shape is represented by a 3D wire-frame whose vertices represent the shape parameters. The motion is described by six parameters: three rotation angles and one 3D translation vector [7, 8]. The mimic is described by action units [9] or facial muscle parameters [10]. The color is described by projecting a real image onto the wire-frame. These parameters are estimated at the sender [10, 11, 12, 13, 14], then coded and transmitted to the receiver. After decoding of the transmitted parameters at the receiver, a synthetic image of the face is generated.

Usually, a generic 3D face model is used. This model has the proportions of an average human face. At the beginning of the videophone sequence, the model has to be adapted automatically to the size, position and orientation of the real face present in the scene. First, the face of the person is segmented in the current image. Then, facial features such as the eye centers, mouth center, chin and cheek borders, nose corners, etc. are extracted [15]. In the final step the model is adapted using the extracted facial features. The adaptation is carried out in two steps known as global adaptation and local adaptation.

In global adaptation, the size, position and orientation of the model are corrected in such a way that the projections onto the image plane of the model's eye and mouth centers coincide with the corresponding facial features. The model deformation is described by three scaling factors S_U, S_V and S_W along the *U*, *V* and *W* axes of the model's local coordinate system, respectively. The orientation of the model is described by three consecutive rotations around the *Y*, *Z* and *X* axes of the world coordinate system by the rotation angles R_Y, R_Z and R_X, respectively. The position of the model is described by the 3D position vector $\vec{G} = (G_X, G_Y, G_Z)^T$ of the origin of the model's local coordinate system respect to the world coordinate system.

In the professional literature, some algorithms are proposed for global adaptation of the face model. In [16], the scaling factor S_U is estimated as the quotient between the distance between the model's projected eye centers and the corresponding distance for the face in the current image. To estimate the scaling factor S_V, the midpoint between the model's projected eye centers must first be obtained and then the distance from this point to the projected mouth center is measured. The factor is estimated as the quotient between the measured distance for the model and the corresponding distance for the face in the current image. In [17], [18], [19] and [20], the scaling factor S_W is estimated as the average between the *U*-axis and *V*-axis estimated scaling factors. In [21], the rotation angle δR_Z for correction of the model orientation around the *Z*-axis of the world coordinate system is equal to the incline of the line which connects the eye centers of the face in the current image with respect to a line which is parallel to the horizontal axis of the image plane. In [18] and [19], the rotation an-

gle δR_Y for correction of the model orientation around the *Y*-axis of the world coordinate system is estimated in four steps. First, a line that crosses the eye centers of the face in the current image is traced. Then, the intersection point of this line with the right cheek border and the intersection with the left cheek border are found. Then, the midpoint between the two points found in the last step is obtained. Finally, the rotation angle δR_Y is estimated from the position difference between the point found in the last step and the midpoint between the eye centers. The rotation angle δR_X for correction of the model orientation around the *X*-axis of the world coordinate system is not estimated in any of the fore-mentioned contributions. All the above mentioned global adaptation algorithms fail when the eyes are closed because the eye centers are difficult to extract under those conditions.

In the local adaptation, a more refined adaptation of the face model to individual features of the face in the current image is carried out. In the professional literature, some algorithms are proposed for local adaptation of the face model. In [18] and [19], the face model is adapted to the chin and cheek borders of the face in the current image. First, each vertex whose projection onto the image plane coincides with the chin and cheek borders of the model (known as control vertices of the chin and cheek borders) is displaced until its projection coincides with the chin and cheek borders of the face in the current image. The remaining vertices are displaced in such a way that those vertices which are near control vertices are displaced a greater distance than those that are further away. In [19], the model's nose is scaled only along the *U*-axis by the scaling factor $S_{U,nose}$ and then displaced using the positions of the nose corners of the face in the current image. First, the distance between the projections of the model's nose corners is obtained. Then, the scaling factor $S_{U,nose}$ applied to the nose vertices along the *U*-axis is the quotient between the measured distance for the model and the corresponding distance for the face in the current image. Finally, the nose vertices are displaced with respect to the model's local coordinate system until the projections of the model's nose corners coincide with the nose corners of the face in the current image.

In this paper, a new face model adaptation algorithm is proposed. For global adaptation, the most significant contribution is that the model can be adapted when the eyes are closed and the person is facing downwards. For this, the outer corners of the eyes will be used instead of the eye centers, and a new method will be introduced to estimate the rotation angle δR_X for correction of the model orientation around the *X*-axis of the world coordinate system. For local adaptation, the most significant contribution is that a new method will be presented to estimate the scaling factor $S_{V,nose}$ of the model's nose vertices along the *V*-axis. The scaling factor $S_{U,nose}$ of the nose vertices along the *U*-axis and the displacement of the nose with respect to the model's local coordinate system are carried out using the method described in [19]. The adaptation of the model to the extracted chin and cheek borders is done using the method described in [18] and [19].

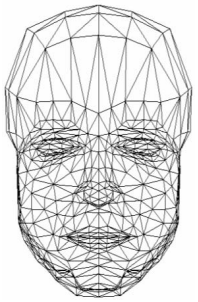


Fig. 1. *Robinson* face model.

The proposed adaptation algorithm will be applied to real videophone sequences to compare its performance with existing algorithms in the professional literature.

This paper is organized as follows. In section 2, the 3D face model is described. In section 3, the global adaptation algorithm is presented. In section 4, the local adaptation algorithm is described. In section 5 and in section 6, experimental results and the conclusions are given, respectively.

2. HUMAN FACE MODEL DESCRIPTION

In this paper, the *Robinson* face model is used (see Fig. 2). The 3D model belongs to a virtual 3D world, where (*X,Y,Z*) is the world coordinate system and (*U,V,W*) is the model coordinate system. Vector $\vec{G}$ describes the position of the model coordinate system with respect to the world coordinate system. Vector $\vec{H}$ describes the 3D position of any vertex i of the model with respect to the world coordinate system. The model's *shape* is described by the positions of all the vertices with respect to the model coordinate system. Vector $\vec{H}^*$ describes the position of vertex i with respect to the model coordinate system, where:

$$\vec{H} = R\vec{H}_i^* + G_i, \tag{1}$$

here *R* is the 3*3 rotation matrix that describes the orientation of the model with respect to the world coordinate system by three consecutive rotations around the *Y*, *Z* and *X* axes of the world coordinate system by the rotation angles R_Y, R_Z and R_X, respectively.

3. GLOBAL ADAPTATION

To correct the orientation, size and position of the model, nine *global adaptation parameters* are estimated: three rotation angles: δR_X, δR_Y and δR_Z which correct the model

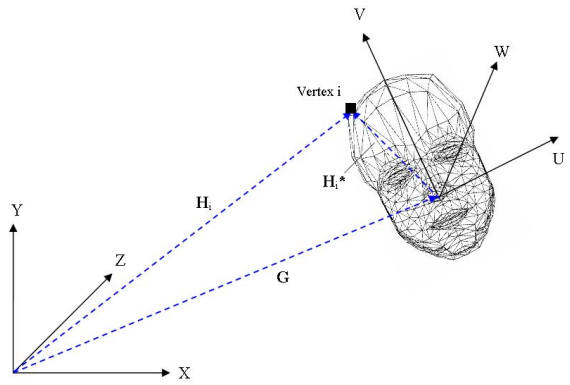
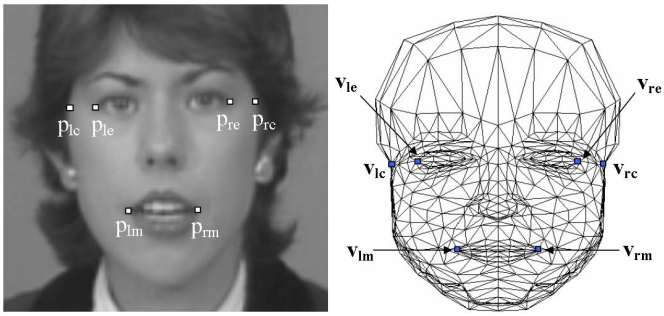


Fig. 2. *Robinson* face model inside the virtual 3D world.



(a) (b)

Fig. 3. Positions used to calculate the global adaptation parameters. (a) Face in the current image. (b) Model.

orientation respect to the world coordinate system, three scaling factors: S_U, S_V and S_W which correct the model shape size, and a 3D translation vector δG that corrects the model position with respect to the world coordinate system. To estimate the nine parameters, the positions of the mouth corners ($\vec{p_{lm}}$ and $\vec{p_{rm}}$ in Fig. 3(a)), outer corners of the eyes ($\vec{p_{le}}$ and $\vec{p_{re}}$ in Fig. 3(a)) and top pixels of the cheek borders ($\vec{p_{lc}}$ and $\vec{p_{rc}}$ in Fig. 3(a)) will be used, as well as the projections onto the image plane of the model's corresponding vertices (see Fig. 3(b)), using the model of perspective projection. Before adaptation the model is facing straight toward the camera, as shown in Fig. 3(b).

In the first step, δR_Y is estimated. First, the quotient between the absolute distance of the left outer corner of the eye and the top pixel of the left cheek border and the absolute distance of the right outer corner of the eye and the top pixel of the right cheek border of the face in the current image is computed as follows:

$$a_{Y,face} = \frac{|\vec{p_{lc}} - \vec{p_{le}}|}{|\vec{p_{rc}} - \vec{p_{re}}|}. \quad (2)$$

A similar quotient is calculated for the model:

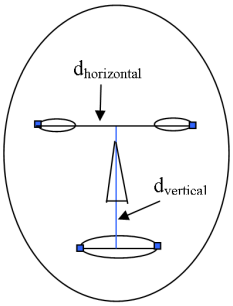


Fig. 4. Width and height of a real person's face.

$$a_{Y,model} = \frac{|\vec{v_{lc}} - \vec{v_{le}}|}{|\vec{v_{rc}} - \vec{v_{re}}|}. \quad (3)$$

If the quotient obtained in (2) is inside the interval $[0.92, 1.08]$, then the rotation angle δR_Y is set to 0 ($\delta R_Y = 0$), because the person is supposed to be facing straight towards the camera. The interval was defined heuristically for typical videophone sequences.

If the quotient obtained in (2) is outside the interval $[0.92, 1.08]$, then the model is rotated in steps of 0.5 degrees until the value of the quotient in (3) is very similar to the quotient calculated in (2). Finally, δR_Y corresponds to the amount of times the model was rotated multiplied by 0.5 degrees.

In the second step, δR_Z is estimated. This angle is supposed to be the angle formed by the line connecting both outer corners of the eyes with respect to a line which is parallel to the image plane's horizontal axis.

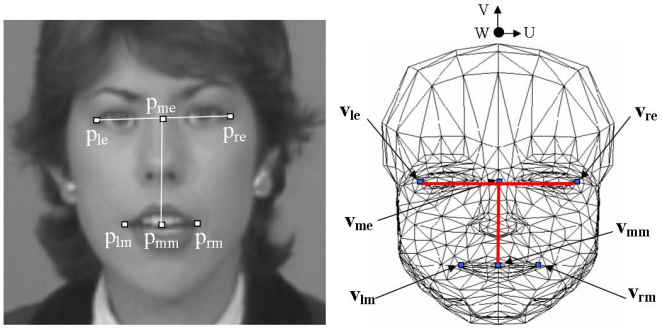
In the third step, δR_X is estimated. As a first approach, this value can only be estimated if previously δR_Y has been found to be 0 ($\delta R_Y = 0$). First, the mean quotient between the width and height of a face is estimated from measurements in real faces. To this end, the quotient c between the distance $d_{horizontal}$ from the outer corner of the right eye to the outer corner of the left eye (see Fig. 4) and the distance $d_{vertical}$ from the midpoint between the outer corners of the eyes and the midpoint between the mouth corners (see Fig. 4) was measured for each person of a large group of people:

$$c = \frac{d_{horizontal}}{d_{vertical}}. \quad (4)$$

The mean quotient between the width and height of a face is assumed to be the mean of c. Assuming a gaussian probability density, the estimated mean of c was $m_c = 1.375$.

Then, the quotient between the width and height of the face in the current image (see Fig. 5(a)) is computed:

$$a_{X,face} = \frac{|\vec{p_{re}} - \vec{p_{le}}|}{|\vec{p_{me}} - \vec{p_{mm}}|}. \quad (5)$$



(a) (b)

Fig. 5. (a) Width and height of the face in the current image. (b) Width and height of the model.

A similar quotient is computed for the model (see Fig. 5(b)):

$$a_{X,model} = \frac{|\vec{v_{re}} - \vec{v_{le}}|}{|\vec{v_{me}} - \vec{v_{mm}}|}. \quad (6)$$

If the quotient obtained in (5) is smaller than the estimated mean quotient between the width and height of a face $m_c = 1.375$, then the rotation angle δR_X is set to 0 (δR_X=0), because the person is supposed not to be facing downward. Currently, we apply a threshold of 1.44 because heuristically it works better for typical videophone sequences.

Then, the model is rotated in steps of 0.5 degrees until the value of the quotient in (6) is very similar to the value calculated in (5). Finally, δR_X is equal to the amount of times the model was rotated multiplied by 0.5 degrees.

In the fourth step, the scaling factor S_V is estimated. First, the distance between the midpoint between the outer corners of the eyes and the midpoint between the mouth corners is measured for the face in the current image (see Fig. 5(a)). Then the corresponding distance over the image plane is measured for the model (see Fig. 5(b)). Finally, the scaling factor is equal to the quotient between these two distances:

$$S_V = \frac{|\vec{p_{me}} - \vec{p_{mm}}|}{|\vec{v_{me}} - \vec{v_{mm}}|}. \quad (7)$$

In the fifth step, the scaling factor S_U is estimated. First, the distance between the outer corners of the eyes is measured for the face in the current image (see Fig. 5(a)). Then the corresponding distance over the image plane is measured for the model (see Fig. 5(b)). Finally, the scaling factor is equal to the quotient between these two distances:

$$S_U = \frac{|\vec{p_{re}} - \vec{p_{le}}|}{|\vec{v_{re}} - \vec{v_{le}}|}. \quad (8)$$

In the sixth step, the scaling factor S_W is estimated. This scaling factor is estimated as the average between the scaling factors along the *U*-axis and *V*-axis, as in [22], [16] and [20], since there is not enough information to calculate its value from a single frontal image:

$$S_W = \frac{1}{2}(S_U + S_V). \quad (9)$$

In the last step, the 3D translation vector $\delta G = (\delta G_X, \delta G_Y, \delta G_Z)^T$ is estimated. This vector describes the translation of the model from its current position to a new one where the projections of the vertices corresponding to the model's mouth corners and outer corners of the eyes coincide with the corresponding positions for the face in the current image. First, a 2D translation vector δg is estimated, which describes the distance between the midpoint between the mouth corners and outer corners of the eyes for the face in the current image and the projected midpoint between the model's corresponding vertices. Finally, assuming that $\delta G_Z = 0$, the components δG_Y and δG_Z of the translation vector are obtained by projecting the 2D translation vector δg into the 3D virtual world at a depth corresponding to the *Z*-coordinate of the midpoint between the model's mouth corners and outer corners of the eyes.

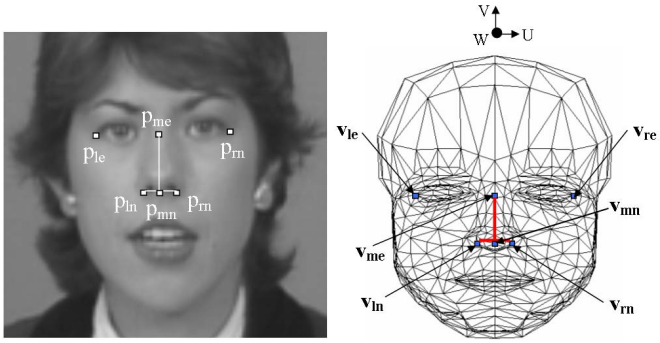
4. LOCAL ADAPTATION

In the local adaptation, first, the shape of the whole model is adapted using the chin and cheek borders, and then only the shape of the model's nose is adapted using the nose corners. The adaptation of the model using the chin and cheek borders is done using a very similar method to the one described in [18] and [19], so it will not be described here. The adaptation of the model's nose consists of a scaling along the *U*-axis and *V*-axis and then a displacement δG^*_{nose} with respect to the model's local coordinate system. In the first step, the scaling factor $S_{U,nose}$ of the nose vertices along the *U*-axis is estimated. First, the distance between the nose corners of the face in the current image is calculated (see Fig. 6(a)). Then, the corresponding distance over the image plane is calculated for the model (see Fig. 6(b)). Finally, the scaling factor is estimated as follows:

$$S_{U,nose} = \frac{|\vec{p_{ln}} - \vec{p_{rn}}|}{|\vec{v_{ln}} - \vec{v_{rn}}|}. \quad (10)$$

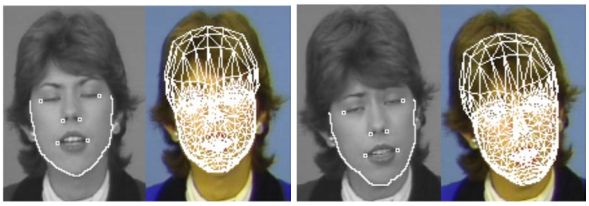
In the second step, the scaling factor $S_{V,nose}$ of the nose vertices along the *V*-axis is estimated. First, the distance between the midpoint between the outer corners of the eyes and the midpoint between the nose corners is calculated for the face in the current image (see Fig. 6(a)). Then, the corresponding distance over the image plane is calculated for the model (see Fig. 6(b)). Finally, the scaling factor is computed as:

$$S_{V,nose} = \frac{|\vec{p_{me}} - \vec{p_{mn}}|}{|\vec{v_{me}} - \vec{v_{mn}}|}. \quad (11)$$



(a) (b)

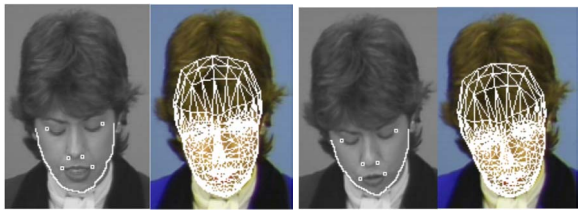
Fig. 6. (a) Width and height of the nose in the current image. (b) Width and height of the model's nose.



(a) (b)

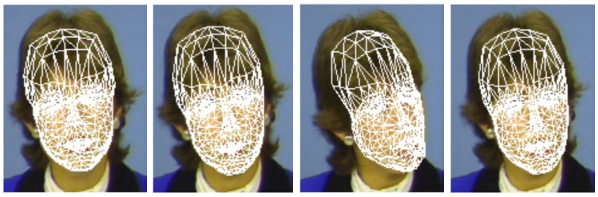
Fig. 7. Adaptation in frames of the test sequence *Claire* where the eyes are closed. (a) 4^{th} frame. (b) 29^{th} frame.

In the last step, the nose vertices are displaced with respect to the model's local coordinate system until the projections of the model's nose corners coincide with the nose corners of the face in the current image. First, a 2D displacement vector is obtained, which describes the distance between the midpoint between the nose corners of the face in the current image and the projected midpoint between the model's corresponding vertices. Then, the 3D displacement vector of the nose vertices δG^*_{nose} is obtained by projecting the 2D displacement vector onto plane N. Plane N passes through the midpoint between the model's nose corners and is perpendicular to the *W*-axis.



(a) (b)

Fig. 8. Adaptation in frames of the test sequence *Claire* where the person is facing downward. (a) 71^{st} frame. (b) 106^{th} frame.



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

Fig. 9. Examples of the *Robinson* model adapted in images of the test sequence *Claire*. (a) 8^{th} frame, (b) 20^{th} frame, (c) 33^{rd} frame, (d) 38^{th} frame.

5. EXPERIMENTAL RESULTS

The proposed algorithm was implemented in C and applied to the first 40 frames of the test sequence *Claire* (CIF, 10 Hz). Using an Intel Pentium IV processor at 1.6 GHz, with 256 MB of RAM and under Windows XP, the mean processing time was 0.015 s.

The reliability of the proposed algorithm was measured to evaluate its performance. To measure the reliability, it was verified if the model was correctly adapted in images where the eyes are closed or the person is facing downward, as well as in other images where the person is not facing straight toward the camera. According to the experimental results, the algorithm was correctly adapted when the person's eyes are closed (see Fig. 7) or the person's head is facing downward (see Fig. 8). Some examples of other obtained results are depicted in Fig. 9.

6. SUMMARY AND CONCLUSIONS

A new automatic face model adaptation algorithm for model-based coding of videophone sequences has been presented, which uses the known or previously extracted positions of the mouth corners, the outer corners of the eyes, the nose corners and the chin and cheek borders. The face model is globally adapted by first rotating, then scaling and finally translating the model using relations between the positions of the mouth corners, outer corners of the eyes and top pixels of the cheek borders. Then, the model is locally adapted to the chin and cheek borders and the nose vertices are locally scaled and displaced. The main contribution of this paper is that the face model can be adapted when the person is facing downward or the person's eyes are closed. The algorithm was applied to 40 frames of the test sequence *Claire* (CIF, 10 Hz). The experimental results show that the reliability and overall quality of the adaptation algorithm are very high, even in cases when the person's eyes are closed or the person is facing downward.

7. ACKNOWLEDGMENTS

The authors are grateful to Prof. H.-G. Musmann for encouraging this work.

8. REFERENCES

[1] K. Aizawa and T. Huang, "Model-based coding: Advanced video coding techniques for very low bit-rate applications," *Proceedings of the IEEE*, vol. 83, pp. 259-271, February 1995.

[2] R. Forchheimer, O. Fahlander, and T. Kronander, "A semantic approach to the transmission of face images," in *Picture Coding Symposium*, Cesson-Sevigne, France, July 1984, paper 10.5.

[3] H. Li, A. Lundmark, and R. Forchheimer, "Image sequence coding at very low bit-rates: A review," *IEEE Transactions on Image Processing*, vol. 3, pp. 589-609, September 1994.

[4] H. Musmann, "A layered coding system for very low bit-rate video coding," *Signal Processing: Image Communications*, vol. 7, pp. 267-278, November 1995.

[5] D. Pearson, "Developments in model-based video coding," *Proceedings IEEE*, vol. 83, pp. 892-906, June 1995.

[6] W.J. Welsh, S. Searby, and J.B. Waite, "Model-based image coding," *British Telecom Technology Journal*, vol. 8, no. 3, pp. 94-106, July 1990.

[7] G. Martinez, "Maximum-likelihood motion estimation of a human face," in *IEEE International Conference on Multimedia and Expo*, Baltimore, USA, 2003, pp. 577-580.

[8] G. Martinez, "Improving the speed of covergency of a maximum-likelihood motion estimation algorithm of a human face," in *IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, 2004.

[9] P. Ekman and V.W. Friesen, "Facial action coding system," Palo Alto, USA, 1978, Consulting Psychologists Press Inc.

[10] J. Fischl, B. Miller, and J. Robinson, "Parameter tracking in a muscle-based analysis-synthesis coding system," in *Picture Coding Symposium*, Lausanne, Switzerland, March 1993, paper 2.3.

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

[12] C.S. Choi, K. Aizawa, H. Harashima, and T. Takebe, "Analysis and synthesis of facial image sequences in model-based image coding," *IEEE Transactions on Circuit, Systems and Video Technology*, vol. 4, pp. 257-275, June 1994.

[13] I. Essa and A. Pentland, "Coding, analysis, interpretation and recognition of facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 757-763, July 1997.

[14] H. Li, P. Roivainen, and R. Forchheimer, "3d motion estimation in model-based facial image coding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 545-556, June 1993.

[15] M. Hess and G. Martinez, "Facial feature extraction based on the smallest univalue segment assimilating nucleus (susan) algorithm," in *Picture Coding Symposium*, San Francisco, California, December 2004.

[16] R. Koch, "Adaptation of a 3d facial mask to human faces in videophone sequences using model-based image analysis," in *Picture Coding Symposium*, Tokyo, Japan, September 1991, pp. 285-288.

[17] L. Zhang, "Automatic adaptation of a face model using action units for semantic coding of videophone sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 6, pp. 781-795, October 1998.

[18] M. Kampmann and R. Farhoud, "Precise face model adaptation for semantic coding of videophone sequences," in *Picture Coding Symposium*, Berlin, Germany, September 1997.

[19] M. Kampmann, "Automatic 3d face model adaptation for model-based coding of videophone sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 3, March 2002.

[20] M. Reijnders, P. van Beek, B. Sankur, and J.C.A. van der Lubbe, "Facial feature localization and adaptation of a generic face model for model-based coding," *Signal Processing: Image Communication*, vol. 7, pp. 57-75, March 1995.

[21] M. Kampmann and 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.

[22] M. Wollborn, M. Kampmann, and R. Mech, "Content-based coding of videophone sequences using automatic face detection," in *Picture Coding Symposium*, Berlin, Germany, September 1997.