Picture Coding Symposium, pp. 33-38, Berlin, September 1997.

Facial Expression Analysis for Model-Based Coding of Video Sequences

Peter Eisert and Bernd Girod

Telecommunications Institute, University of Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
Email: {eisert,girod}@nt.e-technik.uni-erlangen.de

ABSTRACT

In this paper we present an algorithm for the estimation of 3D motion and the analysis of facial expressions of a speaking person. To constrain the flexible body motion, a 3D model of the person is used specifying shape, texture and facial motions. The surface of the model is built with triangular B-splines leading to a small number of control points that determine the appearance of the person. A set of animation parameters characterizes the facial expression. The proposed algorithm estimates changes in the animation parameters between two successive frames using a hierarchical optical flow based method. The computational requirement for this solution is low due to the linear structure of the algorithm.

1. INTRODUCTION

Model-based coding is a promising approach for very low bit-rate video compression [1, 2, 3]. Motion parameters of objects are estimated from video frames using three-dimensional models of the objects. These models describe the shape and texture of the objects. At the decoder, the video sequence is synthesized by rendering the models at the estimated positions.

The head of a speaking person cannot be modeled as a rigid body. Local deformations due to facial expressions must be taken into consideration when analyzing the 3D facial motion [4, 5]. It is assumed that the facial expression can be represented by a linear combination of small elementary local movements. These movements are described by facial animation parameters (FAPs). Examples of face parameterization are the Facial Action Coding System [6] and the system of the MPEG-4 SNHC group [7] that is used in this work.

In our coder all these parameters are estimated simultaneously using a hierarchical optical flow based method. The optical flow constraint is combined with the parameterized 3D motion equations for each object point. The determination of the motion as a function of the FAPs is simplified due to the use of triangular B-splines for the head surface construction.

The paper is organized as follows. We first describe the head model that is used for the 3D facial expression analysis. In section 3 the camera model and the basic geometry of our virtual scene is shown. We then present the optical flow based algorithm for the global and local motion estimation and finally experimental results are given.

2. HEAD MODEL

The three-dimensional scene used for parameter estimation and rendering of the synthetic images consists of a camera model and a head model. The head model specifies the 3D shape and facial expressions of a speaking person but also constrains the motion due to local deformations. Like other well-known facial models [8, 9], our proposed model also consists of a number of triangles onto which texture is mapped to obtain a photorealistic appearance. The shape of the surface is, however, defined by second order triangular B-splines [10, 11] to make the modeling of facial expressions easier.

2.1. Triangular B-Splines

The set of vertices that define the shape of our model has a very large number of degrees of freedom. Therefore, very complex objects can be modeled with sharp edges and discontinuities. The face of a person, however, has a smooth surface and facial expressions result in smooth movements of surface points due to the anatomical properties of tissue and muscles. These restrictions on curvature and motion can be modeled by splines which satisfy specific continuity constraints. It has been shown [12] that B-splines are well suited for the modeling of facial skin. This type of splines exhibits some interesting properties useful for the implementation of the head model:

- Smoothness:
  A B-spline of order n is C^{n-1}-continuous. For our model we use second order splines leading to a C^1-continuity of the surface.

- Local control:
  Movement of single control points influence the surface just in a local neighborhood which simplifies the modeling of facial expressions.

- Affine invariance:
  An affine transformation of the surface can be obtained by applying the same transformation on the control points. Facial movements are now defined by the transformation of a small number of control points instead of applying transformations on each vertex which reduces the computational complexity.

Normal B-splines or NURBS which are commonly used in computer graphics suffer from one well-known drawback. They are defined on a rectangular topology and therefore do not allow a local refinement in areas that are more curved. To overcome this restriction, triangular B-splines [10] are used. This new spline scheme is based on triangular patches which can easily be refined while still preserving the above mentioned properties of normal B-splines. For rendering we approximate the smooth spline and subdivide each patch into a number of flat triangles. The number of triangles can be varied to get either a good approximation or higher frame rate during rendering. The resulting triangular mesh is defined by a discrete set of vertices v_j on the mathematically defined spline surface. Only at these discrete positions on the surface we have to compute the B-spline basis functions [11] which can be done off-line. Once the basis functions are determined the position of the vertices can be calculated by

$$\mathbf{v}_j = \sum_{i \in I_j} N_{ji} \mathbf{c}_i, \text{ such that } \sum_{i \in I_j} N_{ji} = 1 \qquad (1)$$

with N_{ji} being the precalculated basis functions of vertex j, I_j the index set of vertex j and $\mathbf{c_i}$ the i th control point.

2.2. Generic Head Model

Our face model is a generic 3D model consisting of 101 triangular B-patches. Teeth and the interior of the mouth are separate 3D objects. The topology of the patches (Figure 1) which is based on the Candide model [9] and the definition of the facial animation parameter units remain fixed for all persons. Only the texture and the position of the control points are changed to adjust the model to an individual person. The shape and texture information for each person is obtained from an individual 3D laser scan.

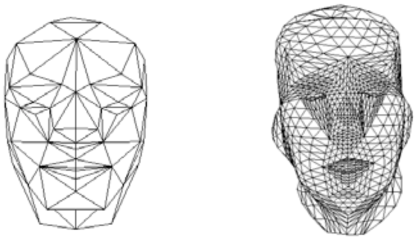

Figure 1: B-patches (left) and subdivided, shaped triangular mesh (right).

For the estimation of facial parameters we must be able to animate our model to create different facial expressions. This task is simplified due to the use of splines because they already constrain the motion of neighboring vertices. For the parameterization of facial expressions the proposal of the MPEG-4 SNHC group [7] was chosen. According to that scheme, every facial expression can be generated by a superposition of 68 action units. These include both global motion like head rotation and local motion like eye or mouth movement. 46 of these 68 FAPs are currently implemented. To model the local movements, our generic face model contains a table for each action unit describing how the control points of the mesh are translated or rotated. This is done only for the small number of control points. For a given set of facial animation parameters specifying an expression of a person, all positions of the control points are updated and then the vertices are computed according to (1) by a linear combination of the control points. Examples of such rendered expres-

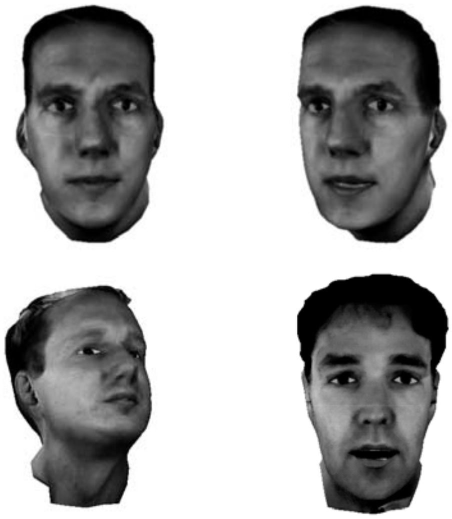

Figure 2: Illustration of different synthesized images.

sions for different persons are depicted in Figure 2.

3. CAMERA MODEL

The second model in our virtual scene is the camera model that describes the projection of the 3D world into the 2D image plane. Here, a perspective projection is assumed that can be described by

$$\begin{aligned} X_p &= X_0 - f_x \frac{x}{z} \\ Y_p &= Y_0 - f_y \frac{y}{z}, \end{aligned} \qquad (2)$$

where x, y and z are the 3D coordinates of an object point and X_p, Y_p the corresponding pixel coordinates in the image plane. f_x and f_y denote the focal length multiplied by scaling factors and X_0 and Y_0 the image center and its translation from the optical axis due to inaccurate placement of the CCD-sensor in the camera. All these coordinates and coordinate systems are shown in Figure 3. For simplicity, normalized pixel coordinates X_n and Y_n are introduced

$$X_n = \frac{X_p - X_0}{f_x}\ ,\ Y_n = \frac{Y_p - Y_0}{f_y}. \qquad (3)$$

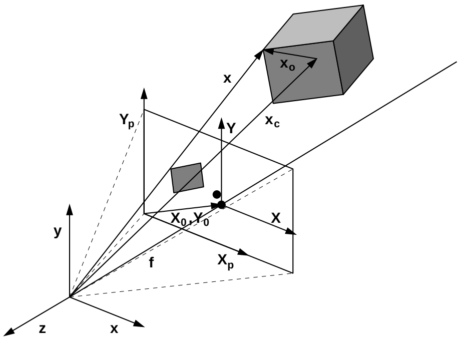

Figure 3: Scene geometry.

4. ESTIMATION OF FACIAL PARAMETERS

In [13] we propose an algorithm for the estimation of rigid body motion from monocular video sequences. This algorithm is now extended to estimate also motion and deformation of flexible bodies. The possible deformations must be restricted by an appropriate model. For the application of facial expression analysis the constraints are given by our head model.

The algorithm uses the whole image by setting up the optical flow constraint equation at each pixel position that contains the head object

$$I_X \cdot u + I_Y \cdot v + I_t = 0 \qquad (4)$$

where $[I_X\ I_Y]$ is the gradient of the intensity at point $[X\ Y]$, u and v the velocity in x- and y-direction and I_t the intensity gradient in temporal direction.

We estimate the facial animation parameters from (4) together with the 3D motion equations of the head's object points [4] without using additional smoothness constraints. For a rigid body motion the motion equation can easily be written in terms of a translation and a rotation. When allowing the body to be flexible, this is no longer possible and the motion equation must be set up at each object point independently. The trajectory of an object point is, however, not independent of its neighbors but is constrained by the head model that describes motion of surface points as a function of facial animation parameters.

The motion of the head is modeled by moving single control points. The new control point position $\mathbf{c}'$ in the 3D space can be determined from the position $\mathbf{c}$ of the previous frame by

$$\mathbf{c}' = \mathbf{c} + \sum_k a_k \mathbf{d}_k \qquad (5)$$

where a_k's are the changes of the facial animation parameters between the two frames that have to be estimated by the algorithm and $\mathbf{d}_k$ the 3D direction vectors of the corresponding movement. The vectors $\mathbf{d}_k$ corresponding to the control point $\mathbf{c}$ are defined in the generic head model and describe how the control point is translated for a specific value of a facial expression parameter k.

Having modeled the shift in control points, the motion of the vertices of the triangular mesh can be determined using (1) and the local object motion is calculated from that using

$$\mathbf{x} = \sum_{m=0}^{2} \lambda_m \mathbf{v}_m = \sum_{j \in J} (\sum_{m=0}^{2} \lambda_m N_{mj}) \mathbf{c}_j, \qquad (6)$$

where λ_m are the barycentric coordinates of the object point in the triangle that encloses that point. The motion equation for a surface point can be represented as

$$\mathbf{x}' = \mathbf{x} + \sum_k a_k \mathbf{t}_k = \mathbf{x} + T \cdot \mathbf{a}, \qquad (7)$$

where $\mathbf{t}_k$'s are the new direction vectors to the corresponding facial animation parameter calculated from $\mathbf{d}_k$ by applying the linear transforms (1) and (6). T combines all the vectors in a single matrix. Let $\mathbf{t}_x$, $\mathbf{t}_y$ and $\mathbf{t}_z$ be the row vectors of this matrix. Combining (7) with the the camera model (2) and using a first order approximation leads to

$$\begin{aligned} X' - X &\approx -\frac{f_x}{z}(\mathbf{t}_x + X_n \mathbf{t}_z)\mathbf{a} \\ Y' - Y &\approx -\frac{f_y}{z}(\mathbf{t}_y + Y_n \mathbf{t}_z)\mathbf{a} \end{aligned} \qquad (8)$$

for the 2D point correspondences. Together with (4) a linear equation at each pixel position can be set up

$$\frac{1}{z}\left[I_X f_x \mathbf{t}_x + I_Y f_y \mathbf{t}_y + (I_X f_x X_n + I_Y f_y Y_n)\mathbf{t}_z\right]\mathbf{a} = I_t \tag{9}$$

with z being the depth information coming from the model. We obtain an overdetermined system that can be solved in a least-squares sense for the unknown facial animation parameters **a**. The size of the system depends directly on the number of implemented FAPs.

Due to the first order approximation of the optical flow constraint and the projected motion equations that we use to obtain a linear set of equations, we can deal only with restricted motion parameters between two successive frames. To avoid this problem we use a hierarchical approach that starts with subsampled versions of the images (88 by 72 pixels) to get a rough estimate of the global motion parameters. The synthetic image is then motion compensated by moving the 3D model according to the estimated values. Then, the estimation of all FAPs is performed on higher resolution images leading to more and more accurate animation parameters. Currently three different levels of resolution are used with a final resolution of 352 by 288 pixels.

5. EXPERIMENTAL RESULTS

The proposed algorithm was tested on both synthetic and real video sequences. First a number of synthetic images with well defined facial animation parameters are rendered. For every pair of images the estimated facial parameters are compared with the correct values. The viewing angle of the camera is about 25^o. The result of one such experiment can be seen in Table 1. Ten animation parameters are changed, the others remain constant. The chosen parameters (given in parenthesis) correspond to global motion of the head (FAP 48,49,50), opening of the jaw (3), movement of eyelids (19,20), eyebrows (35,36) and lip corners (12,13). The FAPs in the table are normalized with respect to their maximum allowed value in order to get values reaching from -1 to 1 (0 corresponds to a neutral expression). Over several frames the average absolute error of the estimated facial parameters is about 0.06% of the maximum value of the parameters.

These very accurate results for synthetic images are also observed in a second experiment. Here a synthetic sequence of a talking head is taken as the input of our estimation algorithm. The facial animation parameters are estimated for every 5th frame and these parameters are then used to reconstruct the sequence by rendering the 3D model. First, only the global motion (head translation and rotation) is estimated and then all FAPs are determined. In Figure 4 the PSNR between the resulting reconstructed image and the original one is plotted over time. The high values of about 70 dB show that the images are nearly identical.

FAP	3	12	13	19	20
Org.	0.400	0.600	-0.700	0.900	0.700
Est.	0.400	0.600	-0.700	0.905	0.699
FAP	35	36	48	49	50
Org.	0.500	-0.500	0.200	-0.100	0.300
Est.	0.500	-0.500	0.200	-0.100	0.300

Table 1: Estimated facial animation parameters in comparison to the correct values.

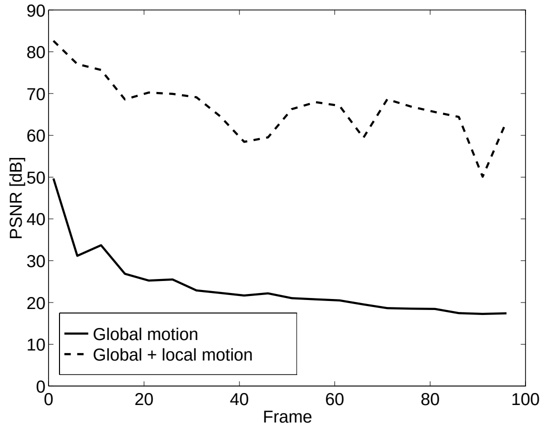

Figure 4: PSNR after global and global + local motion compensation.

In another experiment a video sequence of a talking person is recorded with a camera at a frame rate at 12.5 Hz and CIF resolution. Again, the facial animation parameters are estimated between two successive frames. In the first frame the model is roughly adjusted by estimating translation and rotation from feature correspondences like eyes and mouth corners. The 3D head model is then animated according to the estimated values and put into a virtual environment. Four frames of the original and the reconstructed video sequence are shown in Figure 6. The PSNR between original and synthetic image is calculated for the facial area and shown in Figure 5. We get an average PSNR of about 26.4 dB. This value is mainly influenced by some model failures. Especially in the mouth area there are some larger differences because the head model used in this experiment contains no teeth.

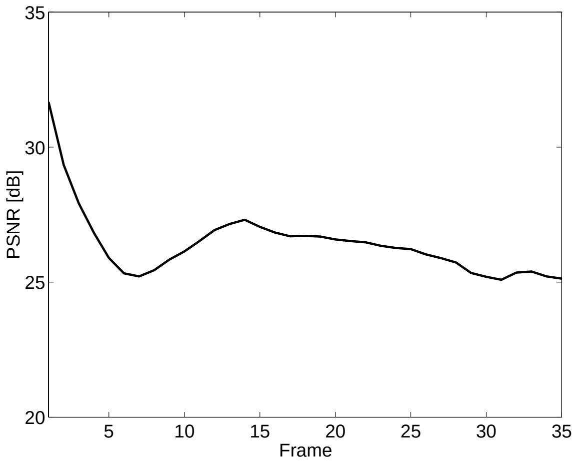

Figure 5: PSNR of the facial area.

If this scheme is used for video coding the resulting bit rate needed for transmission is of high interest. To get an estimate for the bit-rate, the differences of the estimated facial expression parameters between two successive frames are quantized with 6 bits and Huffman coded. This results in an average bit-rate of about 0.57 kbit/s at a frame rate of 12.5 Hz.

6. CONCLUSIONS

In this paper we have presented an algorithm for the estimation of facial animation parameters from monocular video sequences. The approach is model-based and uses a 3D model specifying shape and texture of a head. The surface of this model is built by triangular B-splines to simplify the modeling of facial expressions. These expressions are described by a set of facial parameters according to MPEG-4 syntax that are estimated by our gradient based motion estimator. Experimental results show that the estimation works well for both synthetic and real video sequences and that very low data rates of less than 1 kbit/s can be achieved for the transmission of head-and-shoulder scenes.

7. REFERENCES

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

[2] K. Aizawa and T. S. Huang, "Model-based image coding: Advanced video coding techniques for very low bit-rate applications", *Proceedings of the IEEE*, vol. 83, no. 2, pp. 259–271, Feb. 1995.

[3] D. E. Pearson, "Developments in model-based video coding", *Proceedings of the IEEE*, vol. 83, no. 6, pp. 892–906, June 1995.

[4] H. Li, P. Roivainen, and R. Forchheimer, "3-D motion estimation in model-based facial image coding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 545–555, June 1993.

[5] J. Ostermann, *Analyse-Synthese-Codierung basierend auf dem Modell bewegter, dreidimensionaler Objekte*, VDI Reihe 10, Nr. 391, VDI-Verlag, 1995.

[6] P. Ekman and W. V. Friesen, *Facial Action Coding System*, Consulting Psychologists Press, Inc., Palo Alto, 1978.

[7] MPEG-4, *SNHC Verification Model 4.0, Document N1666*, Apr. 1997.

[8] F. I. Parke, "Parameterized models for facial animation", *IEEE Computer Graphics and Applications*, vol. 2, no. 9, pp. 61–68, Nov. 1982.

[9] M. Rydfalk, *CANDIDE: A Parameterized Face*, PhD thesis, Linköping University, 1978, LiTH-ISY-I-0866.

[10] G. Greiner and H. P. Seidel, "Modeling with triangular B-splines", *ACM/IEEE Solid Modeling Symposium*, pp. 211–220, 1993.

[11] G. Greiner and H. P. Seidel, "Splines in computer graphics: Polar forms and triangular B-spline surfaces", *Eurographics*, 1993, State-of-the-Art-Report.

[12] M. Hoch, G. Fleischmann, and B. Girod, "Modeling and animation of facial expressions based on B-splines", *Visual Computer*, vol. 11, pp. 87–95, 1994.

[13] P. Eisert and B. Girod, "Model-based 3D motion estimation with illumination compensation", in *Proc. International Conference on Image Processing and its Applications*, Dublin, Ireland, Jul. 1997, vol. 1, pp. 194–198.

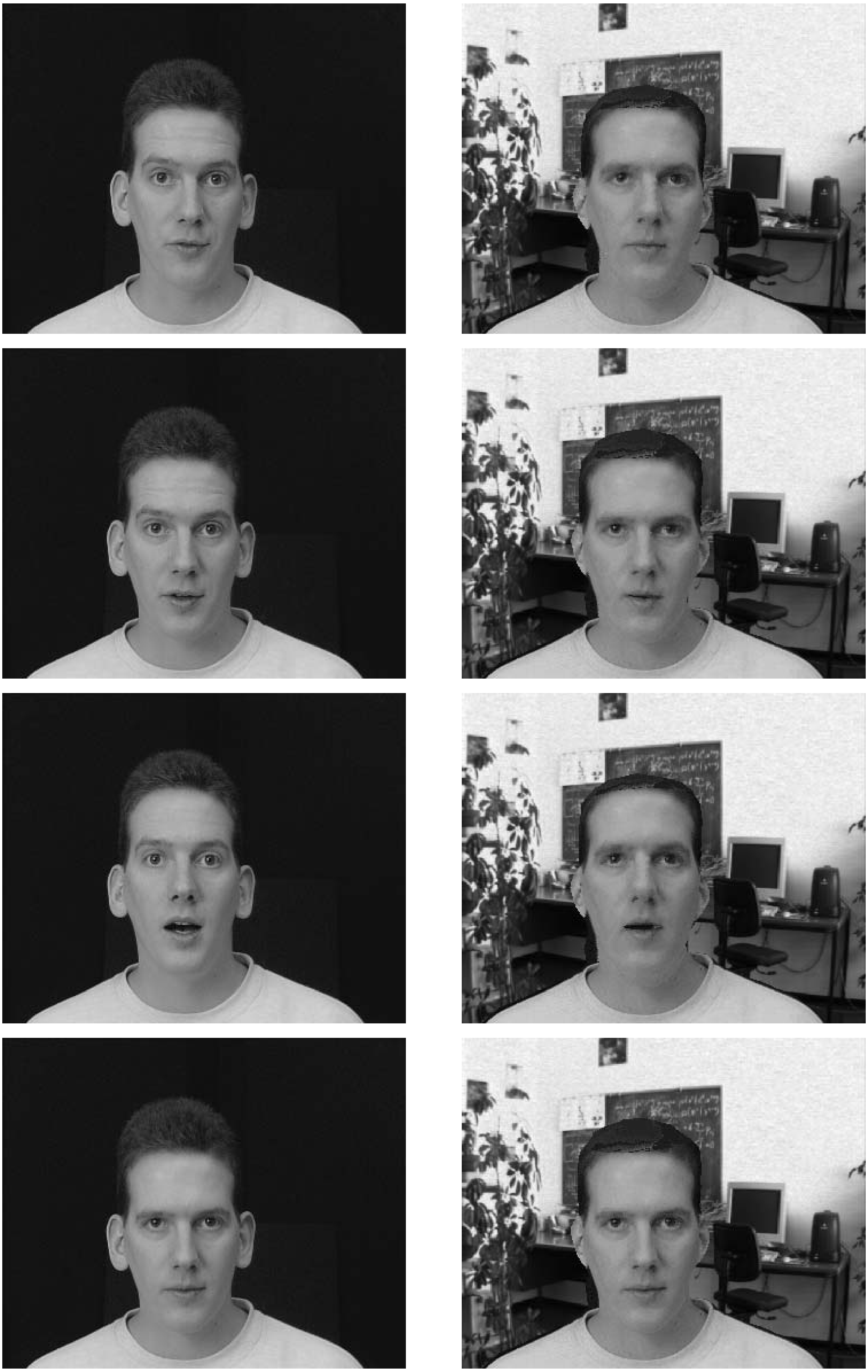

Figure 6: Four frames from a recorded talking head video sequence (left) and corresponding synthetically rendered 3D model that is put in a virtual office (right).