How Far Away are We From the Virtual Actor and the Virtual Scene?

C.-E. Liedtke, S.Weik, J.Wingbermühle, M. Kampmann, L. Falkenhagen, W. Niem
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover - Germany
email: weik@tnt.uni-hannover.de

Abstract

This paper discusses 3-D generation techniques in view of virtual actors and scenes. Naturally looking virtual actors and scenes are obtained using current 3-D reconstruction techniques that describe real objects by 3-D models. Four techniques are introduced: For the generation of virtual sets, 3-D models of rigid objects are generated using a shape from silhouette approach. Arbitrary time-varying scenes are reconstructed using a stereo vision technique. When dealing with humans, this knowledge can be exploited for improving the 3-D models. Therefore, model-based reconstruction techniques for persons are described using body animation for whole persons and facial animation when facial expressions are of interest. The naturalness of these 3-D models is illustrated, and their benefit for movie production is discussed.

Keywords: 3-D reconstruction, body animation, facial animation, virtual actor, virtual set, 3-D modelling

1 Introduction

Virtual scenes and virtual actors can be seen as sets of data which can be handled and manipulated by a digital computer. They become alive and can be viewed at from various positions using visualization equipment like computer monitors, specialized VR equipment or by the presentation of images or image sequences printed on paper or written on films. They are widely used in labs and studios to produce purely virtual worlds or can be mixed with natural scenes, which are recorded by cameras from the real world. Many methods exist, how real scenes can be combined and mixed with synthetic scenes [3]. In this contribution we will concentrate on 3-D computer generated virtual objects and actors only, which have the advantage, that they can be positioned and act freely in 3-D space and can be observed from one or more arbitrarily chosen positions.

Virtual actors and scenes are not just a matter of the far future, they are already used in a variety of applications for different purposes and with different (technical) constraints. Some examples will be mentioned in the following.

Virtual actors are used for display of home furniture equipment, for simulations in ergonomic tests like car seats, bath tubs, etc., for the planning and simulation of military strategies, the planning of industrial handling and production processes. An example is the software system Jack[2], which has been developed from the University of Pennsylvania and is marketed by an US company. The purpose of Jack is to provide a high-fidelity human model to populate a virtual scene and to interact with it. Another area is the entertainment industry. Interactive use of virtual objects and actors in the game industry requires extreme efficiency for the rendering process, leading to virtual actors like Lara Croft [15] in the surrealistic computer game *Tomb Raider*. Another request can be high-fidelity modelling and nice appearance as in the case of the Japanese teenage popstar, dancer and singer Kyoko Date [14]. In all these cases computing efficiency has to be traded against natural appearance. The actors and scenes look synthetic.

In the movie industry there is a high demand for perfectly natural appearance. Examples are synthetic stunts for special effects in action scenes which are not possible in a real physical world or the appearance of virtual creatures like the dinosaurs in *Jurassic Park*. In addition there is a request for persons to act, which either do not live any more or who are for various reasons not available for the requested task. A famous example is the revival of Humphrey Bogart and Marilyn Monroe in the 7-minute-film *Rendez-vous à Montréal*, which was presented already in 1987 by Nadia Magnenat Thalmann and Daniel Thalmann [20]. The demand for natural appearance of an actor which can be animated without any restrictions poses extreme demands to the precision of the required modelling process. The creation of human skin, the growing of hair, the modelling of cloth wrinkling, the bending of a scarf in an air breeze become scientific projects. Outstanding work in this direction has been done by Prof. Thalmann from the MIRALab at the University of Geneva [21]. The progress, which has been achieved over the past years is promising. Nevertheless, the questions, which physical principles have to be simulated to which extent in order to produce a subjectively natural appearance of a virtual actor and a virtual scene has not yet completely been solved and are still a matter of current basic research in Computer Graphics. The computing time required so far only permits the off-line production of movies and is still far from real time processing as would be needed for interactive systems.

Another approach to the realization of naturally appearing virtual actors and virtual scenes comes from the telecommunication industry. Telecommunication involves persons which are situated in different places and who communicate by word, facial and body expressions. The goal of present and future research is to develop technical systems which suggest the (virtual) presence of the remote communication partner. For the technical presentation system the communication partner becomes a *virtual actor* in a *virtual scene*. Since communication is always online, the processing has to be done in realtime. The virtual partner has to look and to behave exactly like the original partner in his remote location; any difference would unduly disturb the communication process.

A significant difference to the virtual actor for the movie industry as mentioned above is, that during communication for any time instant the original person exists in real life. Therefore the technical system can observe the original and needs just to reconstruct the visual appearance of the body surface but does not need to simulate the underlying and invisible physical processes. The primary goal of 3-D modelling in telecommunications becomes the correct reconstruction of the 3-D scene and the 3-D persons from the image sequences of several observing cameras. The automated analysis process can be reduced to a parametric surface reconstruction of the visible parts, where the model parameters have continuously to be adapted to the visual appearance of the original.

The viewing position and orientation of the virtual scene including the virtual actors is usually limited to the viewing positions of the observing cameras in the real scene. Invisible parts cannot be reconstructed. If for future communication systems the user will be permitted to choose an arbitrary viewing angle of his partner, as is used in any natural local face to face communication, a complete surface reconstruction of the partner is required. The reconstruction of the parts of the scene and the actors which are not seen by the observing cameras during the communication session has to be done in a preceding off-line step. During the communication session itself only the 3-D motions and the shape and surface properties of the parts visible to the observing cameras can be adapted. For this purpose surface models of virtual actors are under development, which appear correctly from arbitrary viewing angles, even if only their motion parameters are controlled, i.e. these models can be freely animated.

The reconstruction of the virtual actor in telecommunications can thus be reduced to the on-line measurement of the spatial motion parameters, the body shape, and its surface properties like colour and texture. In order to prevent reconstruction errors, which could produce implausible appearances or motions the degrees of freedom of the model description have to be reduced by model assumptions. These model assumptions contain the prior knowledge about meaningful relations between the parts of the data, which form the actor and represent a human being or the parts which describe the scene. As an example, they may contain information about the spatial relations between the parts of the human body, their connectivity, about the maximum speed of motion, the maximum tilt of joints or they may describe which parts of the control points describing the local shape can be affected, when a single muscle in the face is constricted. These model assumptions are used in different detail and on different levels of abstraction. Often they refer to meaningful and explicit parts of the scene, like *arm*, *head*, *eyes*, *nose*, etc., in the description of a human, the positions of *wall* and *roof* of a house, etc. Using these model assumptions, the control parameters of the virtual scene and the virtual actor are reduced and become more abstract as well. As an example, instead of referring to the 3-D positions of hundreds of vertices of the triangular mesh, which describe the shape of the head, they can all together be referred to by just one parameter describing the *bending of the head*.

The parameters, which control the appearance of these scene parts and the virtual actor are usually obtained by measurements from the observed original. Instead of using these measured parameters one could use parameters from different sources instead, for instance interactively created parameters or parameter sets from a different scene. The result would be, that objects could be placed in other positions, facial expressions and body motions could be manipulated interactively or adapted to other persons, etc. To a limited extent the virtual actor and the virtual scene can be changed compared to the original scene without looking *unnatural*.

Since our major field of application is telecommunications in the following different aspects of our approaches to the reconstruction of virtual scenes and virtual actors will be explained. Similar approaches as described here are considered for standardization in communication techniques (MPEG-4 [10]) for the transmission of models of humans.

Section 2 describes the automatic 3-D reconstruction of rigid objects by taking several camera views from all around the object. When objects are moving or moving actors appear in the scene this approach cannot be used any more. Therefore in Section 3 a general approach to the automatic 3-D reconstruction of time varying scenes is described. If it is known that the moving part of the scene is a human, the reconstruction can be improved by using model based techniques specialized for humans, which are described in Section 4. An important part in visual communications are facial expressions. Section 5 describes in this context our approach to model based reconstruction of facial expressions.

2 Automatic 3-D reconstruction of rigid objects

Virtual scenes are composed of virtual 3-D models. The manual generation of such 3-D models, e.g. by using a CAD tool, is time consuming and therefore cost expensive. Hence, techniques for automatic 3-D reconstruction will simplify the building of virtual environments a lot, because they rapidly provide 3-D models of natural rigid objects [19][23][5]. Shape as well as texture are automatically estimated and are represented in a parameterized 3-D model, which can be visualized on a computer from arbitrary viewpoints as done e.g. while rendering a virtual studio scene [3].

We developed a new method for the automatic 3-D reconstruction of rigid objects, which simply uses calibrated camera images of the object from multiple views. The calibrated images are obtained for example with a turntable environment as shown in Fig. 1 or by using an underlying calibration pattern as shown in Fig. 2.

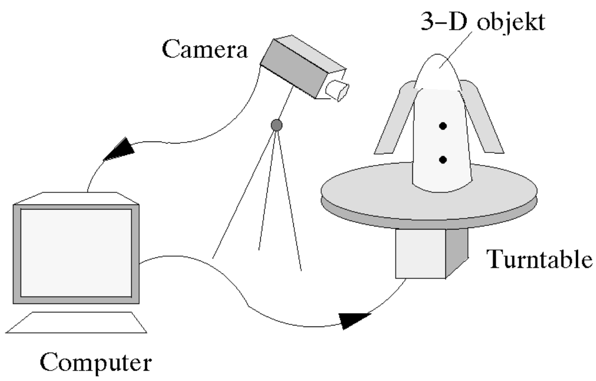

Figure 1: Calibrated turntable environment

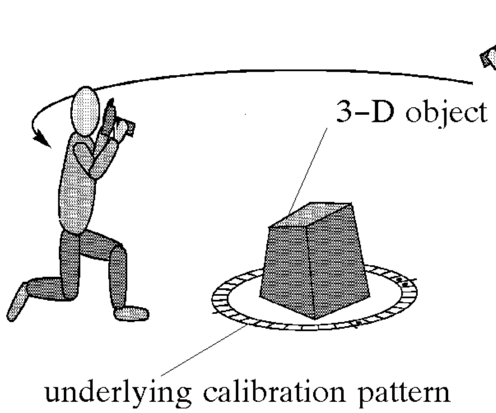

Figure 2: Environment with underlying calibration pattern

In the turntable environment, the camera is stationary mounted in front of the platform and its position and orientation is calibrated once in advance. Then for image acquisition the turntable with the real object on top is rotated in predefined steps until the object is completely registered.

In case of using the underlying calibration pattern, for image acquisition the camera is freely moved around the object. The camera calibration is obtained individually for each image by evaluating the calibration points of the underlying pattern.

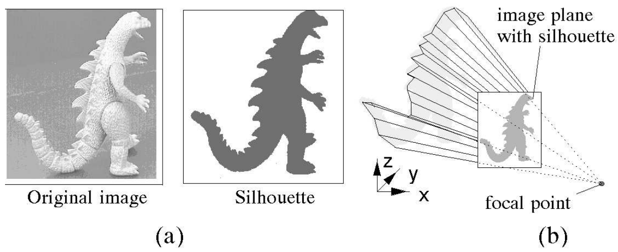

Figure 3: Shape-from-silhouette

Once the images are obtained, the 3-D reconstruction is performed in two steps: shape estimation and texture estimation [4][18].

The shape is estimated by evaluating the silhouettes from the calibrated views by using a technique known as *shape-from-silhouette* or *method of occluding contours* : At first, the silhouettes are obtained from the images by using colour keying (Fig. 3a), which is facilitated by a uniformly coloured background. Then volume cones are formed by the lines of sight through the silhouette contour points and the focal points of the individual views (Fig. 3b). The intersection of the single volume cones form the final volume model. The volume model surface is approximated by a locally adaptive wireframe, which consists of smaller triangles in surface regions of higher curvature and larger triangles for smoother surface parts.

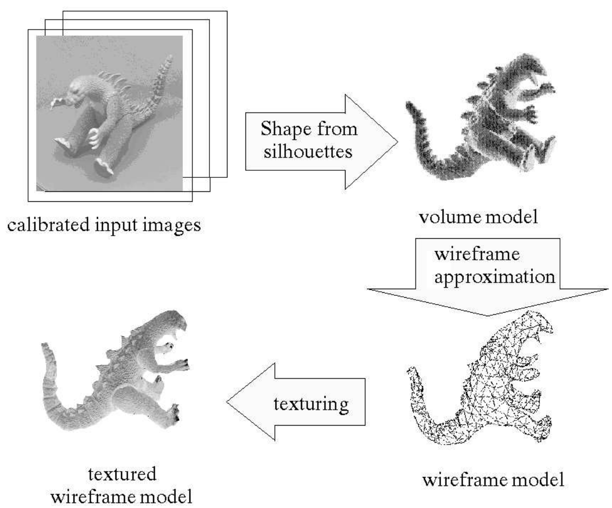

Figure 4: Reconstruction of 3-D objects from calibrated views

The texture of the object is estimated and added to the wireframe model in order to obtain a natural look (see Fig. 4). For this purpose, each triangle of the wireframe is assigned to a sector of the original images. Those image sectors (texture maps) are obtained by projecting each triangle of the wireframe model into the camera plane of an input image. Applying texturing significantly enhances the realism by adding detail to the 3-D model.

The technique has been successfully tested for a variety of objects. In Figs. 5 and 6 synthetic generated views of different automatically reconstructed objects are shown. In Fig. 14 the object *doll* from Fig. 6 is situated in a synthetic environment.

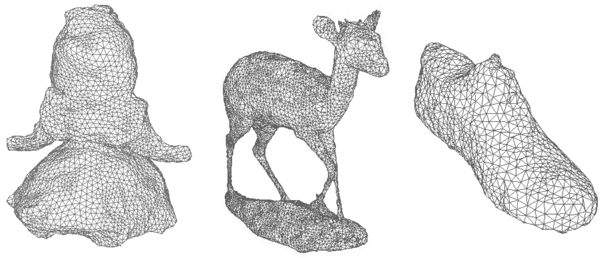

Figure 5: Synthetically generated views of automatically generated 3-D models (wireframes)

Figure 6: Synthetically generated views of automatically generated 3-D models (textured)

3 Automatic 3-D reconstruction of time varying scenes

So far, the 3-D reconstruction of rigid objects was described. When more general scenes with several moving objects are considered, the 3-D reconstruction concept has to be extended. Due to object motion, it is not possible to successively capture images from different view points.

By using a multi-camera approach [7], where the different views are captured simultaneously, also moving objects can be reconstructed. The rigid 3-D models that are described by their shape and surface colour, are now extended to dynamic models by introducing the motion of the objects (Fig. 8).

Such a dynamic reconstruction approach is used for the development of an advanced telecommunication system with 3-D telepresence in the European ACTS-PANORAMA project. There, it is used to compute new views of the communication partner according to the viewers head position. This provides the feeling of physically being close to the communication partner and therefore introduces a new quality in telecommunications.

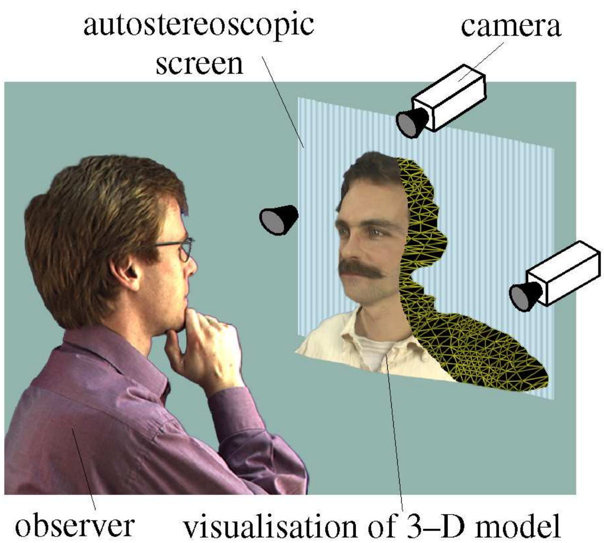

Figure 7: Trinocular camera setup for telecommunication application

A calibrated trinocular camera setup is used for observing the communication partner and his environment (Fig. 7). The cameras are arranged around an autostereoscopic screen that shows two images simultaneously, one for the left and one for the right eye. Lenticular lenses in the screen spatially separate the two views in such a way that each view is directed to one of the eyes of the viewer and therefore avoid to wear special glasses for 3-D viewing.

The observed scene is reconstructed by analysing the trinocular image sequences (Fig. 8). The reconstruction can be subdivided into two phases: the initialisation phase, where new objects occurring in the scene are analysed, and the update phase, where temporal changes of already reconstructed objects are analysed and the quality of 3-D models is successively improved.

In the initialisation phase, a set of depth maps is analysed from the input triple [8]. These depth maps represent the distance between the focal point of the camera and the

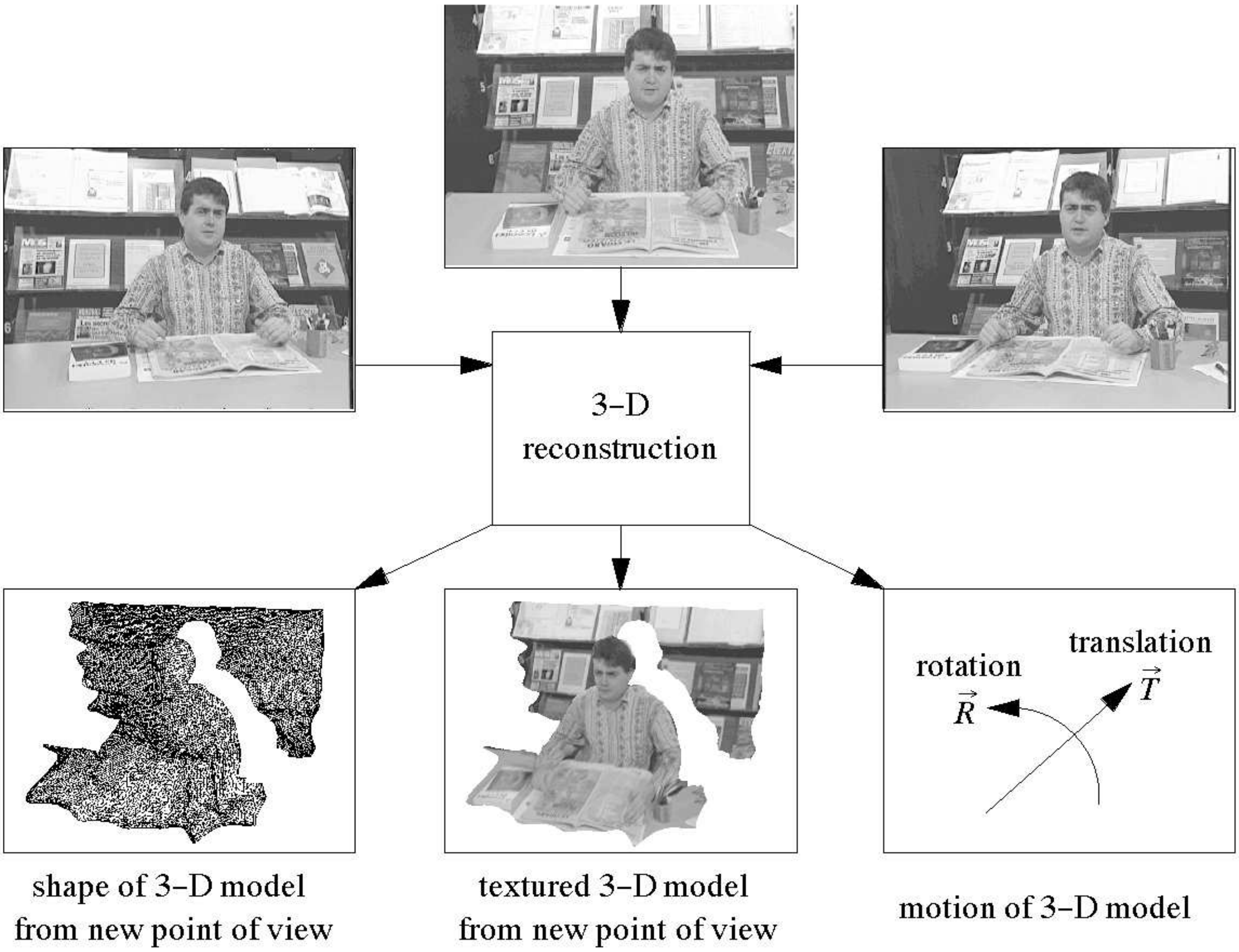

Figure 8: Block diagram of the scene reconstruction using textured 3-D models that are analysed from a trinocular image sequence of the scene

point in the scene for each picture element. The depth values are back-projected into 3-D space, resulting in a cloud of 3-D points. These points are interpolated in a surface model that is further approximated using a triangular mesh that represents the shape of the objects [24]. The texture of the objects is initialised by mapping parts of the image triple onto each triangle [18].

The update phase starts by analysing the motion of the objects, assuming that most of the changes result from motion of rigid objects [12]. In a next step, the shape of the 3-D models is updated. This update considers flexible deformations of the object which cannot be described by motion and improves the shape over time. Finally the texture of the 3-D models is updated, compensating the changes of the objects surface colour. The resulting 3-D model is stored in order to analyse the next image triple and is used for the computation of images from arbitrary points of view.

This approach provides realistic looking virtual images of arbitrary moving scenes. The position of virtual view points is restricted to an area that is covered by the real camera triple, because the scene is reconstructed from these positions. For example, the 3-D models are not closed at their back when only front views are captured and therefore no meaningful virtual back view can be computed.

4 Model based reconstruction of persons

If one wants to permit arbitrary viewing positions a complete closed 3-D model of the communicating virtual actor has to be provided. In order to copy the natural movements of the original communication partner to the virtual actor, the latter needs to be animated in its motion and controlled by the parameters, which have been derived from the original person's movements. For this purpose a modelling technique has been developed, which permits model based motion animation.

Virtual actors consist of an internal, hierarchical skeleton and a surrounding surface mesh describing the actual shape of the real actor [25]. The internal structure which is a simplified human skeleton is linked to the external surface and controls its motion dependent shape adaptation (Fig. 9). The external surface consists of a triangular mesh which is organized in rigid vertex rings along the internal bone structure (Fig. 10 and 11).

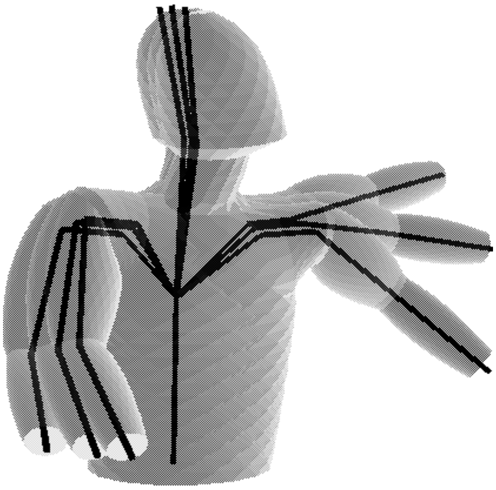

Figure 9: Initial template in several postures

The internal skeleton is organized in a hierarchical way; each element is connected to exactly one parent bone but can have one or more children, e.g. the upper torso has the lower torso as its parent and the neck and the left and right shoulder as children. The elements are connected by joints which store the actual position with up to three degrees of freedom per joint. The values are determined relatively to to the respective parent, thus moving a part automatically moves all children as well. The relative coordinates also enable us to check for motional constraints i.e. suppression of physically impossible postures.

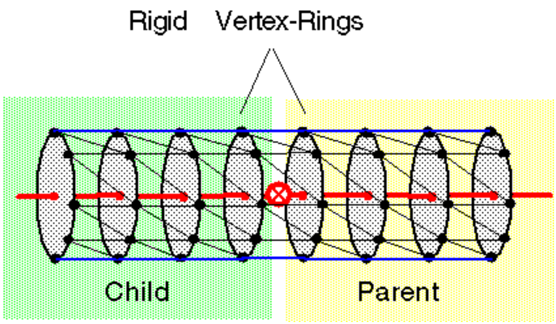

Figure 10: Flexible surface model (stretched)

Each element is surrounded by a triangular mesh whose vertices are arranged in rigid rings along the respective bone (Fig. 10 and 11). The rings are tilt and bent when an element is moved in its parent joint. Thus the overall motion is interpolated from ring to ring, Fig. 11 shows the principle.

The movable person-template can then be animated using the motion of a real actor. To make the template look like

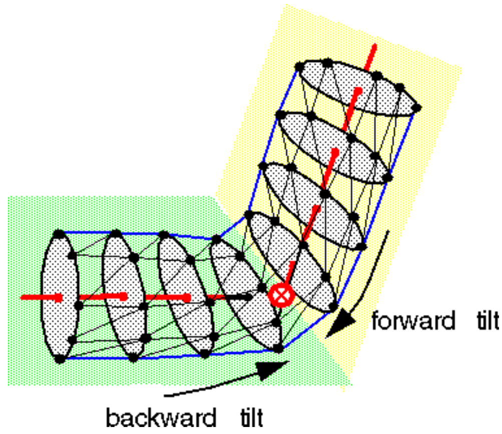

Figure 11: Flexible surface model (bent)

the original the shape and texture have to be adapted. The 3-D shape of the real actor is measured either employing a stereoscopic approach or a laser scanner, resulting in a cloud of surface sample points. The template is then inflated to this 3-D data set; each triangle of the template is moved along its surface normal until the deviation to the measured 3-D surface points vanishes.

Figure 12: Inflated and textured template

After having adapted the shape the template is textured with original images of the actor, i.e. the image information is projected onto the triangular mesh; when the mesh is moved and deformed the texture information follows the same motion. The inflated and textured model is depicted in Fig. 12. Fig. 13 shows the virtual actor in different states, animated using real motion of a typical conferencing scene. The upper two rows show the original sequence and the internal skeleton whereas the lower two rows depict the initial and the inflated and textured virtual model respectively. A completely virtual scene shows Fig. 14 where two of those virtual actors are situated in a synthetically created environment. The presented method for reconstruction of persons is described in more detail by the authors in [16] and [26].

Figure 14: Two virtual actors in a synthetic scene

5 Model based reconstruction of facial expressions

Facial expressions are an important part of personal communication. For telecommunication applications like 3-D telepresence or semantic coding (sometimes also called model based coding) of videophone sequences at very low bit rates[17][1][22], the natural facial expression of the communication partner have to be reconstructed by the animation of a virtual actor. In order to achieve a high degree of naturalness, a parametric description of facial expressions is often carried out. This description by mimic parameters allows a limitation of all possible facial expressions to natural looking facial expressions. One example for a parametric description of facial expressions is the use of the human facial muscles [9]. The degree of contraction of each muscle is described by a parameter. Moreover, Action Units are used as mimic parameters [6]. Action Units are defined as noticeable visible changes in the facial expression of a human person. They are attached to one or more facial muscles. For the reconstruction of the natural facial expression of the communication partner, the corresponding mimic parameters of the original person are estimated by image analysis. With these estimated mimic parameters an animation of a virtual actor is possible achieving the same facial expression as the original person.

For reconstruction of facial expressions, the human face is represented by a generic 3-D face model consisting of a mesh of triangles (Fig. 15 [9]). This face model has the proportions of an average human face. In order to adapt this generic face model to the size and to the individual anatomy of the face of the original person (Fig. 16), facial features like the positions of eyes, nose and mouth as well as contours of chin and cheeks are estimated by image analysis [11][13]. According to the estimated facial features, the proportions of the generic face model are changed for adaptation to the individual face. After adaptation, the face model is textured by projecting the corresponding image area of the person's face onto the face model's surface. An animation of facial expressions is carried out by a deformation of the mesh of triangles of the face model. For deformation the vertices of the mesh of triangles are shifted. The magnitude of shifting is controlled by the values of the mimic parameters. Only those vertices are shifted which are in the sphere of action of the respective mimic parameter e.g. in the sphere of action of a specific facial muscle.

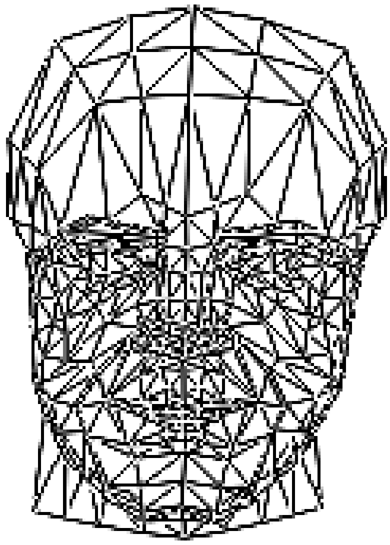

Figure 15: Face model

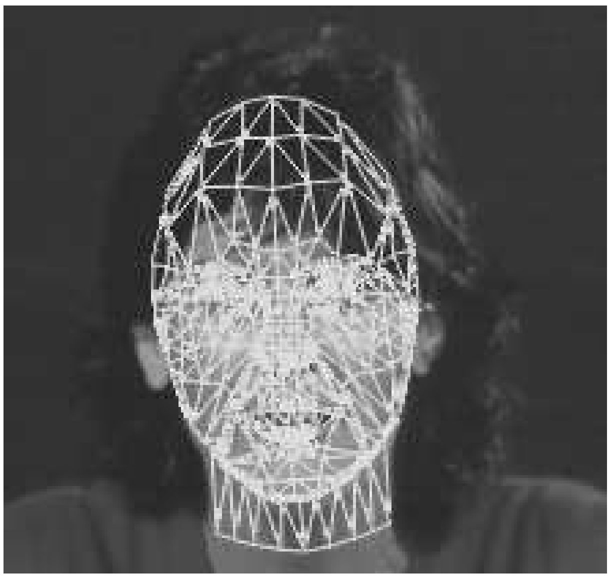

Figure 16: Adapted face model over original image

Fig. 17 shows some examples for animated facial expressions. Texture from the original image (Fig. 17 (a)) has

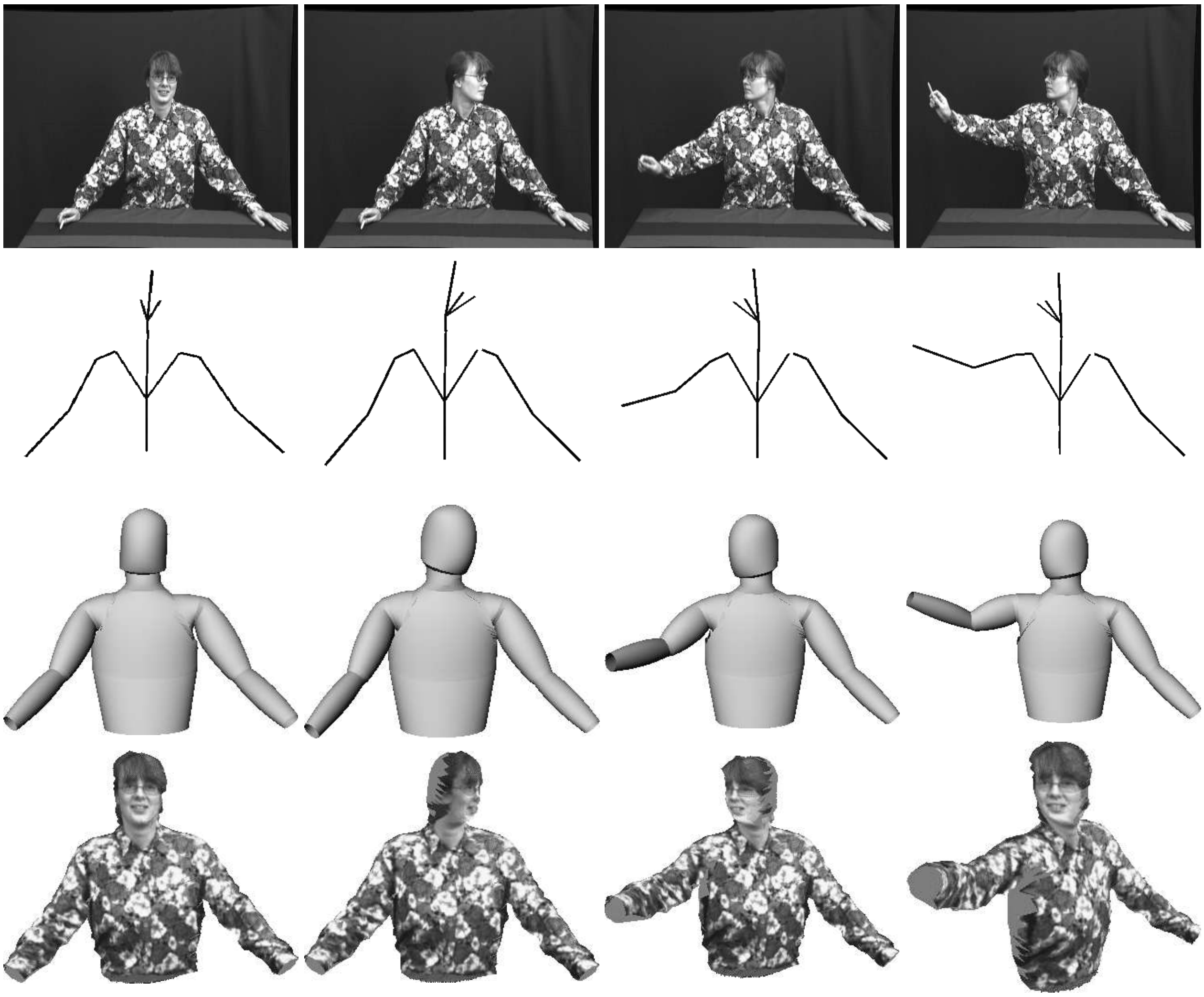

Figure 13: Virtual actor following real motion

been projected onto the face model. For the animation of facial expressions (Fig. 17 (b)(c)(d)), an animation of the human facial muscles is carried out. Eleven facial muscles are chosen. According to the contraction of these muscles the mesh of triangles is deformed by shifting corresponding vertices of the face model.

6 Conclusions

Several methods are under development in telecommunication techniques which aim at the representation of the communication partner as virtual actor in a virtual scene. Since telecommunication is always an on-line process, the model adaptation and the rendering of the scene including the actors have to be done in realtime. The reconstruction techniques have therefore been developed for a high degree of computing efficiency. Communication systems require that the view of the communication partner is transmitted without any visible distortion. Therefore a high degree of naturalness is requested from these techniques. The methods which are presently developed for telecommunications techniques are bound to a human original, which is continuously observed and analysed. Parameters which are extracted automatically from a scene analysis are used to control the virtual actor in the receiving set of the communication partner.

At the present stage of development those techniques provide a very efficient way for a naturally looking reconstruction of humans. Control parameters from various sources can be used to animate the models. The control parameters are described on different levels of abstraction dependent on the extent of model assumptions which have been employed. Animations leading to natural appearance of virtual actors become the easier the higher the abstraction level of the control parameter becomes. For various reasons the tendency of future research in telecommunications is to use more elaborate methods aiming at control parameters of higher abstraction level. Therefore the barriers in using the approaches mentioned above for realizing the virtual actor in applications for the movie industry

(a) (b) (c) (d)

Figure 17: Animation of facial expressions: (a) original image; (b) (c) (d) images with various animated facial expressions

may be shifted as more abstract high level reconstruction techniques will be used to control the analysis and the rendering process.

7 Acknowledgements

This work has been supported with grants from the European Community and the Deutsche Telekom.

References

[1] K. Aizawa, T. Huang, *Model Based Image Coding: Advanced Video Coding Techniques for Very Low Bit-Rate Applications*, Proceedings of the IEEE, Vol. 83, No. 2, February 1995, pp. 259-271.

[2] N. Badler, C. B. Philips, B. L. Webber: *Simulating Humans*, Oxford University Press, Oxford/UK, 1993.

[3] L. Blonde: *The MONA LISA Project* in Image Processing for Broadcast and Video Production. Workshops in computing series, Springer, Hamburg 1994, ISBN 3–540–19947–0.

[4] H. Busch: *Automatic Modelling of Rigid 3D Objects Using an Analysis by Synthesis System*, SPIE Proceedings Visual Communication and Image Processing IV, 1989.

[5] Cyberware: *www.cyberware.com* Monterey, CA 93940, http://www.cyberware.com/ .

[6] P. Ekman, W. Friesen, *Facial Action Coding System*, Consulting Psychologists Press, Palo Alto CA., 1977.

[7] O. Faugeras, *Three-Dimensional Computer Vision: A Geometrical Viewpoint*, The MIT Press, Cambridge, Massachusetts, 1993, ISBN 0-262-06158-9.

[8] L. Falkenhagen, *Block-Based Depth Estimation from Image Triples with Unrestricted Camera Setup*, IEEE Workshop on Multimedia Signal Processing, Princeton, New Jersey, USA, June 23-25, 1997.

[9] J. Fischl, B. Miller, J. Robinson, *Parameter tracking in a muscle-based analysis/synthesis coding system*, Picture Coding Symposium (PCS '93), Lausanne, Switzerland, No. 2.3, March 1993.

[10] C. Horne: *SNHC Verification Model 3.0*, (Status: Approved), ISO/IEC JTC1/SC29/WG11 N1545, MPEG 97, Feb. 1997.

[11] M. Kampmann, *Precise face model adaptation for semantic coding of videophone sequences*, submitted to 1997 Picture Coding Symposium, Sept. 1997, Berlin, Germany.

[12] R. Koch, *Dynamic 3-D Scene Analysis through Synthesis Feedback Control*, IEEE PAMI, Vol. 15, Nr. 6, S556-568, Juni 1993.

[13] 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 Communications, Vol. 9, No. 3, March 1997, pp. 201-220.

[14] *http://www.etud.insa-tlse.fr/~mdumas/kyoko.html*

[15] *http://www.tombraider.com*

[16] Liedtke, Weik, Wingber, Kopernik: *3D-Modellierung bewegter Personen für künftige Videokonferenzsysteme*, submitted paper for DAGM97

[17] H. Musmann, *A layered coding system for very low bit rate video coding*, Signal Processing: Image Communication, Vol. 7, Nos. 4-6, November 1995, pp. 267-278.

[18] W. Niem, R. Buschmann: *Automatic Modelling of 3D Natural Objects from Multiple Views* in Image Processing for Broadcast and Video Production. Workshops in computing series, Springer, Hamburg 1994, ISBN 3–540–19947–0.

[19] W.Niem, J.Wingbermühle: *Automatic Reconstruction of 3D Objects Using a Mobile Monoscopic Camera* in Proceedings of the International Conference on Recent Advances in 3-D Imaging and Modelling, Ottawa, Canada May 1997.

[20] N. M. Thalmann, D. Thalmann: *Rendez-vous a Montreal* http://miralabwww.unige.ch/Java-Less.html

[21] N. M. Thalmann, S. Carion, M. Courchesne et al.: *Virtual Clothes, Hair and Skin for Beautiful Top Models*, in Proc. Computer Graphics International, June 24-28, 1996, Pohang, Korea, IEEE Computer Society Press, June 1996, pp. 132–141.

[22] D. Pearson, *Developments in Model-Based Video Coding*, Proceedings of the IEEE, Vol. 83, No. 6, June 1995, pp. 892-906.

[23] M. Proesmans, L. Van Gool, A. Oosterlinck: *One-shot active 3d shape acquisition* in Proceedings 13th IAPR International conference on pattern recognition robotic systems, vol. III C, pp. 336–340, August 25–26, 1996, Vienna, Austria.

[24] T. Riegel, R. Manzotti, F. Pedersini, *3-D Shape Approximation of Objects in Multiview Image Sequences*, Int. Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging - IWSNHC3DI'97, Rhodes, Greece, 1997.

[25] D. Thalmann, J. Shen, E. Chauvineau: *Fast Realistic Human Body Deformation for Animation and VR Applications* in Proc. Computer Graphics International, Pohang, Korea, IEEE Computer Society Press, June 1996, pp. 166–174.

[26] Wingbermühle, Weik, Kopernik: *Highly Realistic Modeling of Persons for 3D Video conferencing Systems*, to appear in MMSP 97 Conference Proceedings