# IMAGE-BASED RENDERING AND TRACKING OF FACES

*Peter Eisert and Jürgen Rurainsky*

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

## 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 at a virtual table. 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 that models temporally changing facial features. Both geometry and image cube information are jointly exploited in facial expression analysis and synthesis.

## 1. INTRODUCTION

Model-based video coding [1, 2, 3] is a technique which allows a very efficient encoding of particular video sequences. Head-and-shoulder scenes typical for video conferencing applications can for example be streamed at only a few kbit/s [4]. This low bitrate is achieved by representing the objects in the scene by 3-D computer graphics models which are transmitted once unless not already cached at the decoder. Temporal changes are described by a small set of parameters specifying object motion, deformation, or other scene changes. Fig. 1 shows the structure of a model-based codec.

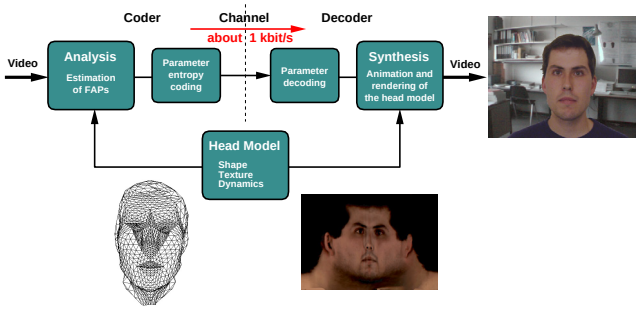


**Fig. 1**. Model-based codec.

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) [5] 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.

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 example in video conferencing with participants meeting in a virtual room as shown in Fig. 2, the viewing direction must be changed afterwards in order to place the people correctly at a shared table. Moreover, head rotations must be emphasized or altered 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.

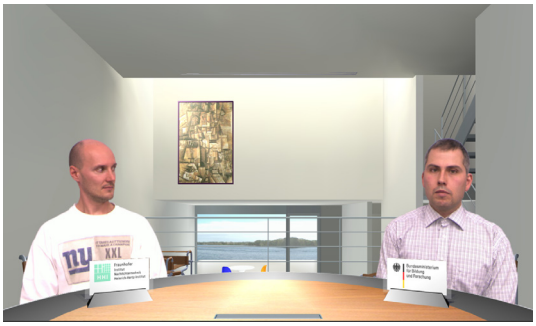

**Fig. 2**. Video conferencing with the participants meeting in a virtual room.

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* [10]. Instead of using a highly accurate geometry, new views of an object are simply in-

terpolated 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 the human face with the enormous capabilities of shape variations, it is almost impossible to capture all their 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 propose a method that combines image-based rendering (IBR) with the use of a 3-D head model. Only head turning with the most dominant image changes are interpolated from a set of initially captured views, whereas other global and local facial motion are represented with a geometry model. This severely restricts the number of required images but enables head rotation of the person as a postprocessing 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, 11]. Although purely geometry-based, it is used in this work to extract and represent local facial expressions. This technique uses a 3-D head model with a single texture map extracted from the first frame of the video sequence. All facial expressions are modeled by 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 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 [12]. 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 caused by facial expressions is 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. 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 [4]. 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.

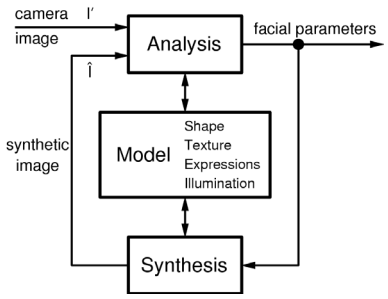


**Fig. 3**. Analysis-synthesis loop of the model-based estimator.

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 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.

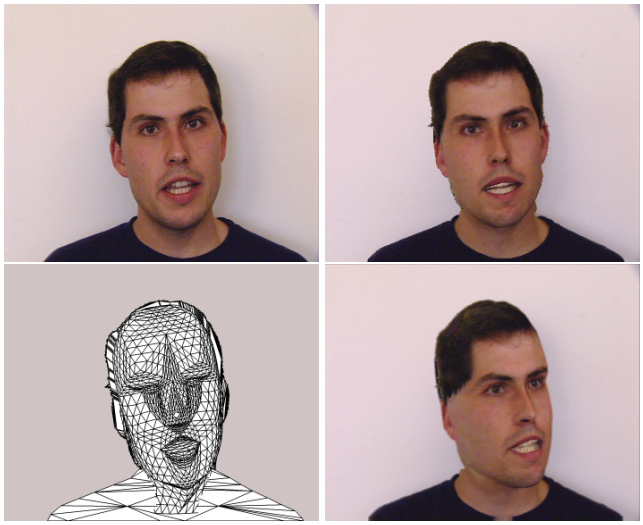

**Fig. 4**. **Upper Left:** One original frame of sequence *Peter*. **Upper Right:** Textured 3-D head model with FAP's extracted from the original frame. **Lower Left:** Wireframe representation. **Lower Right:** Synthesized frame with head rotated additional 20 degrees compared to the original, showing 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 [13, 14] 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. Facial expressions are in this paper represented with the generic head model of the model-based approach. 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 or roll, which usually show less variations, are also modeled by geometry-based warping in the same way as the local deformations.

### 3.1. Initialization of the Image Cube

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

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 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.

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 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. Local deformations due to facial expression 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. This way, natural looking images can be synthesized showing facial expressions and a correct silhouette even for large modifications of the

head rotation angles.

### 3.3. 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 from camera images must be slightly modified in order to avoid inconsistent values for the two approaches and to obtain a smooth blending. 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

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

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.

### 3.4. 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 and then starts talking. The left hand side of Fig. 6 shows an image of this initialization while the image in the middle depicts a frame later in the sequence showing facial motion. The local changes of the facial features are warped to the new desired head orientation using 3-D geometry information. It is then smoothly blended into the frame of the image cube corresponding to this orientation. The right hand side of Fig. 6 shows the result of the warping with additional blending. Both local mouth deformations and the head orientation are recovered quite accurately. This model can then be placed into a virtual conferencing system as shown in Fig. 2.

**Fig. 6**. **Left:** Initial frame from the image cube. **Middle:** One frame of the sequence with local facial action. **Right:** New frame with different head orientation obtained by geometry-based warping and image-based interpolation.

## 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 and local motion are modeled with a generic 3-D head model. The image-based component is embedded into a gradient-based estimation technique that uses the entire image information in a hierarchical framework.

## 5. ACKNOWLEDGMENTS

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

## 6. 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, "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.

[12] 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.

[13] 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.

[14] 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.