AUTOMATIC MODELLING OF NATURAL SCENES FOR GENERATING SYNTHETIC MOVIES

Reinhard Koch

Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover
Hannover, Federal Republic of Germany

A model based analysis by synthesis algorithm is presented. It combines image analysis and synthesis techniques to obtain a three dimensional scene description and to generate synthetic movies out of that scene description. A 3D model world is generated automatically out of TV image sequences containing 3D objects with naturally textured surfaces. The objects are projected into the 2D image domain and compared with subsequent images of the sequence to be analyzed. Differences between input and synthesized images serve to extract shape, motion and surface texture parameters and to adapt the 3D model scene. The model data base can be manipulated to generate synthetic movies with highly realistic, natural looking images.

1. INTRODUCTION

The classical techniques of computer graphics generate artificial 3D objects using CAD techniques and synthesize 2D images based on synthetic textures using solid model rendering. See Foley, van Dam [1] on the basic principles.

The techniques of image analysis try to extract relevant features out of natural images to gain information about the real 3D scene. The main problem here is the ambiguity in estimating the depth from a series of 2D images. This problem has been investigated by many researchers in the past. Special attention has been given to extracting 3D object motion out of a sequence of images. A compendium on estimating the depth of a scene from image sequences can be found in Jarvis [2]. A review on motion estimation is presented by Aggarval and Nandhakumar [3].

Computer graphics and image analysis in this context represent inverse problems, giving transformations from the 3D scene space into the 2D image space and vice versa. Fig. 1. reveals the relationship between both fields.

Even though the problems are related, there is a gap between the world of computer graphics and the world of image analysis.The computer graphics rendering process uses artificial object texture for image synthesis which can be detected as such by a human observer in most cases. Image analysis on the other hand relies on natural images as input to create a 3D model scene that should as closely as possible represent the real objects in the scene. If the extracted model scene is a good approximation of the real scene, it should be possible to synthesize a natural looking image out of the model scene. Differences between natural and synthesized images are interpreted as differences in the scene and from that the model description can be updated.

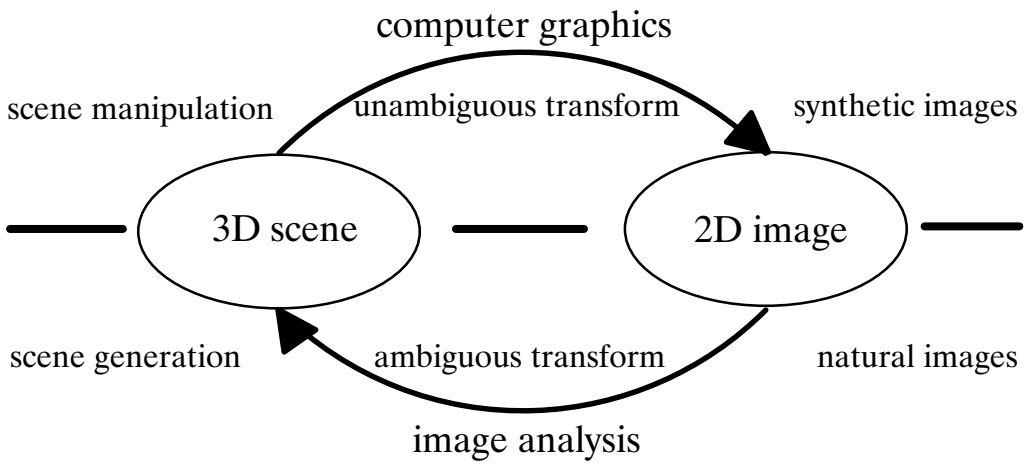


Fig. 1. The relationship between computer grapics and image analysis.

This feature opens a way to develop a model based analysis by synthesis algorithm described here. It integrates both analysis and synthesis and is therefore applicable to a variety of applications ranging from solid modelling over the automatic extraction of object shape and motion out of TV image sequences to the generation of movies containing natural views of outdoor scenes. Emphasis is placed on automatic object acquisition and scene adaptation to allow efficient creation of the model scene.

In the following, the 3D model world and some restrictions for the analysis part are defined. It is then shown how the model world is initialized and some techniques to generate model objects are explained. The ability of the algorithm to track moving objects and to adapt the 3D object shape is discussed. Finally the capability of the algorithm to generate realistic image sequences using naturally textured objects is demonstrated.

2. DEFINITION OF THE 3D MODEL WORLD

In order to generate the model world its properties must be defined. The model world is a representation of the real world. It consists of a model camera, a model illumination and a model scene description. The camera has pinhole geometry using perspective transform, no lens distortions are considered. For the model illumination ambient lighting with constant intensity is assumed. This restriction is neccessary for the analysis part only. Once the objects have been generated, any illumination model for the image synthesis may be used.

The scene consists of 3D objects that move in space. Each object is defined by a set of three parameters: position, shape, and surface color. The objects consist of flexibly connected, deformable components. Each component has a 3D shape built out of control points spanning a triangular mesh on the component surface. Through shifting individual control points the component shape can be deformed. All components of an object are joined together forming a single surface. The interconnection between the components has a flexible surface. The faces of the triangles contain the natural surface texture. Therefore a projection of the object surface onto a model image looks exactly like the real object. Fig. 2. illustrates the model world description.

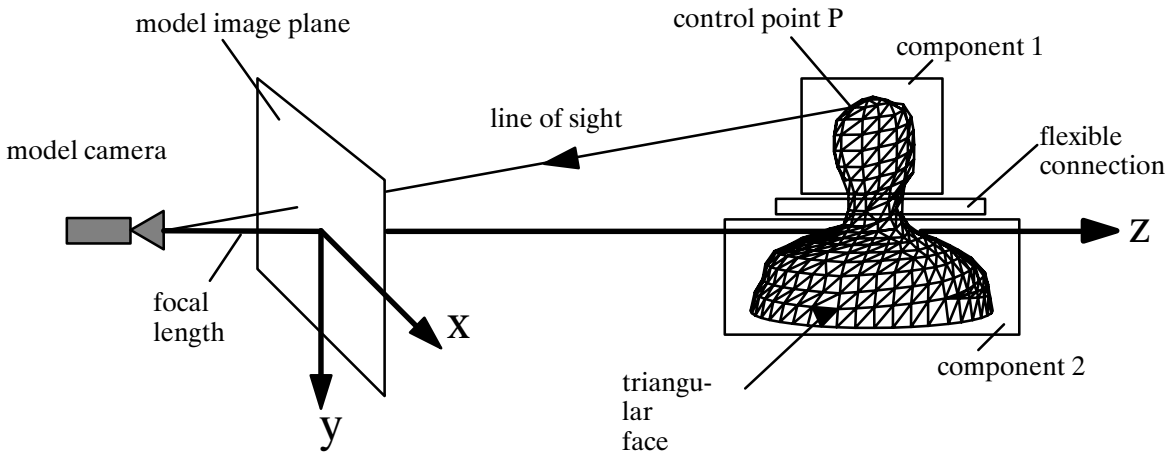


Fig. 2. The 3D model world with one object consisting of two components.

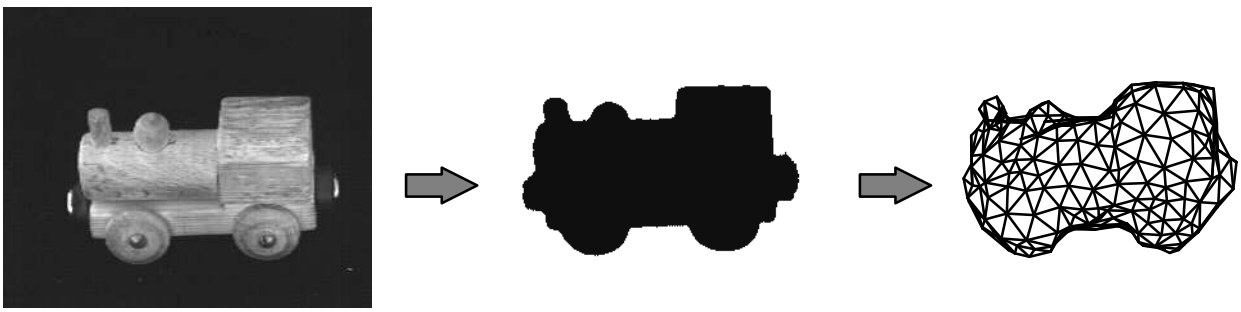
3. AUTOMATIC OBJECT GENERATION

The first step is to acquire the different model objects. The initial objects may be generated using several methods. Depending on the available image material and scene conditions the appropriate technique can be chosen. Two types of imaging situations are considered in the following chapter.

3.1. Monocular camera moves and takes views of a resting object from known positions

Busch [4] describes an algorithm to automatically extract 3D object descriptions out of multiple views without human interaction, using a controlled laboratory environment. Some restrictions are imposed on the objects: they should be rigid and opaque. An automatic camera calibration mechanism serves to determine the camera positions. With this algorithm it is possible to generate even complex objects in two steps.

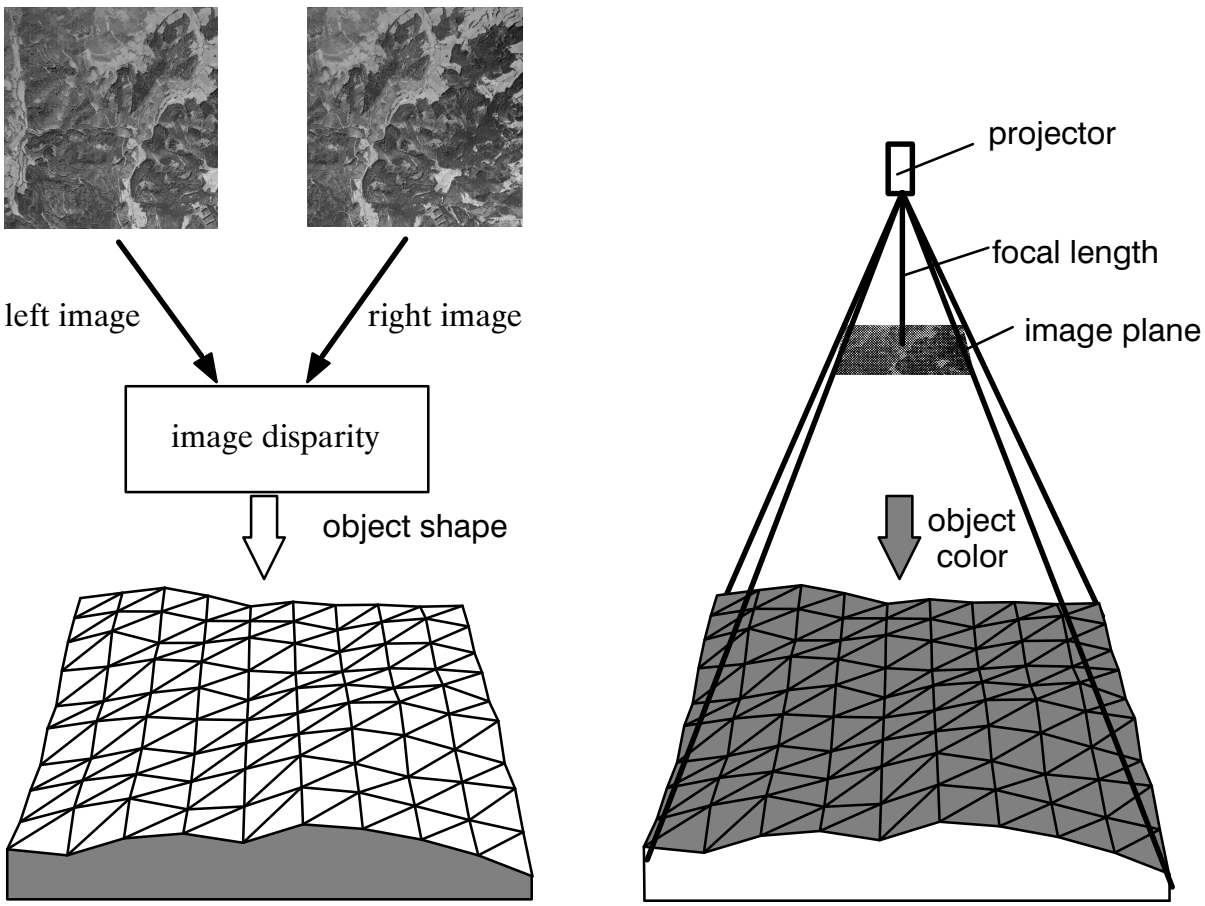
First a volumetric representation is obtained using 3D silhouette clipping. A bounding volume is defined and the object is assumed to be inside this volume. The object is viewed from different sides, each view creating one object silhouette. All the volume elements (voxels) outside the silhouettes are cleared and only the voxels inside the object remain. This yields a voxel representation of the object. This voxel representation must be converted into a surface representation for further processing. Therefore the surface of the voxel object is scanned and control points are selected that span a triangular surface mesh. Figure 3. demonstrates the object generation process for a wooden toy locomotive.



3a. Image input frames 3b. Object silhouette 3c. 3D model object

Fig. 3. Object generation using multiple views [4].

A special case of object generation using multiple views is the creation of range maps from stereoscopic aerial photography. Two overlapping images of outdoor terrain are taken from an airplane and a 3D terrain object is built using a depth map obtained from image disparity measure. The depth map is obtained using standard photogrammetric methods. The implementation used here is an automatic image correlator developed by Sasse and Altrogge [5]. The depth map is easily converted into a 3D object representation. The value of the map simply contains the z−value of a control point at the corresponding (x,y)−position in the image. The number of control points available determine the spatial resolution of the terrain object. Fig. 4a. demonstrates the object generation using stereoscopic images.



4a. Terrain model generation from disparity range maps.

4b. Overlay of object texture using texture mapping.

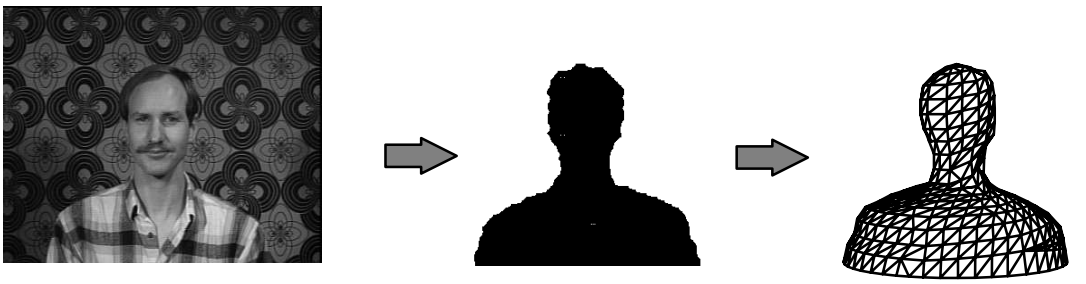
Fig. 4. Object generation for background terrain modelling.

Once the object shape is determined, natural texture is applied to the object. Therefore the image texture is projected onto the object shape to yield a complete object description, as shown in Fig. 4b. Each triangular face of the object surface holds a triangular image patch. For each face the corresponding triple of control points is projected into the image plane. The image area inside this triple is selected as reference image for that face. During image synthesis this projection is reversed and a natural looking image is produced.

3.2. Objects are moving arbitrarly in front of a still camera

This scene type is used for analysing unknown object motion and object shape. The presented algorithm was primarly designed for this purpose, namely for the automatic extraction of motion and shape of a speaker moving in front of a video telephone. A video telephone coding system that uses 3D modelling to reduce transmission bandwith and to further enhance image quality is described by Musmann et al. [6]. A person moving in front of a camera with head and shoulders visible is to be modelled. A monocular or stereoscopic image sequence may be used. In a first step the object silhouette is obtained either through the detection of changed image areas in subsequent image frames (monocular view) or through object segmentation using depth maps (stereoscopic view). Either method arrives at a segmentation of the image into physical object silhouettes. The image segmentation is described in further detail by Hötter and Thoma [7] and Waldowski [8].

Next step during object generation is to guess the unknown object shape in order to arrive at a 3D description. In the case of human beings as viewed by video telephone sequences, a rotational symmetric object may be generated. This rough estimate has to be refined throughout the adaptation phase. Finally the object texture is projected onto the object surface as described before. Fig. 5. shows the process of initial object generation for the head−shoulder scene typical in video telephony.



5a. Image input frames 5b. Object silhouette 5c. 3D model object

Fig. 5. Object generation for the head−shoulder scene.

4. THE ANALYSIS BY SYNTHESIS ALGORITHM

Once the initial objects are generated, they must be adapted throughout the image sequence. Object motion is to be compensated and the model objects derived from the initial object generation procedures need refinement in shape. This can be accomplished using the analysis by synthesis algorithm as shown in Fig. 6. The basic algorithm was published by Kappei and Liedtke [9, 10]. It compares subsequent images of the input sequence with a synthesized model image of the model object.

The position and shape of the objects can be adapted by shifting the control points in space. The control point shift is analyzed using observation points. An observation point is a point on the object surface containing color intensity and spatial gradient information of the reference image that is projected onto the model object. The gray level differences between the image synthesized from the model and the input sequence are evaluated to drive the adaptation process. The algorithm estimates full 3D motion and shape parameters for each model component. For each observation point on the object surface a linea-

rized equation is set up which relates the 3D motion of the observation point to luminance changes in the image plane. A signal model is applied that uses spatial and temporal image gradients to estimate the observation point motion from changed image intensity. To enhance the stability and convergence of the solution and reduce computational effort, only selected surface points are accepted as observation points. Therefore surface points that carry substantial gradient information and exceed a given noise threshold are selected as observation points.

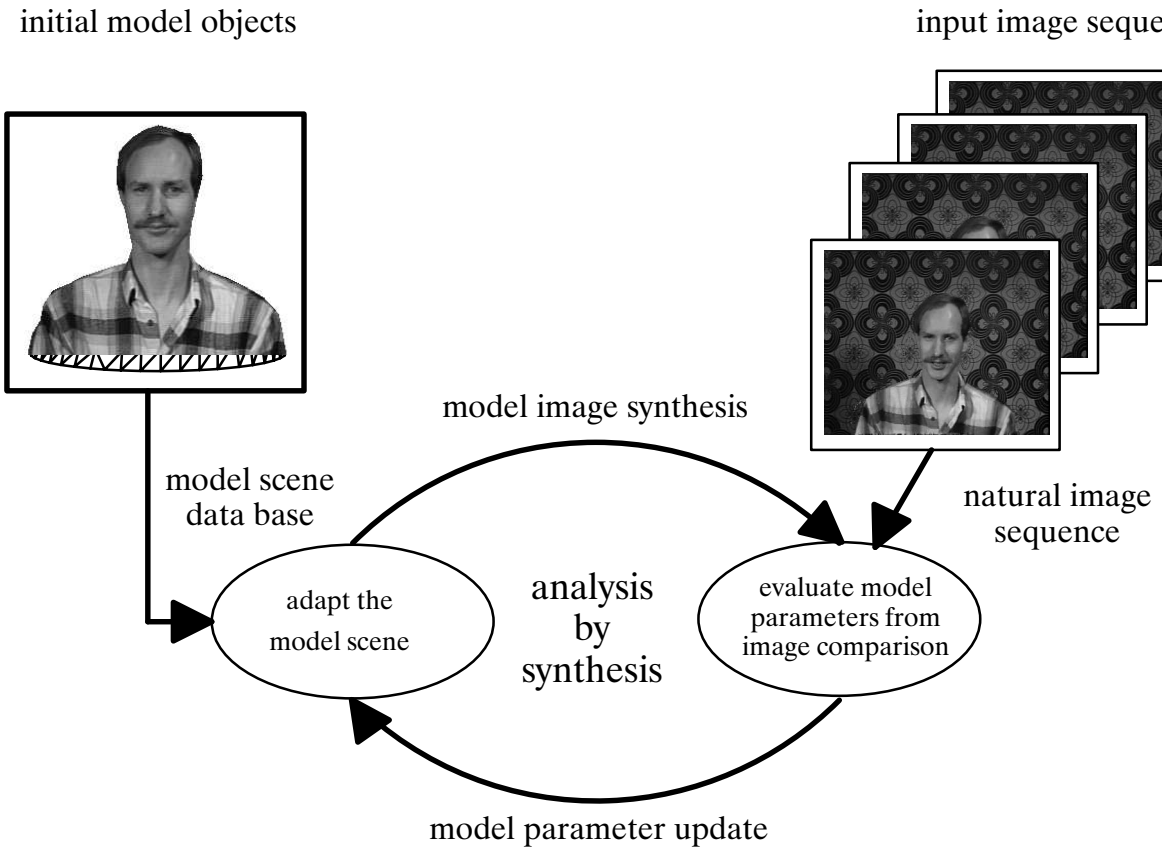


Fig. 6. The model based analysis by synthesis algorithm.

The equations for 3D motion estimation are stated in the following section. See [10] for further details. Translation is defined as motion of center of object gravity. The rotation is defined as rotation around the center of object gravity. The 3D motion of each observation point **P** is described as

$$\mathbf{P'} = \mathbf{[R](P - C) + C + T} \tag{1.}$$

where
- $\mathbf{P}$ = position vector of observation point before motion
- $\mathbf{P'}$ = position vector of observation point after motion
- $\mathbf{C}$ = center of object gravity
- $\mathbf{T}$ = translation vector
- $\mathbf{[R]}$ = rotation matrix

For small rotations the linearized 3D motion equation for the position elements can be expressed:

$$\begin{aligned} P'_x &= P_x - (P_y - C_y)R_z - (P_z - C_z)R_y + T_x \\ P'_y &= (P_x - C_x)R_z + P_y - (P_z - C_z)R_x + T_y \\ P'_z &= (P_x - C_x)R_y + (P_y - C_y)R_x + P_z + T_z \end{aligned} \tag{2.}$$

where $P_{x,y,z}$ = elements of old position vector of observation point
$P'_{x,y,z}$ = elements of new position vector of observation point
$C_{x,y,z}$ = elements of center of object gravity
$T_{x,y,z}$ = elements of translation vector to be estimated
$R_{x,y,z}$ = elements of rotation vector to be estimated

Perspective transform of the position vector into the image plane yields:

$$B_x = P_x F / P_z$$
$$B_y = P_y F / P_z \quad (3.)$$

where $B_{x,y}$ = position in image coordinates
F = focal length of camera

The use of a linear signal model relates image intensity changes to changes of observation point projection into the image plane:

$$I' = I + G_x(B_x' - B_x) + G_y(B_y' - B_y) \quad (4.)$$

where I = luminance intensity at position of B
I' = estimated luminance intensity at position B'
G_x, G_y = spatial luminance gradient at position B

When substituting equation (2.) and (3.) in (4.) and linearizing the resulting equation a relationship between the 3D motion parameters and the image intensity change is found:

$$\begin{aligned} I' - I &= FG_x / P_z T_x \\ &+ FG_y / P_z T_y \\ &- F(P_x G_x + P_y G_y) / P_z^2 T_z \\ &- F(P_x G_x(P_y - C_y) + P_y G_y(P_y - C_y) + P_z G_y(P_z - C_z)) / P_z^2 R_x \\ &- F(P_y G_y(P_x - C_x) + P_x G_x(P_x - C_x) + P_z G_x(P_z - C_z)) / P_z^2 R_y \\ &- F(G_x(P_y - C_y) - G_y(P_x - C_x)) / P_z^2 R_z \end{aligned} \quad (5.)$$

This equation is formulated for each observation point and solved for the 6 motion parameters using least squares fit. Assuming rigid components for motion estimation one parameter set for each component can be estimated. About 100 observation points spread across the component surface are sufficient for reliable motion estimation even in noisy image sequences. The global error in frame differences is compensated with this method.

In a similar way local object deformations can be estimated to adapt the object shape. Residual local errors in the frame difference are assumed to be due to local shape deformations of the object. The shape of the object can be manipulated through shifting individual control points of the object surface. Therefore all observation points grouped around each control point are used to estimate the shape deformation. The control point is moved along the line of sight until the local frame difference is mimimized. This leads to a local adaptation of the object shape throughout the sequence. The next step is to compensate the flexible surface of the object, i.e. the mimic in the persons face. The local displacement vector of each control point is estimated in a plane perpendicular to the line of sight and the control points are shifted accordingly.

The final step in adapting the model szene is to add texture to newly visible parts of the object. When the person rotates the head or opens the mouth, new surfaces are visible and the object texture must be updated.

This strategy to adapt the objects is further described by Liedtke et al. [11]. Some results using that strategy are displayed in Fig. 7. An input image sequence of a speaker in front of a planar background is analyzed. The model scene consists of two objects: a planar background object and a speaker object. The speaker object shape is adapted throughout the sequence and the object motion is tracked. Image 10 and image 50 of the image sequence are shown with the corresponding model objects from the front view (left) and side view (middle). The speaker object is projected untextured and with artificial illumination to display the object shape. On the right side the synthesized image is displayed. No differences between input image and synthesized image are visible.

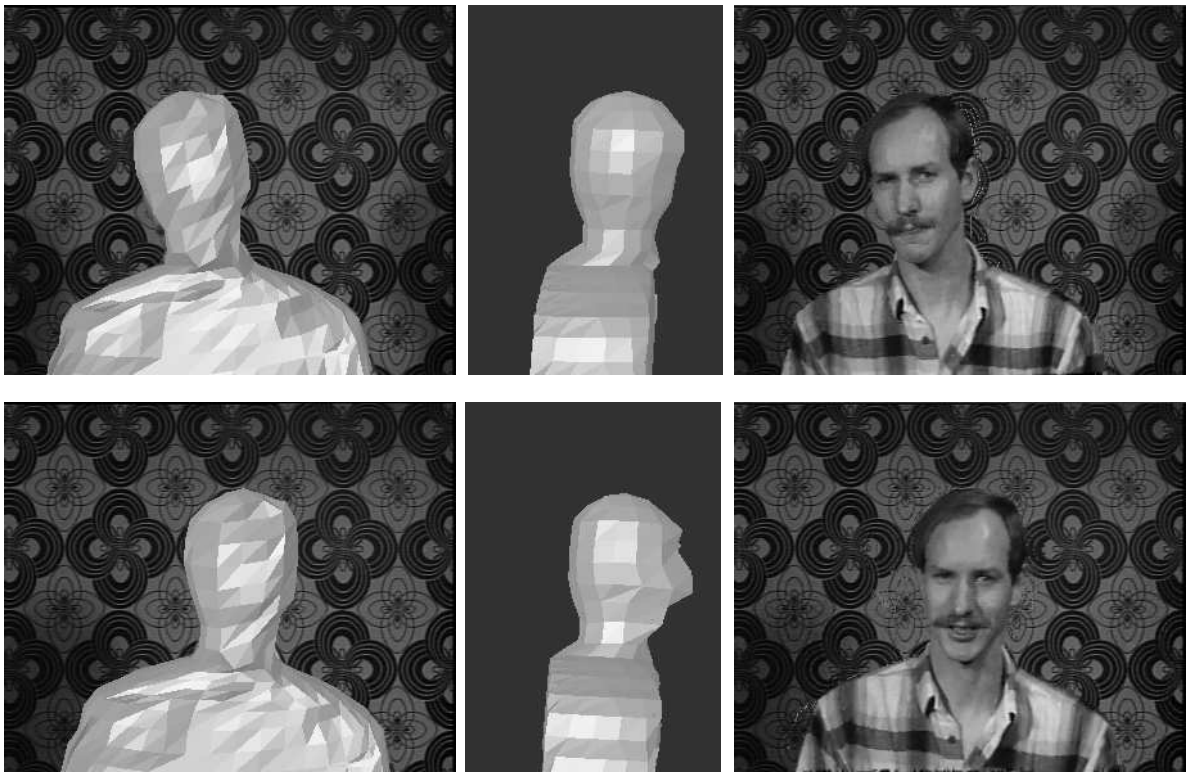


Fig. 7. a|b|c d|e|f — Modelling results: (a) model image front view and (b) model image side view and (c) synthesized image after processing of 10 input images; (d) model image front view and (e) model image side view and (f) synthesized model image after processing of 50 input images [11].

5. GENERATING SYNTHETIC MOVIES OUT OF THE MODEL SCENE

Up to now object generation and adaptation using input image sequences have been described. A 3D scene database consisting of several model objects has been built. Those different types of objects may now be used to synthesize new image sequences out of the model scene. Large static background objects may be combined with small moving objects to create synthetic movies. It is possible to generate stereoscopic views from a speaker that is placed into an artificial generated environment, not just onto a background

image. Realistic views of outdoor terrain may help landscape architects to judge their work during design. Many different applications may be envisioned. Two examples will be provided here. In Fig. 8. the speaker model from the previous section is placed into an outdoor environment and forced to perform synthetic head movements. The speaker is moved around and the head is bent 20 degree to illustrate the ability to manipulate and still get natural looking images.

Fig. 8a. Speaker placed into landscape.

Fig. 8b. Speaker moved and head bent 20 degree.

The second example in Fig. 9. shows one image out of 800 from a synthetic movie simulating a flight across the Black Forest in West Germany. A standard 5x5 km^2 terrain model was overlaid with an aerial photography of 5000x5000 pixels, yielding a resolution of 1 m^2 for the image texture. The camera was moved about 100 m above ground and with a relative speed of 500 km/h. The high resolution of the input image leads to a highly realistic, natural looking image sequence.

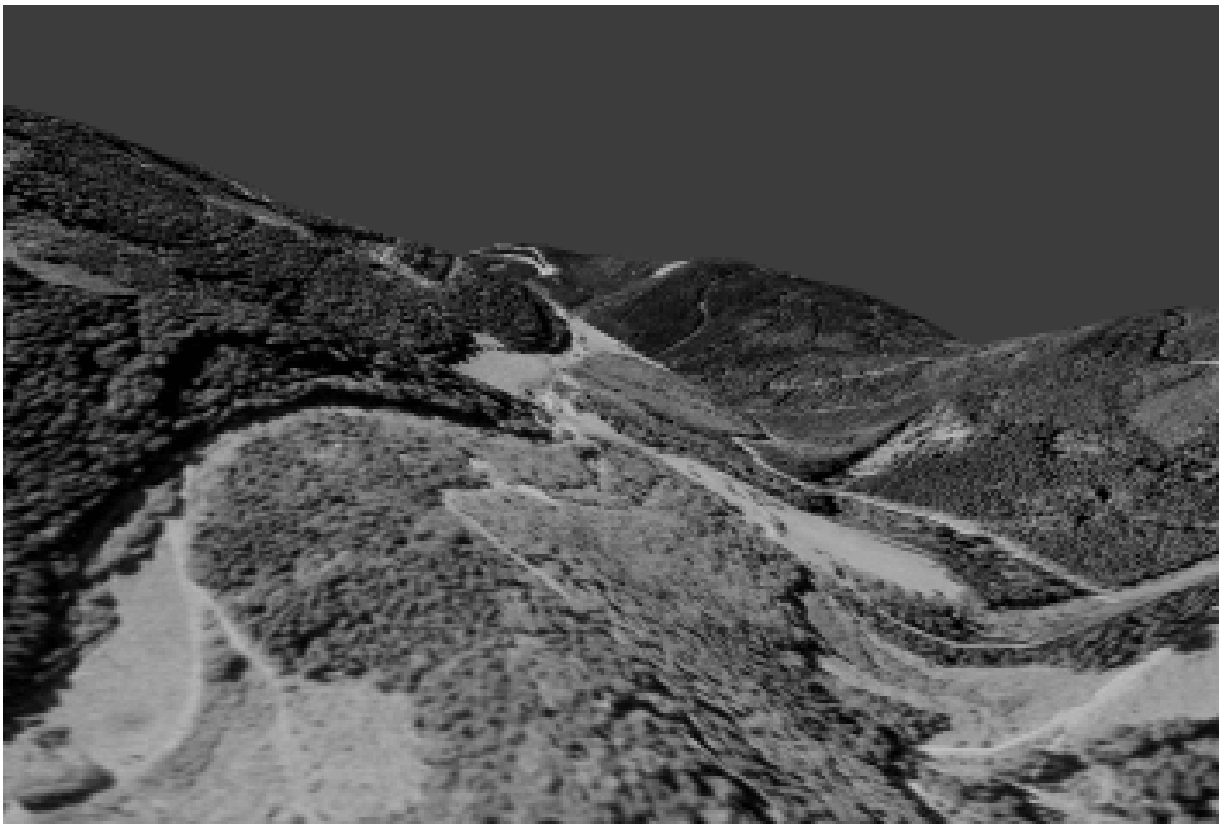


Fig. 9. Frame out of a synthetic movie simulating a flight across the Black Forest.

ACKNOWLEDGEMENTS

The work has been supported by the Forschungsinstitut der Deutschen Bundespost TELEKOM.

REFERENCES

[1] J.D. Foley, A. van Dam, Fundamentals of Interactive Computer Graphics (The Systems programming series). Reading, MA., Addison−Wesley Publ. Comp., 1982.

[2] R.A. Jarvis, "A Perspective on Range Finding Techniques for Computer Vision", in IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI−5, No. 2, pp. 122−139, March 1983.

[3] J.K. Aggarval, N. Nandhakumar, "On the Computation of Motion from Sequences of Images − A Review", Proceedings of the IEEE, Vol. 76, No. 8, 1988.

[4] H. Busch, "Automatic Modelling of Rigid 3D Objects Using an Analysis by Synthesis System", SPIE Symposion on Visual Communications and Image Processing, Nov. 1989, Philadelphia, PA, USA.

[5] V. Sasse, G. Altrogge, "Realisation of automatic correlation within a digital stereo plotter", in Proc. of the Int. Society of Photogrammetry and Remote Sensing, Commision II, Kyoto 1988.

[6] H.−G. Musmann, M. Hötter, J. Ostermann, "Object−oriented analysis−synthesis coding of moving images", in SIGNAL PROCESSING, image communication, No. 2, Nov. 1989.

[7] M. Hötter, R. Thoma, "Image Segmentation Based on Object Oriented Mapping Parameter Estimation", in SIGNAL PROCESSING vol. 15, No. 3, pp. 315−334, Oct. 1988.

[8] M. Waldowski, "Segmentation of a typical video phone scene based on stereo image pairs", Rank Prize Funds Mini−symposium on Model Based Image Coding. Broadway, England, Dec. 1988.

[9] F. Kappei, C.−E. Liedtke, "Modelling of a Natural 3−D Scene Consisting of Moving Objects from a Sequence of Monocular TV Images", SPIE, Cannes, 1987.

[10] F. Kappei, "Modellierung und Rekonstruktion bewegter dreidimensionaler Objekte in einer Fernsehbildfolge", Ph.D.−Thesis, Universität Hannover, 1989.

[11] C.−E. Liedtke, H. Busch, R. Koch, "Automatic Modelling of 3D Moving Objects from a TV Image Sequence", SPIE/SPSE Symposium on Electronic Imaging, Santa Clara, CA., USA, Feb. 1990.