Geometry-Assisted Image-based Rendering for Facial Analysis and Synthesis

Peter Eisert * and Jürgen Rurainsky

Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute
Einsteinufer 37, D-10587 Berlin

Abstract

In this paper, we present an image-based method for the tracking and rendering of faces. We use the algorithm in an immersive video conferencing system where multiple participants are placed in a common virtual room. This requires viewpoint modification of dynamic objects. Since hair and uncovered areas are difficult to model by pure 3-D geometry-based warping, we add image-based rendering techniques to the system. By interpolating novel views from a 3-D image volume, natural looking results can be achieved. The image-based component is embedded into a geometry-based approach in order to limit the number of images that have to be stored initially for interpolation. Also temporally changing facial features are warped using the approximate geometry information. Both geometry and image cube data are jointly exploited in facial expression analysis and synthesis.

Key words:
facial animation, image-based rendering, model-based coding, face tracking

Image-based rendering is a technique which has received a considerable interest in computer graphics for the realistic rendering of complex scenes. Instead of modeling shape, material, reflection of objects as well as light sources and light exchange with high accuracy and sophisticated physical models, image-based rendering synthesizes new views of a scene by interpolating among multiple images taken with one or multiple cameras. Examples of such approaches are lightfields (1) or concentric mosaics (2; 3). The use of real pictures leads to naturally looking scenes and allow the reproduction of fine structures (e.g., hair, fur, leaves) that are difficult to model with polygonal representations. Also,

*

Email address: eisert@hhi.fhg.de (Peter Eisert).
URL: http://iphome.hhi.de/eisert (Peter Eisert).

the rendering complexity is independent from the scene content since interpolation is performed on pixels instead of polygons. As a result, sophisticated scenes can naturally be rendered with limited computational complexity.

One drawback of image-based rendering, however, is the high demand on storage and memory capacity. In order to allow free navigation and to avoid rendering artifacts, a very large number of images has to be captured, stored, and used for interpolation. Datasets of hundreds of giga bytes overstrain even today's computers. However, in image-based rendering, it is possible to trade the number of required images with the accuracy of an additionally used approximate geometry model (4). The more geometry information is used, the less images are needed for a particular quality. No geometry information requires many images whereas a highly accurate model is needed if only a few images are exploited. One extreme is a textured polygonal model with an accurate geometry and a single image as texture map. Other approaches that combine geometry and image information are, e.g., the lumigraph (5) and surface light fields (6).

Image-based rendering techniques are most often used for synthesizing new views from a static scene. In order to reduce the amount of data, the temporal dimension of the 7-dimensional plenoptic function (7) and other parameters in the scene are usually neglected. However, image-based rendering is not restricted to movements of a virtual camera but any degree of freedom can be represented by sampling from stored data.

In this paper, we use image-based techniques in order to realistically animate faces. Face animation, however, has traditionally been addressed by deforming and moving 3-D geometry models. A triangle mesh defines the person's shape (8; 9) and texture mapping ensures visual quality (10; 11). With extensive use of computer graphics techniques, highly realistic head models can be realized (12). Advantages of this approach are a compact description which can be exploited in model-based video coding (13; 14; 15) and simple facial animation capabilities by locally moving vertices. However, the realistic rendering of hair and the dynamic representation of facial features is not easy to solve by pure geometry modeling.

Researchers have therefore tried for a long time to use previously captured images in order to increase the quality of the synthesized views. Especially the facial features like eyes and mouth have been modeled by images. An early approach in model-based coding is the clip-and-paste method (16; 14; 17) where facial expressions are synthesized by copying templates of facial features from a codebook onto a 3-D model representing global head motion. All variations must be stored in a database which can grow significantly if visual artifacts and jitter shall be avoided. The number of templates can be reduced by adding model information to account for motion compensation. Active appearance models (18; 19), e.g., describe facial feature changes by a combination of image movements and dynamic textures and are successfully employed for realistic speech driven visual synthesis (20; 21; 22; 23). Even more geometry information is exploited in morphable head models (24) which interpolate new views from a database of 3-D head scans specifying both geometry and texture information.

In contrast to the above mentioned approaches, we combine geometry warping with image-based rendering in order to describe global head motion and to render a correct outline even in presence of hair. In order to reduce the memory requirements, only head turning with the most dominant image changes is interpolated from a set of initially captured views, whereas other global head motions are represented with a geometry model. Similarly, the jaw movement which affects the silhouette of the person viewed from the side is also represented by geometry deformations. In contrast to (25), local expressions and motion of the mouth and eyes are directly extracted from the video, warped to the correct position using the 3-D head model, and smoothly blended into the global head texture. The additional use of geometry in image-based rendering severely restricts the number of images required but enables head rotation of the person as a postprocessing step in applications like virtual conferencing.

The reminder of the paper is structured as follows. First, we describe possible applications for the presented approach, which are later used as reference for experimental results. We then show the method for pure model-based facial analysis and synthesis which is used for tracking of the face and initialization of the image-based dataset. In Section 3 we then present all extensions for image-based rendering and the modifications to the tracking algorithm that ensures robust estimation in the long term run. Experimental results finally show the applicability and accuracy of the approach.

1 Facial Analysis and Synthesis for Model-based Coding and Virtual Conferencing

Although the algorithms for rendering and tracking can be used for any application related to facial animation like text-driven animation (26), man-machine interfaces, and avatar control (27), we focus in this context on the application of virtual conferencing which has some implications on the settings and the experiments made.

In virtual conferencing, multiple distant participants can meet in a virtual room as shown in Fig. 1. The use of a synthetic 3-D computer graphics scene allows more than two partners to join the discussions even if they are far apart from each other. Each partner is recorded by a single camera and the video

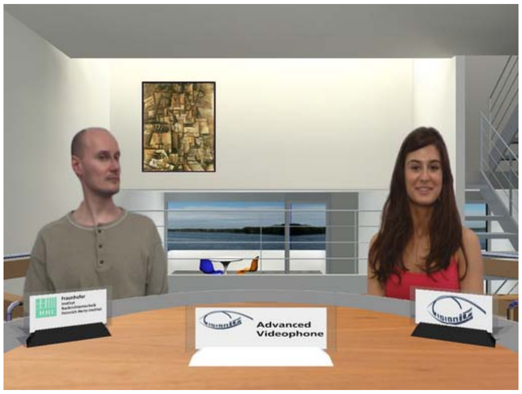

Fig. 1. Video conferencing with the participants meeting in a virtual room.

objects are inserted into the artificial scene. In order to place multiple participants into a common room, viewpoint modification is necessary which requires information about 3-D structure of the person. This geometry information can be estimated from multiple frames or a-priori knowledge is utilized by means of a rough generic head model, as it is done in this work.

A common method for representing head-and-shoulder video sequences three-dimensionally is model-based coding. For this technique, computer models of all objects and people in the scene are created. The models are then animated by motion and deformation parameters. Since temporal changes are described by a small parameter set, very efficient coding and transmission can be achieved. Head-and-shoulder scenes typical for video conferencing applications can for example be streamed at only a few kbit/s (28).

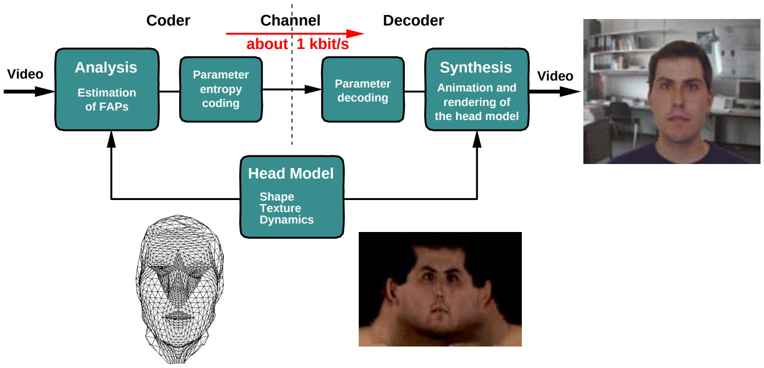

Fig. 2. Structure of a model-based codec.

For head-and-shoulder video sequences, a textured 3-D head model describes the appearance of the individual. Facial motion and expressions are modeled by the superposition of elementary action units each controlled by a corresponding parameter. In MPEG-4, there are 66 different facial animation parameters (FAP's) (29) that specify the temporal changes of facial mimics. These FAP's are estimated from the video sequence, encoded and transmitted over a network. At the decoder, the parameters are used to animate the 3-D head model and synthesize the video sequence by rendering the deformed computer model. Fig. 2 shows the structure of a model-based codec (28).

The initialization of a model-based codec usually starts with the fitting of the 3-D head model to the first frame of the video sequence and the extraction of a texture map from the image. Often, a facial mask is adapted to the video content (30; 31; 32; 33). This mask represents the facial area without hairs, ears, neck, or the back of the head and models local deformations caused by facial expressions. Areas outside the mask cannot be synthesized which might lead to artificially looking images, especially at the silhouette of the person. In our previous work (28), we have therefore used a complete 3-D head model and represented the fine structures of hair with billboarding techniques. These partly transparent planes, however, are only visually correct from near frontal views. Although more than the frontal facial area is modeled, the maximum range of head rotation is limited.

This limitation restricts applications, where full control of the head motion is desired. For the virtual conferencing scenario with participants meeting in a virtual room as shown in Fig. 1, the viewing direction must be changed afterwards in order to place the people correctly at a shared table. Moreover, head rotations can additionally be emphasized on small displays to allow to follow communication of distant partners. Even head pose changes based on audio signals for visualizing speaker attention is possible. With all these modifications, new facial areas become visible that are not captured with the current frame.

In order to render the people in the virtual room correctly, texture map updates have to be performed during the video sequence. The stitching of different image parts, however, requires an accurate head geometry. The more the head is turned from its original position, the more accurate the geometry has to be. Especially at the silhouette, errors are early visible and hairs with their sophisticated structure make an accurate modeling even more difficult.

These problems can be avoided with the combined image- and geometry-based tracking and animation technique described in Section 3. Fine structures like hair or uncovered areas are simply interpolated from previously recorded frames. In principle, all variations can be described by image-based rendering. Without loss of generality, we restrict the image-based representation in this scenario to head turning, since viewpoint modifications in the virtual room are mainly conducted around a vertical rotation axis. Also head turning shows most occlusions and uncovered areas in comparison to nodding or head rolling. This simplifies the stored image data to a one-dimensional array of images which can be easily handled by today's computers.

In the next section, the model-based tracking is described, which is used for initialization of this image-cube. Is also serves as basis for the combined approach and needs only slight modifications in order to incorporate the image-based techniques into the facial analysis component.

2 3-D Model-based Facial Expression Analysis and Synthesis

In this section, we will briefly describe our original 3-D model-based coding system (28; 34). Although purely geometry-based, it is used in this work to represent global head motion (except for head turns) and jaw movements, which severely effects the silhouette if the person is viewed from a sideways directions. This technique uses a 3-D head model with a single texture map extracted from the first frame of the video sequence. All temporal changes are modeled by motion and deformations of the underlying triangle mesh of the head model according to the given facial animation parameters (29). These parameters are estimated from the camera image using a gradient-based approach embedded into a hierarchical analysis-by-synthesis framework.

2.1 Gradient-Based Facial Mimic Analysis

The estimation of facial animation parameters for global and local head motion makes use of an explicit, parameterized head model describing shape, color, and motion constraints of an individual person. This model information is jointly exploited with spatial and temporal intensity gradients of the images. Thus, the entire area of the image showing the person of interest is used instead of dealing with discrete feature points, resulting in a robust and highly accurate system.

The image information is added with the optical flow constraint equation

$$\frac{\partial I}{\partial X} d_x + \frac{\partial I}{\partial Y} d_y = I - I', \tag{1}$$

where $\frac{\partial I}{\partial X}$ and $\frac{\partial I}{\partial Y}$ are the spatial derivatives of the image intensity at pixel position $[X\ Y]$. $I' - I$ denotes the temporal change of the intensity between two time instants $\Delta t = t' - t$ corresponding to two successive frames in an image sequence. This equation, obtained by Taylor series expansion up to first order of the image intensity, can be set up anywhere in the image. It relates the unknown 2-D motion displacement $\mathbf{d} = [d_x,\ d_y]$ with the spatial and temporal derivatives of the images.

The solution of this problem is under-determined since each equation has two new unknowns for the displacement coordinates. For the determination of the optical flow or motion field, additional constraints are required. Instead of using heuristical smoothness constraints, explicit knowledge from the head model about the shape and motion characteristics is exploited. A 2-D motion model can be used as an additional motion constraint in order to reduce the number of unknowns to the number of motion parameters of the corresponding model. The projection from 3-D to 2-D space is determined by camera calibration (35). Considering in a first step only global head motion, both d_x and d_y are functions of 6 degrees of freedom

$$\mathbf{d} = \mathbf{f}(R_x, R_y, R_z, t_x, t_y, t_z). \tag{2}$$

If local head motion like jaw movements or facial expressions are also modeled the displacement vector $\mathbf{d}$ becomes also a function of N additional facial animation parameters

$$\mathbf{d} = \mathbf{f}(R_x, R_y, R_z, t_x, t_y, t_z, FAP_0, \ldots, FAP_{N-1}). \tag{3}$$

Combining this motion constraint with the optical flow constraint (1) leads to a linear systems of equations for the unknown FAP's. Solving this linear system in a least squares sense, results in a set of facial animation parameters that determines the current facial expression of the person in the image sequence.

2.2 Hierarchical Framework

Since the optical flow constraint (1) is derived assuming the image intensity to be linear, it is only valid for small motion displacements between two successive frames. To overcome this limitation, a hierarchical framework can be used (28). First, a rough estimate of the facial motion and deformation parameters is determined from sub-sampled and low-pass filtered images, where the linear intensity assumption is valid over a wider range. The 3-D model is motion compensated and the remaining motion parameter errors are reduced on frames having higher resolutions.

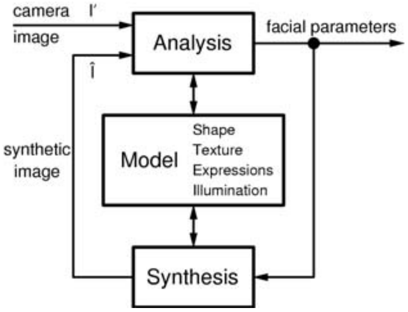

Fig. 3. Analysis-synthesis loop of the model-based estimator.

The hierarchical estimation can be embedded into an analysis-synthesis loop as shown in Fig. 3. In the analysis part, the algorithm estimates the parameter changes between the previous synthetic frame $\hat{I}$ and the current frame I' from the video sequence. The synthetic frame $\hat{I}$ is obtained by rendering the 3-D model (synthesis part) with the previously determined parameters. This approximate solution is used to compensate for the differences between the two frames by rendering the deformed 3-D model at the new position. The synthetic frame now approximates the camera frame much better. The remaining linearization errors are reduced by iterating through different levels of resolution. By estimating the parameter changes with a synthetic frame that corresponds to the 3-D model, an error accumulation over time is avoided.

2.3 Experimental Results

In this section some results for the model-based facial expression analysis are presented. A generic head model is adapted to the first frame of a CIF video sequence by varying shape parameters. A texture map is also extracted from this image. For each new frame, a set of 19 facial animation parameters and 4 motion parameters for the body are estimated using the proposed method. These parameters are transmitted and deform a generic head model in order to model the facial motion. The upper left of Fig. 4 shows an original frame of this sequence; on the right hand side the corresponding synthesized view from the head model is depicted. The lower left image illustrates the triangle mesh representing geometry of this model. As long as the viewing direction is similar to the original camera orientation, synthesized images match the original ones quite accurately. However, if the head model is rotated afterwards in order to simulate viewpoint modifications the silhouette of the model show distortions due to the planar approximation of hair by billboards. This is depicted in the lower right of Fig. 4, where the head is rotated by 20 degrees compared to the correct orientation.

3 Image-based Tracking and Rendering

In this section, we describe an extension of the pure geometry-based estimation and rendering of Section 2. By adding image-based interpolation techniques, the maximum range of head rotation can be broadened while preserving the correct outline, even in presence of hair. In contrast to other image-based techniques in facial animation like active appearance models (18; 19) that describe local features like mouth or eyes by a set of images, we use the captured set of video frames to realistically render the non-deformable parts of the head outside the face. In order to keep the number of images used

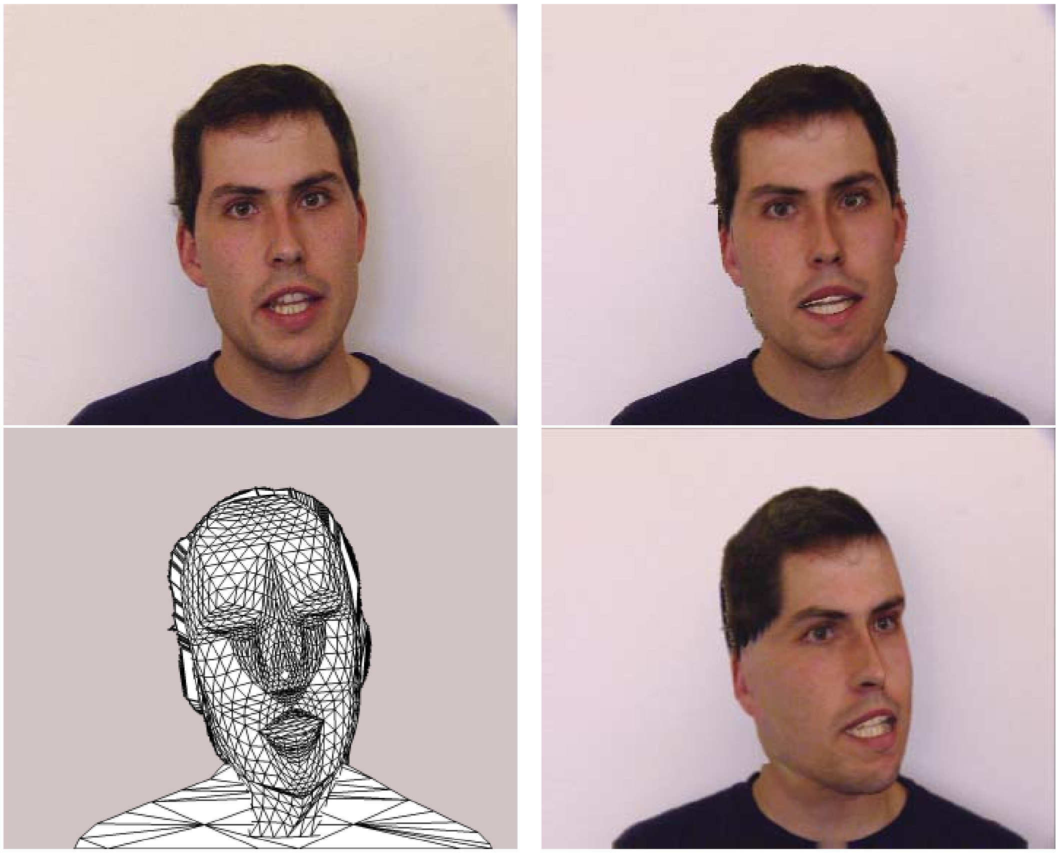

Fig. 4. **Upper Left:** One original frame of sequence *Peter*. **Upper Right:** Textured 3-D head model with FAP's extracted from the original frame. **Lower Left:** Wireframe representation. **Lower Right:** Synthesized frame with head rotated additional 20 degrees compared to the original, showing silhouette artifacts.

for image-based interpolation low, we only capture the one degree of freedom related to head turning. Other global head movements like pitch, roll or head translation, which usually show less variations, are modeled by geometry-based warping as described in Section 2. However, for other applications, more degrees of freedom can be interploated from images exactly in the same way. In that case, a multi-dimensional array of images has to be stored. The selection which frame to keep for interpolation can be derived directly from the tracker information and the desired requirements about sampling density.

3.1 Initialization of the Image Cube

For the initialization of the algorithm, the user has to turn the head to the left and the right as shown in Fig. 5. This way, we capture the appearance of the head from all sides for later interpolation. For simplification, we assume that a neutral expression is kept during this initialization phase; at least no expression altering the silhouette like opening of the jaw is permitted. The

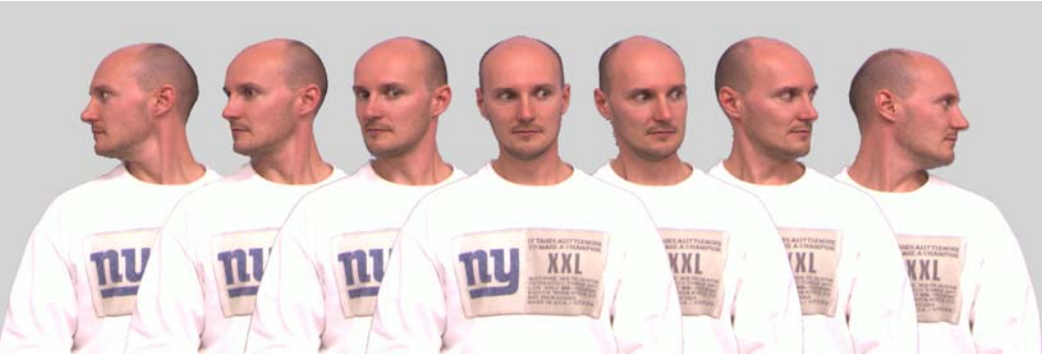

Fig. 5. Initial sequence with head rotation exploited for image-based rendering of new views.

person is then segmented from the background and all these images are collated in a 3-D image cube with two axes representing the X- and Y-coordinate of the images. The third axis of the image cube mainly represents the rotation angle R_y which need not be equidistantly sampled due to variations in the head motion.

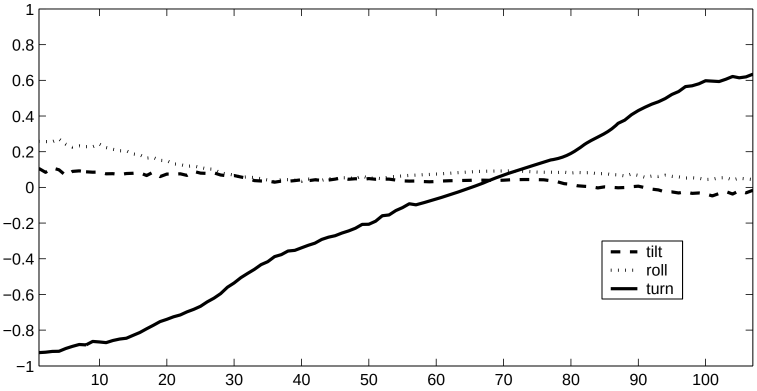

Fig. 6. Estimated rotation parameters for the head turn shown in Fig. 5.

For each of these frames, the rotation angle needs to be determined approximately using the a-priori knowledge of the end position of almost $\pm 90^0$. For that purpose, the global motion is estimated using the approach described in Section 2. The result is a parameter set for each frame specifying the six degrees of freedom with the main component being head rotation around the y-axis. Fig. 6 shows a result for 110 frames of a video sequence, where the user turns the head around the vertical axis. A parameter of one corresponds to a rotation of 90 degrees. From this estimate, it can easily be seen that head turning is the dominant motion but that other movements are additionally present and need to be considered for the final pose synthesis.

With this parameter set, the position and orientation of the triangle mesh in each frame is also known. For the shape adaptation, only the facial area responsible for modeling facial expressions needs to be quite accurate. The outline at the top and back (which shows up if the head is turned) of the head can be of approximate nature since image content recovers the details. It must only be assured, that the 3-D model covers the entire segmented person. Alpha blending is used to show a detailed outline even with a rough geometry model. The wireframe for this model is illustrated in Fig. 7.

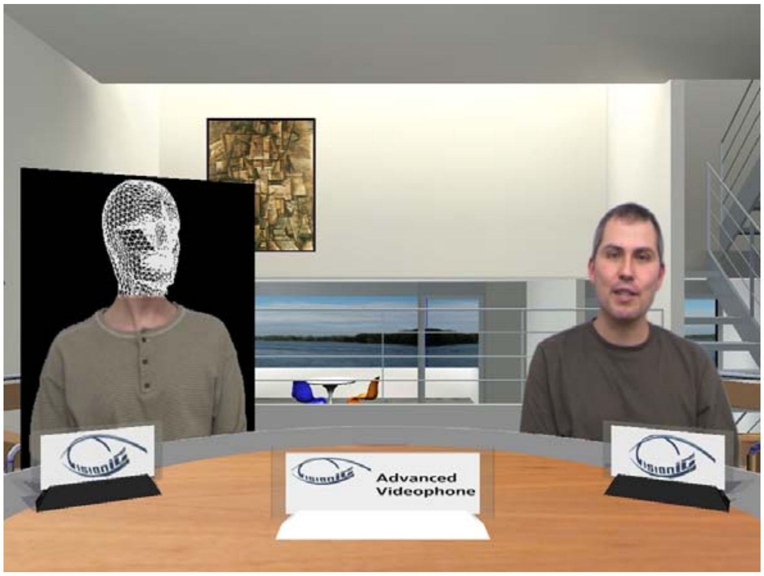

Fig. 7. Wireframe used for geometry warping of the head.

3.2 Rendering of New Frames

The rendering of new frames is performed by image-based interpolation combined with geometry-based warping. Given a set of facial animation parameters, the frame of the image cube having the closest value of head rotation is selected as reference frame for warping. Thus, the dominant motion changes are already represented by a real image without any synthetic warping. Deviations of the deisred global motion parameters from the stored values of the initialization step are compensated using 3-D geometry. This combination of geometry warping with image-based interpolation allows a very flexible trade-off between accuracy and size of the image cube. In principle, a single texture is sufficient if the underlying generic head model is very precise. At the other end with a very large number of images, no geometry is required at all in order to interpolate natural pose modifications. In our case, we combine a limited number of frames (about 100) with a very rough geometry model used for geometry-based interpolation between those views. An analysis of the trade-off between number of images and depth accuracy can be found in (36).

Head translation and head roll can be addressed by pure 2-D motion, only head pitch needs some depth dependent warping. As long as the rotation angles are small which is true in most practical situations, the quality of the geometry can be rather poor. Also local deformations due to jaw movements are here represented by head model deformations as in the original model-based approach of Section 2. In order to combine both sources, alpha blending is used to smoothly blend between the warped image and the 3-D model.

3.3 Representation of Eye and Mouth Movements

Realistic rendering of moving eyes and mouth is difficult to achieve. In this paper, we therefore use the original image data from the camera to achieve realistic animation of face features. The area around the eyes and the mouth is cut out from the camera frames, warped to the correct position of the person in the virtual scene using the 3-D head model, and smoothly merged into the synthetic representation using alpha mapping. This process requires knowledge of the exact position of eyes and mouth in the original video to prevent jitter of facial features. We use the model-based motion estimation scheme described in Section 2 in order to accurately track the facial features over time. For the tracking, realistic hair is not required and the restricted motion of a person looking into a camera reduces the demands on a highly accurate 3-D model for that purpose. Once the features are localized, the corresponding image parts are cut out and combined with the previous steps.

Thus three different techniques are used for different facial areas. The texture of the main head parts except for eye and mouth regions are taken from the image cube representing the person for all possible head turns. 3-D model-based warping is then applied to model the other 5 global head movements (R_x, R_z, t_x, t_y, t_z) as well as the opening of the jaw. Finally local eye and mouth motion is represented by image information captured at the current time instant by a video camera. This way, natural looking images can be synthesized showing facial expressions and a correct silhouette even for large modifications of the head rotation angles.

3.4 Image-based Motion Estimation

Since two different techniques – image- and geometry-based interpolation – are used to render novel views, the estimation of facial animation parameters (head tracking) from camera images must be slightly modified in order to avoid inconsistent values for the two approaches and to obtain a smooth blending of all three sources. The optical-flow constraint equation is therefore replaced by

$$\frac{\partial I}{\partial X} d_x + \frac{\partial I}{\partial Y} d_y + \frac{\partial I_{ibr}}{\partial R_y} \Delta R_y = I - I', \tag{4}$$

with the additional dependence from $\frac{\partial I_{ibr}}{\partial R_y}$. Instead of describing temporal image changes purely by warping with displacements $\mathbf{d}$, head rotation around the y-axis is modeled by moving the reference frame in the image cube. Intensity changes between neighboring images in the image cube are given by $\frac{\partial I_{ibr}}{\partial R_y}$. The dependence from R_y is taken from the estimates of the initialization phase. In contrast to (2), the displacement vector is now only a function of 5 unknowns for global head motion

$$\mathbf{d} = \mathbf{f}(R_x, R_z, t_x, t_y, t_z, FAP_0, \ldots, FAP_{N-1}) \tag{5}$$

with head rotation R_y being excluded. With the additional term in the optical flow constraint (4) all parameters can be estimated in the same way as described in Section 2.1. In the hierarchical framework, also the image cube must be downsampled in all three directions. All other components remain the same and allow the estimation of all FAP's consistently with the initially captured frames of the image cube.

3.5 Experimental Results

In this section, we show some results obtained with the proposed tracking and rendering technique. A video sequence is recorded showing the head and upper body of a person. In the beginning, the person rotates the face to the left and right as shown in Fig. 5 and then starts talking. From the initial part with about 110 frames, the video cube is created from the segmented images and the global head motion is estimated for each of these frames.

For the rendering of new views in a virtual conferencing scenario, the current position and orientation of a person's head as well as jaw movements are tracked with the method described in Section 3.4. The pose of the person can simply be altered by changing the rigid body motion parameters obtained from the real data. The resulting head turn angle R_y determines which frame to use from the image cube for texture mapping. The remaining motion parameters are used for geometry-based warping using the selected texture map. The resulting image shows the person from a different direction and head orientation compared to the original camera image. This is illustrated in Fig. 8, where different views are rendered from a single camera image by changing the estimated global motion parameters. Please note that also occluded areas like the ears are correctly reproduced due to the usage of an image cube with previous frames. Also people with a lot of hair can be rendered naturally as shown in Fig. 9.

As described in Section 3.3, local facial expressions are modeled by clipping the corresponding region from the camera frames and, after warping, pasting

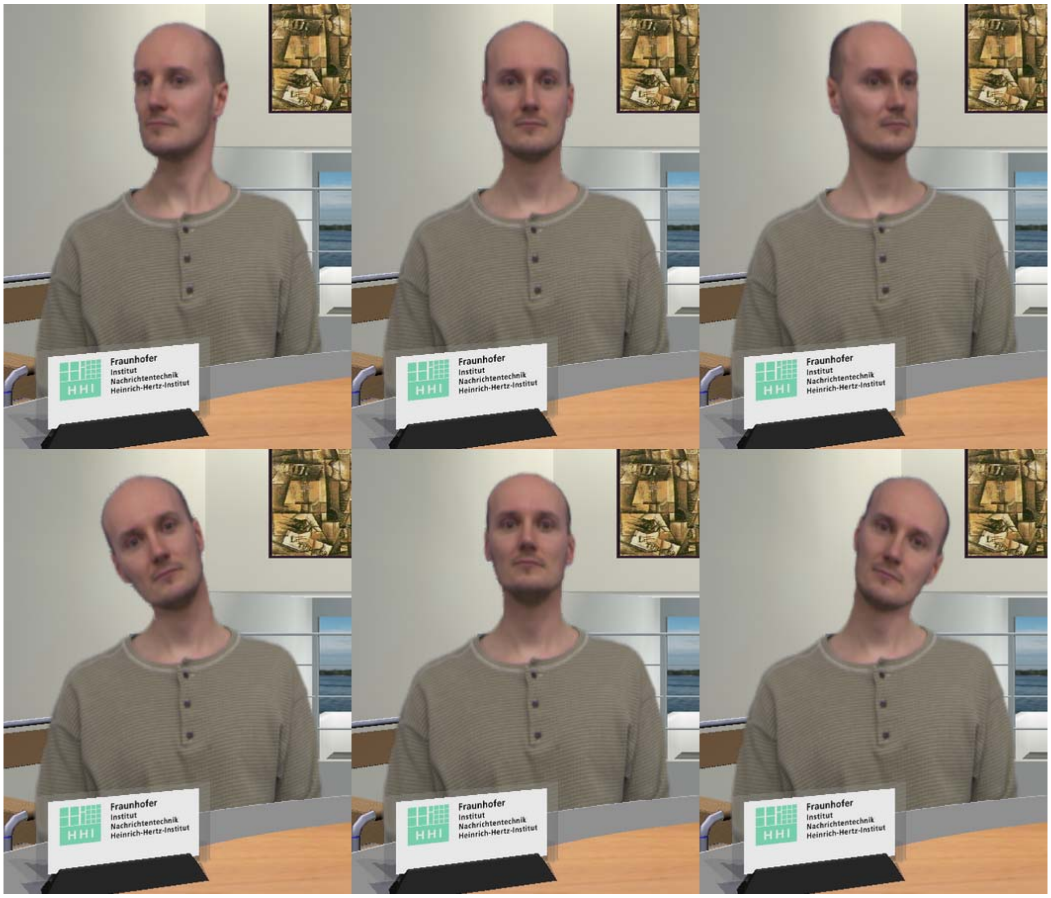

Fig. 8. Different head positions created from a single camera frame using the proposed method. The viewing direction in the virtual scene is not identical to the original camera position.

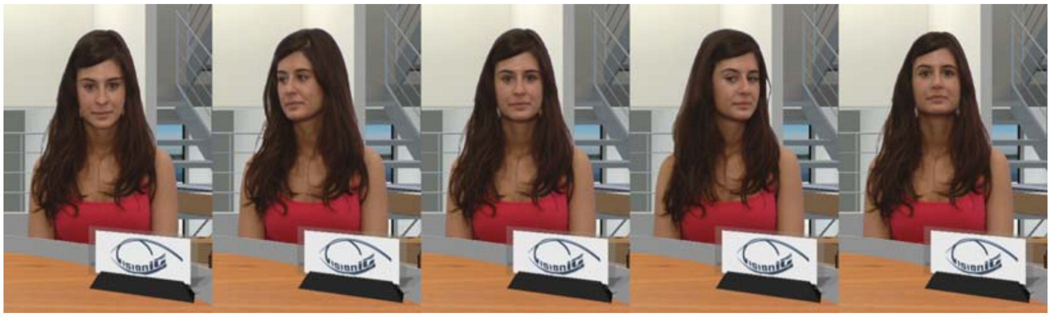

Fig. 9. Different head positions. Hair is correctly reproduced if the head is turned.

them into the synthetic scene representation. The usage of the entire image for model-based face tracking assures an accurate extraction of these features. Fig. 10 shows different frames rendered at different time instants. During rendering, the head pose can be modified even if the camera captures the user from the same viewing direction. This is illustrated in Fig. 11, where the upper row shows the frames of the camera whereas the lower row depicts a part of the synthesized frame from the virtual environment. In these frames, the

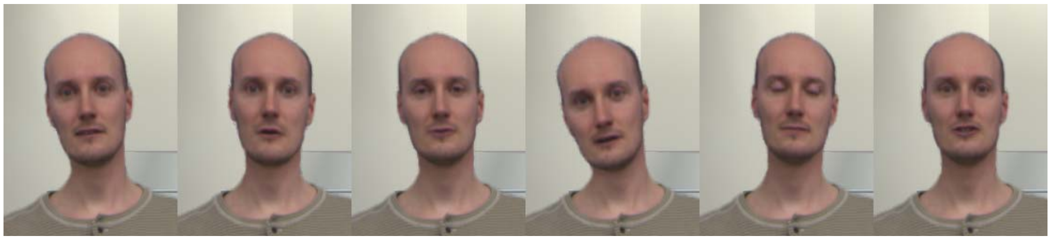

Fig. 10. Different facial expressions synthesized with the proposed method which combines image-based techniques with 3-D model-based components.

viewing direction of the novel views does not coincide with the one of the real camera. No artifacts caused by the different rendering techniques are visible due to smooth alpha blending.

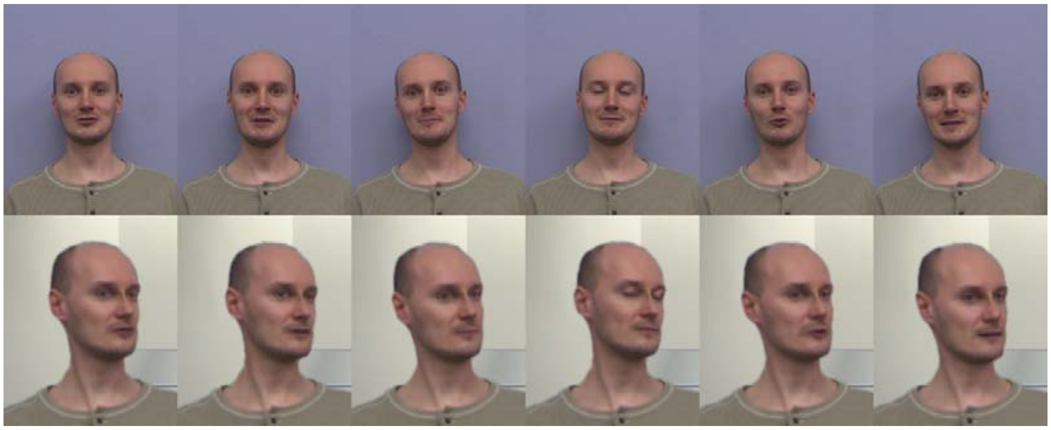

Fig. 11. Upper row: camera frames. Lower row: synthesized frames with different head poses.

Since a wide range of different head poses is covered in the image cube, large changes compared to the real orientation can be applied later on for the rendering of new views in the virtual scene. This enables many enhancements compared to conventional systems. For small displays, e.g., head motion is very small if a user looks at different people in the room. In order to show the other participants the current focus of a local user these head motions can be enhanced and adjusted to the chairs' positions at the virtual table. If one user is connected with a conventional terminal without tracking capabilities, also synthetic head motion can be added to show a visually more pleasing result. For that purpose, we added a speech detector, which selects the person currently speaking. The head of the user with no tracking capabilities is then turned to this person. Fig. 12 shows two views from such a system. In the upper image, both people show a neutral position. As soon as the person on the right side starts speaking, the head of the left person is turned towards him. For a video conferencing application, the entire rendering of the 3D scene with image-based warping of the video textures runs in real-time at 25 frames per second on a standard PC.

4 Conclusions

In this paper, we have presented a method for the analysis and synthesis of head-and-shoulder scenes in the context of virtual video conferencing. We have extended a 3-D model-based coding approach with image-based rendering techniques, in order to obtain naturally looking images even for large modifications of the viewing direction. In order to reduce the demands on memory and capturing, only one degree of freedom related to head rotation around the vertical axis is described by image-based warping. Other global motions are modeled with a generic 3-D head model. Local facial expressions are added using clip-and-paste techniques. Although in the experiments only head rotation is interpolated by previously stored images, the approach is fully scalable and can be extended to describe more facial motion parameters by natural image information. The image-based component is embedded into a gradient-based estimation technique that uses the entire image information in a hierarchical framework for accurate facial motion analysis.

References

[1] M. Levoy, P. Hanrahan, Light field rendering, in: Proc. Computer Graphics (SIGGRAPH), New Orleans, USA, 1996, pp. 31–42.

[2] S. Peleg, M. Ben-Ezra, Stereo panorama with a single camera, in: Proc. Computer Vision and Pattern Recognition, Ft. Collins, USA, 1999, pp. 395–401.

[3] H.-Y. Shum, L.-W. He, Rendering with concentric mosaics, in: Proc. Computer Graphics (SIGGRAPH), Los Angeles, USA, 1999, pp. 299–306.

[4] H.-Y. Shum, L.-W. He, A review of image-based rendering techniques, in: Proc. Visual Computation and Image Processing (VCIP), Perth, Australia, 2000, pp. 2–13.

[5] S. J. Gortler, R. Grzeszczuk, R. Szeliski, M. F. Cohen, The Lumigraph, in: Proc. Computer Graphics (SIGGRAPH), New Orleans, USA, 1996, pp. 43–54.

[6] D. N. Wood, D. I. Azuma, K. Aldinger, B. Curless, T. Duchamp, D. H. Salesin, W. Stuetzle, Surface light fields for 3D photography, in: Proc. Computer Graphics (SIGGRAPH), New Orleans, USA, 2000, pp. 287–296.

[7] E. H. Adelson, J. R. Bergen, The plenoptic function and the elements of early vision, in: M. Landy, J. A. Movshon (Eds.), Computational Models of Visual Processing, The MIT Press, 1991.

[8] M. Rydfalk, Candide: A parameterized face, Ph.D. thesis, Linköping University, liTH-ISY-I-0866 (1978).

[9] F. I. Parke, Parameterized models for facial animation, IEEE Computer Graphics and Applications 2 (9) (1982) 61–68.

[10] K. Waters, A muscle model for animating three-dimensional facial expressions, in: Proc. Computer Graphics (SIGGRAPH), Vol. 21, Anaheim, CA, USA, 1987, pp. 17–24.

[11] Y. Lee, D. Terzopoulos, K. Waters, Realistic modeling for facial animation, in: Proc. Computer Graphics (SIGGRAPH), Los Angeles, USA, 1995, pp. 55–61.

[12] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D. H. Salesin, Synthesizing realistic facial expressions from photographs, in: Proc. Computer Graphics (SIGGRAPH), Orlando, Florida, 1998, pp. 75–84.

[13] R. Forchheimer, O. Fahlander, T. Kronander, Low bit-rate coding through animation, in: Proc. Picture Coding Symposium (PCS), Davis, California, 1983, pp. 113–114.

[14] W. J. Welsh, S. Searsby, J. B. Waite, Model-based image coding, British Telecom Technology Journal 8 (3) (1990) 94–106.

[15] D. E. Pearson, Developments in model-based video coding, Proceedings of the IEEE 83 (6) (1995) 892–906.

[16] K. Aizawa, H. Harashima, T. Saito, Model-based analysis synthesis image coding (MBASIC) system for a person's face, Signal Processing: Image Communication 1 (2) (1989) 139–152.

[17] S. Chao, J. Robinson, Model-based analysis/synthesis image coding with eye and mouth patch codebooks, in: Proceedings of Vision Interface, Banff, Alberta, 1994, pp. 104–109.

[18] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, in: Proc. European Conference on Computer Vision (ECCV), Freiburg, Germany, 1998.

[19] R. Gross, I. Matthews, S. Baker, Constructing and fitting active appearance models with occlusions, in: Proc. IEEE Workshop on Face Processing in Video, Washington, USA, 2004.

[20] B. J. Theobald, G. C. Cawley, I. A. Matthews, J. A. Bangham, Near-videorealistic synthetic visual speech using non-rigid appearance models, in: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 5, Hongkong, 2003, pp. 800–803.

[21] M. Odisio, G. Bailly, Shape and appearance models of talking faces for model-based tracking, in: Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures, Nice, France, 2003, pp. 143–148.

[22] I. A. Ypsilos, A. Hilton, A. Turkmani, P. Jackson, Speech-driven face synthesis from 3d video, in: Proc. IEEE Symposium on 3D Data Processing, Visualisation and Transmission, 2004.

[23] Y. Chang, T. Ezzat, Transferable videorealistic speech animation, in: Proc. ACM Eurographics, Los Angeles, USA, 2005, pp. 29–31.

[24] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: Proc. Computer Graphics (SIGGRAPH), Los Angeles, CA, USA, 1999, pp. 187–194.

[25] P. Eisert, J. Rurainsky, Image-based rendering and tracking of faces, in: Proc. International Conference on Image Processing (ICIP), Vol. I, Genova, Italy, 2005, pp. 1037–1040.

[26] J. Rurainsky, P. Eisert, Text2video: Text-driven facial animation using MPEG-4, in: Proc. Visual Computation and Image Processing (VCIP), Beijing, China, 2005.

[27] O. Schreer, R. Tanger, P. Eisert, P. Kauff, B. Kaspar, R. Englert, Real-time avatar animation steered by live body motion, in: Proc. 13th International Conference on Image Analysis and Processing, Cagliari, Italy, 2005.

[28] P. Eisert, B. Girod, Analyzing facial expressions for virtual conferencing, IEEE Computer Graphics and Applications 18 (5) (1998) 70–78.

[29] ISO/IEC FDIS 14496-2, Generic Coding of audio-visual objects: (MPEG-4 video), Final Draft International Standard, Document N2502 (1999).

[30] 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 Communication 9 (3) (1997) 201–220.

[31] J. Ahlberg, Extraction and coding of face model parameters, Ph.D. thesis, University of Linköping, Sweden, lIU-TEK-LIC-1999-05 (1999).

[32] D. DeCarlo, D. Metaxas, Deformable model-based shape and motion analysis from images using motion residual error, in: Proc. International Conference on Computer Vision (ICCV), Bombay, India, 1998, pp. 113–119.

[33] M. Hess, G. Martinez, Automatic adaption of a human face model for model-based coding, in: Proc. Picture Coding Symposium (PCS), San Francisco, USA, 2004.

[34] P. Eisert, MPEG-4 facial animation in video analysis and synthesis, International Journal of Imaging Systems and Technology 13 (5) (2003) 245–256, invited paper.

[35] P. Eisert, Model-based camera calibration using analysis by synthesis techniques, in: Proc. International Workshop on Vision, Modeling, and Visualization, Erlangen, Germany, 2002, pp. 307–314.

[36] J.-X. Chai, X. Tong, S.-C. Chan, H.-Y. Shum, Plenoptic sampling, in: Proc. Computer Graphics (SIGGRAPH), Los Angeles, USA, 2000, pp. 307–318.

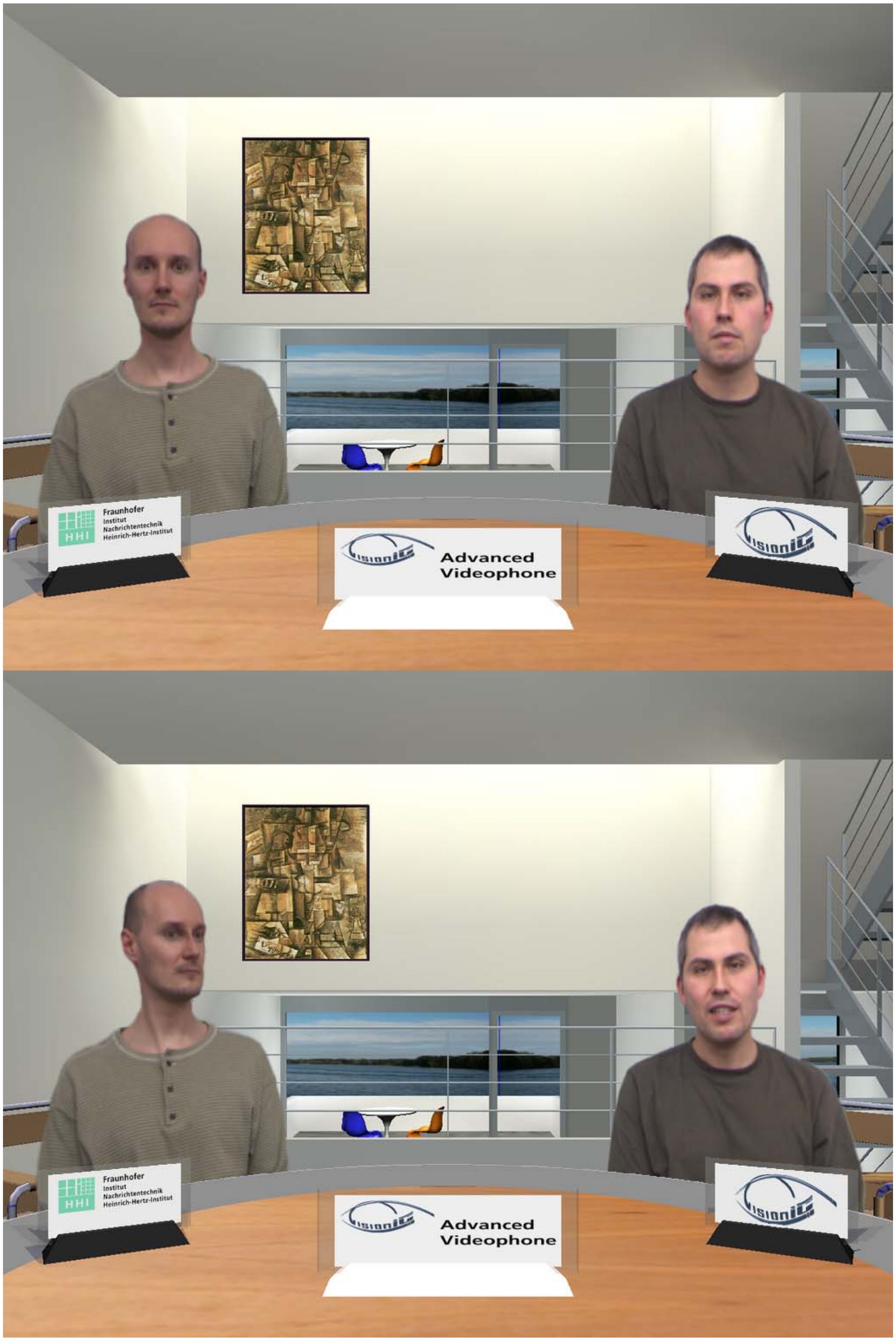

Fig. 12. Virtual conferencing with speech-assisted scene manipulation. As soon as one person starts speaking the other turn his head towards the speaking person.