RENDERING AND ANALYSIS OF FACES USING MULTIPLE IMAGES WITH 3D GEOMETRY

Peter Eisert and Jürgen Rurainsky

Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute
Image Processing Department
Einsteinufer 37, D-10587 Berlin, Germany
Email: eisert@hhi.fhg.de

ABSTRACT

In this paper, we present a method for the analysis and synthesis of faces that combines image-based rendering with geometry-based warping. In contrast to conventional model-based coding systems that usually rely on textured 3-D triangle mesh models for the representation of human heads, we additionally incorporate a large number of past images in order to interpolate new views. The use of real camera frames for facial synthesis leads to very natural looking results. However, the large number of variations in the face cannot all be covered by past frames. Therefore, we use image-based interpolation only to represent head turning, which usually shows most variations due to uncovered areas. The remaining motion is modeled by 3-D geometry-based warping. Similarly, the tracking of facial expression and motion has to be adapted to the combined model- and image-based approach. Experiments illustrate the applicability of the algorithms for the application of virtual conferencing.

1. INTRODUCTION

The use of 3-D models for the efficient representation of people in video conferencing applications has been investigated in model-based video coding [1, 2, 3]. For this technique, computer models of all objects and people in the scene are created, which 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 [4].

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) [5] that specify the

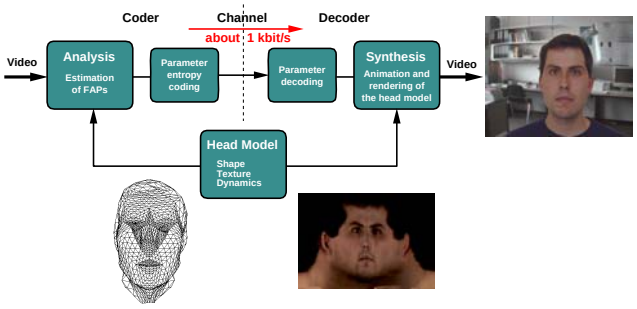

Fig. 1. Structure of a model-based codec.

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. 1 shows the structure of a model-based codec.

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. Often, a facial mask is adapted to the video content [6, 7, 8, 9]. 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 [4], 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 also 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. 2, the viewing direction must be changed af-

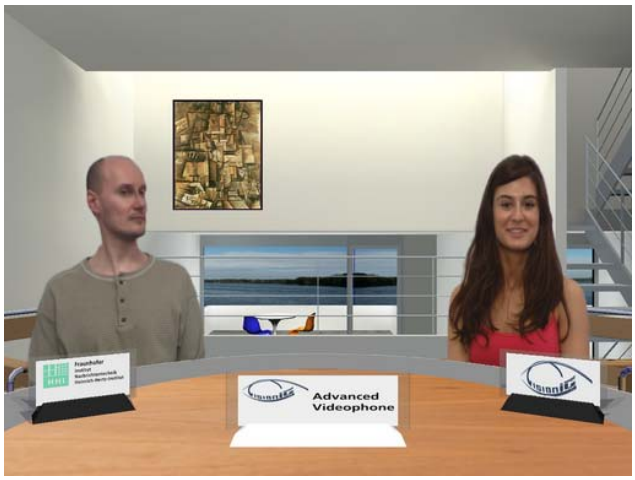

Fig. 2. Video conferencing with the participants meeting in a virtual room.

terwards 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. 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.

A technique which allows a realistic rendering of even complex surfaces is image-based rendering. Instead of using a highly accurate geometry, new views of an object are simply interpolated from a large number of images. If enough images are available, 3-D shape information is even unnecessary. On the other hand, with increasing number of degrees of freedom in object motion, the number of required images grows exponentially. For a human face with the enormous capabilities of shape variations, it is almost impossible to capture all possible combinations. 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 [10]. No geometry information requires many images whereas a highly accurate model is needed if only a few images are exploited.

In this paper, we combine image-based rendering with the use of a 3-D head model. 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. In contrast to [11], 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 post-processing step in applications like virtual conferencing. Since the new frame is interpolated from real images, the uncovered areas at the side of the head look correctly and also hair with their sophisticated structure are accurately reproduced.

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 [4, 12]. 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 [5]. These parameters are estimated from the camera image using a gradient-based approach embedded into a hierarchical analysis-by-synthesis framework.

For the image-based tracker and renderer described in Section 3, the basic system is extended to deal also with the initial image sequence representing head rotation. A modified gradient-based estimator is embedded into a similar architecture in order to combine image-based rendering with geometry modeling.

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 [13]. 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 a function of N facial animation parameters including those for global head motion [4]

$$\mathbf{d} = \mathbf{f}(FAP_0, \ldots, FAP_{N-1}). \tag{3}$$

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. 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. 3 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 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. 3, where the head is rotated by 20 degrees compared to the correct orientation.

Fig. 3. **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 artifacts at the silhouette.

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 [14, 15] 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 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.

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. 4. This

Fig. 4. Initial sequence with head rotation exploited for image-based rendering of new views.

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 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 has not to be equidistantly sampled due to variations in the head motion.

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. 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 need to be quite accurate. The outline at the top and back 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 mapping is used to show a detailed outline even with a rough geometry model.

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 other global motion parameters from the stored values of the initialization step are compensated using 3-D geometry. 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 convincing results also in these regions. 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 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-based 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

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

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

$$\mathbf{d} = \mathbf{f}(R_x, R_z, t_x, t_y, t_z) \tag{5}$$

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. 4 and then starts talking. From the initial part with about 120 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. 5, 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.

As described in Section 3.3, local facial expressions are modeled by clipping the corresponding region from the camera frames and, after warping, pasting 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. 6 shows different frames rendered at different time instants. Again, the 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. 5. 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.

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

Fig. 6. Different facial expressions synthesized with the proposed method which combines image-based techniques with 3-D model-based components.

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. The image-based component is embedded into a gradient-based estimation technique that uses the entire image information in a hierarchical framework.

Acknowledgements

The work presented in this paper has been developed with the support of the European Network of Excellence VISNET (IST Contract 506946).

5. REFERENCES

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

[2] 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.

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

[4] P. Eisert and B. Girod, "Analyzing facial expressions for virtual conferencing," *IEEE Computer Graphics and Applications*, vol. 18, no. 5, pp. 70–78, Sep. 1998.

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

[6] 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 Communication*, vol. 9, no. 3, pp. 201–220, Mar. 1997.

[7] J. Ahlberg, *Extraction and Coding of Face Model Parameters*, Ph.D. thesis, University of Linköping, Sweden, 1999, LIU-TEK-LIC-1999-05.

[8] D. DeCarlo and 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, Jan. 1998, pp. 113–119.

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

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

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

[12] P. Eisert, "MPEG-4 facial animation in video analysis and synthesis," *International Journal of Imaging Systems and Technology*, vol. 13, no. 5, pp. 245–256, Mar. 2003, invited paper.

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

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

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