Automatic Creation of Flexible Antropomorphic Models for 3D Videoconferencing

S.Weik, J. Wingbermühle, W. Niem
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover - Germany
email: weik/wingber/niem@tnt.uni-hannover.de

Abstract

This contribution describes the automatic creation of highly realistic flexible 3D models of participants for distributed 3D videoconferencing systems. The proposed technique uses a flexible mesh template surrounding an interior skeleton structure, which is based on a simplified human skeleton. The vertices of this template are arranged in rigid rings along the bones of the skeleton. Using 3D data obtained by a shape from silhouette approach the size and shape of the mesh template is adapted to the real person. Texture mapping of the adapted mesh using real camera images leads to a natural impression. The mesh organization in rigid rings allows an efficient surface deformation according to the skeleton movements. Once the resulting model is transmitted, it can be animated subsequently using the simple parameter set of the interior skeleton structure. Results obtained with real image data confirm the eligibility of the animated person models in terms of realism and efficiency for 3D videoconferencing applications.

Index Terms - Antropomorphic Modeling, Flexible Deformation, Shape From Silhouette

1. Introduction

Current videoconferencing systems use 2D images to transmit and display visual information between different conference partners. As far as video is concerned such a system only enables the user to perform bilateral information exchange. The only way for more than two persons to take part in a 2D videoconference is to broadcast the image information to all other participants and get the image streams from every other conferee. This requests an enormous technical effort, mainly bandwith, and doesn't really give the impression of physically beeing close to each other. Another drawback is the fixed viewpoint of each participant's image stream which is determined by his local camera system and can not be changed arbitrarily at the receiver's site.

For future distributed videoconferencing systems it is therefore desirable not only to transmit 2D video, audio and shared applications. Rather one would wish to create a 3D virtual copy of a real 3D scene for each user, containing highly realistic models of the participants. These models should look like, act and behave similar to the actual conferees.

The virtual environment would be provided and managed by a central session server. Each participant of such a virtual conference would run a client application and could distribute his own model and the respective movement via the central server. The virtual environment and the models of the conferees are shared once in the beginning among the participants, the subsequent movements of the conferees are transmitted continously from the clients to the server. The server redistributes the actual motion to all clients providing the clients with the possibility to update their model of the complete scene.

Figure 1 depicts the main elements that are necessary for the transceiver of such a 3D virtual teleconferencing system. The main functionality is covered by a movable antropomorphic template. It consists of an interior skeleton structure that controls the movements of the model. The internal bone structure is covered by a triangular mesh describing the surface of the model. The scale and shape of this human model template is then adapted to 3D data of a real speaker person yielded by a shape from silhouette approach. In addition the triangles of the surface mesh can carry image information from the original camera images to increase the realism of the models. The adapted model is transmitted once in the beginning of the conference. In later stages only the motion parameters are estimated and transmitted for animation purposes. Several approaches for determination of the body movement can be found in [9, 12, 13].

This paper deals with the creation of such models and is organized as follows: Section 2 gives an overview of re-

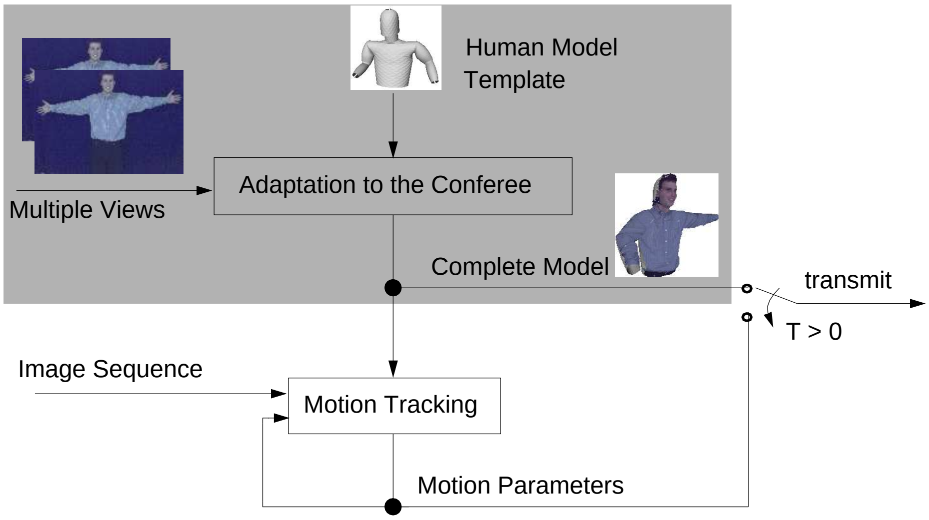

Figure 1. Overview of the transceiving part of the conferencing system

lated work in the same area, followed by some basics on *shape from silhouettes*, which is employed here to make the 3D scan of the persons. In section 4 the initial structure of the human body template, consisting of the internal skeleton and the covering triangle mesh, is described in detail. The adaptation of the antropomorphic template to the 3D measured data of an individual speaker person is discussed in section 5. The paper finishes with some experimental results we achieved with the proposed approach.

2. Related Work

The difficulty of dealing with the 3D modeling of complicated, non rigid objects like humans is evidenced by the extensive recent research on the topic. The application areas for antropomorphic modeling can roughly be divided into four main areas:

1. The creation of human models with a perfectly natural appearance for the movie industry
2. Simulations of human behaviour in ergonomic and medical studies
3. Human avatars and virtual agents
4. Telecommunication applications like 3D virtual videoconferencing systems

In the first field human models are used in high quality movie production. In most applications these virtual actors are mixed with images of a real scene taken from a camera. The goal here is to create sufficiently realistic human models to make the audience believe to see real world images[2][3][4]. Virtual actors can be positioned and moved freely in any real or virtual environment. The resulting scene description can be viewed from one or more arbitrarily chosen viewpoints. This helps to overcome certain restriction of the real physical world, e.g. dangerous stunts can be simulated without endangering actors or already deceased actors can live on in current movies.

In the second area antropomorphic models help to simulate medical and ergonomic tests for car seats, bath tubs, etc. Further Applications are the evaluation of simulated crash tests for the car industry. An example is the software system Jack[1], which has been developed from the University of Pennsylvania and is marketed by an US company.

The third kind of applications deals with the enhancement of man-machine communication. The objective is to make the means of interaction more resemblant to every day human conversation. Current approaches employ e.g. lifelike talking heads on a computer screen combined with the possibility of receiving input or transforming text to speech including lip and jaw animation[7]. Such an virtual agent could read messages or email to the user and in general could function as a person-like interface to the otherwise impersonal machine handling [5][6].

The last area of application for virtual human models stems from the telecommunication industry. The objective of present and future research is to develop technical systems which suggest the (virtual) presence of remote communication partners, which are situated in different places and who communicate by word, facial and body expressions. One important restriction for such a system is the real-time constraint which implies that the model should be simple and efficient though highly realistic. The objective

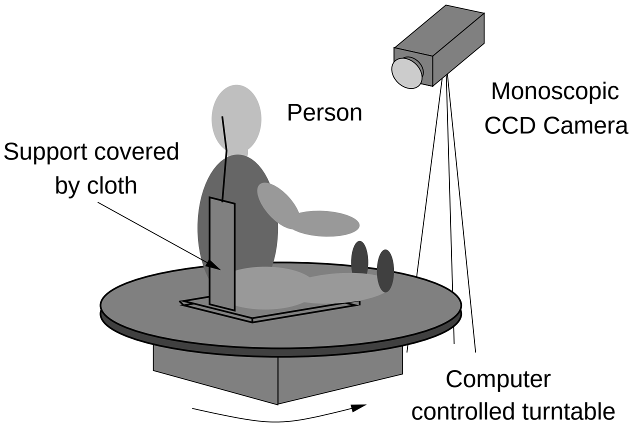

Figure 2. Measurement environment

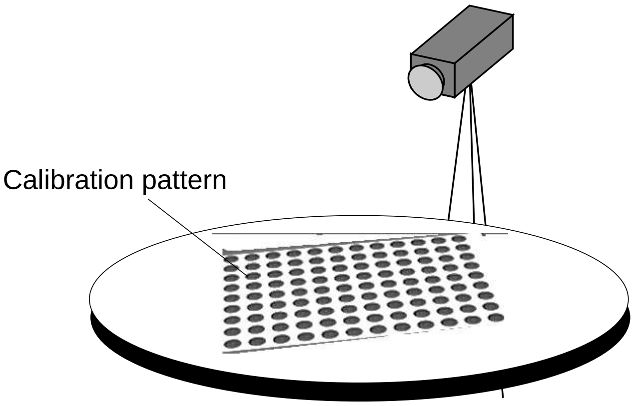

Figure 3. System calibration

is to render the models with up to several frames per second. Otherwise the element of communication would be lost. In particular the complex virtual actors described previously take large computational time for each frame, thus not meeting the realtime constraint. The approach presented in this contribution uses a simple flexible surface model which achieves both high performance and the desired quality. Examples for that kind of application can be found in [8][14].

3. Shape from silhouette

The "shape from silhouettes" or "method of occluding contours" approach is a well known technique for the automatic reconstruction of 3D objects from multiple camera views [16],[18],[17]. In the following section the measurement environment and the reconstruction technique are described in more detail.

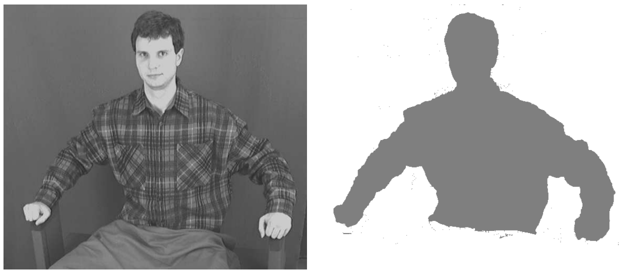

Figure 4. Input image and segmentation mask

3.1. Measurement environment and system calibration

The environment used for this work consists of a stationary CCD TV camera in front of a turntable with a stabilizing back-rest on top, which can be rotated in controled steps (figure 2). During image acquisition the back-rest is covered by cloth. The background is monochrome in order to facilitate the segmentation of the object against the background. Before the acquisition of an object on the turntable can be started, the geometric relation between the CCD camera and the turntable has to be computed, i.e. the system has to be calibrated. The calibration of this image acquisition system supplies the extrinsic parameters of the camera, which are the position and orientation of the camera, and the intrinsic parameters, which are focal length, 3rd order radial lens distortion and the shift of the optical center.

For that purpose, a precisely known test pattern consisting of black circles on a white background is placed on the turntable as shown in figure 3. The size and position of the test pattern are measured. The detection of the corresponding ellipses in the image taken by the CCD camera and the subsequent estimation of all parameters describing the acquisition process is performed using a calibration method e.g. as proposed by Tsai [19].

3.2. Volume reconstruction

The principle of the silhouette-based volumetric reconstruction can be divided into three steps. In the first step, the silhouette of the real object must be extracted from the input images as shown in figure 4. In controlled environments the segmentation of the real object against the background can be facilitated by using a monochrome background ("blue screen technique").

In the second step, a volumetric cone is constructed using the focal point of the camera and the silhouette as shown in figure 5a. The convex hull of the cone is formed by the lines of sight from the camera focal point through all con-

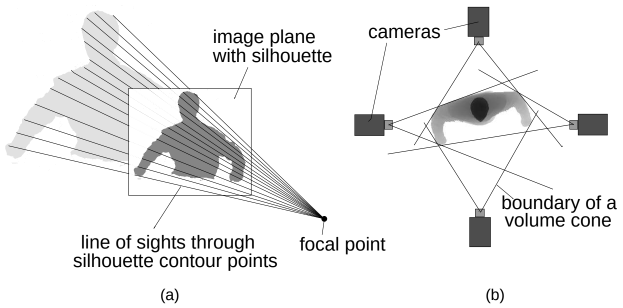

Figure 5. Volume reconstruction: (a) Construction of a volumetric cone, (b) Topview of the cone intersection

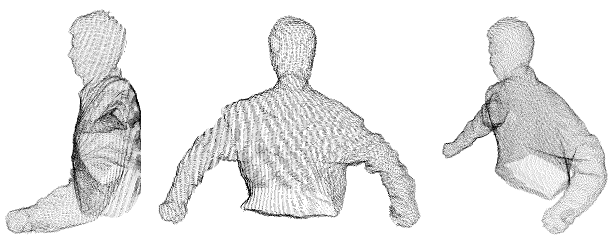

Figure 6. Rendered views of a point cloud representing the volume model surface

tour points of the object silhouette. For each view point such a volumetric cone is constructed, and each cone can be seen as a first approximation of the volume model.

In the last step, the volumetric cones from different view points are intersected in 3D and form the final approximation of the volume model. This is performed with the knowledge of the camera parameters, which give the information of the geometrical relation between the volumetric cones. In figure 5b a two dimensional top view of the intersection of the cones is shown. In figure 6 rendered views of a 3D point cloud representing the volume model surface are shown.

4. Antropomorphic Model

Basically our models consist of two predefined major components which can be easily adapted to an individual person. Firstly of a skeleton, that is a simplified version of the real human interior structure. Secondly of a flexible triangle surface mesh that is linked to the skeleton structure and it's movement. In the following section the initial model-template is described in more detail.

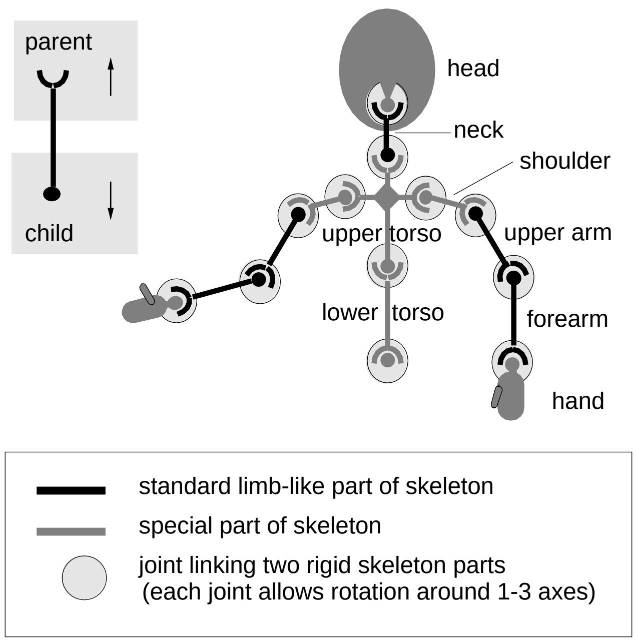

Figure 7. Model of human bone structure

4.1. Internal Skeleton Structure

The used model currently consists of a hierarchy of skeleton parts as shown in figure 7. The root of the hier-

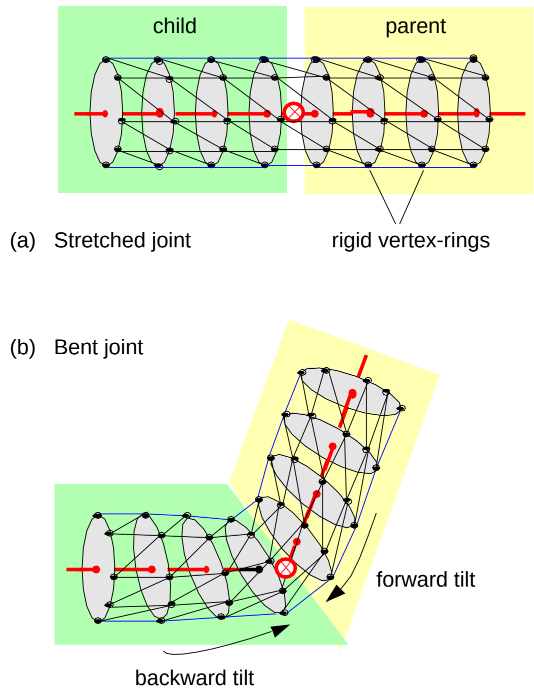

Figure 8. Ring-oriented deformation model

archy is situated in the lower torso. Each skeleton part is connected to it's parent by a joint with up to three degrees of freedom. A joint's current bending and twisting is represented by three rotation angles with respect to three predefined axes [10]. In order to avoid physically impossible positions minimum and maximum values are provided for each angle. The sole transmission of all joint angles enables the receiver to animate the model.

4.2. Triangle Surface Description

Together with the skeleton structure a surface mesh template is created. For each part of the skeleton the vertices of this initial triangle mesh are arranged to rigid rings which are lined up along the bone. Each vertex-ring has a relative position and orientation in it's respective part's local coordinate system. If the skeleton is bent or twisted at a certain joint the transformation is decreasingly propagated to the vertex-rings on both sides of the joint, i.e. the slices tilt on the bone as shown in figure 8. The propagation range varies for different joints and may even be different for the parent- and childside. Thus the appropriate deformation model can be chosen for different joints, e.g. head/neck or upper arm/forearm. In figure 9 the initial template consisting of the internal bone structure and surrounding triangle mesh is depicted.

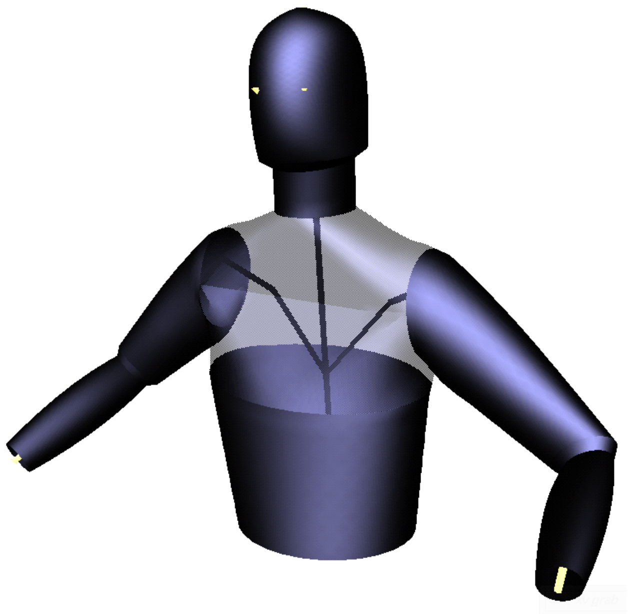

Figure 9. Initial template with semi-transparent upper torso showing the internal skeleton structure

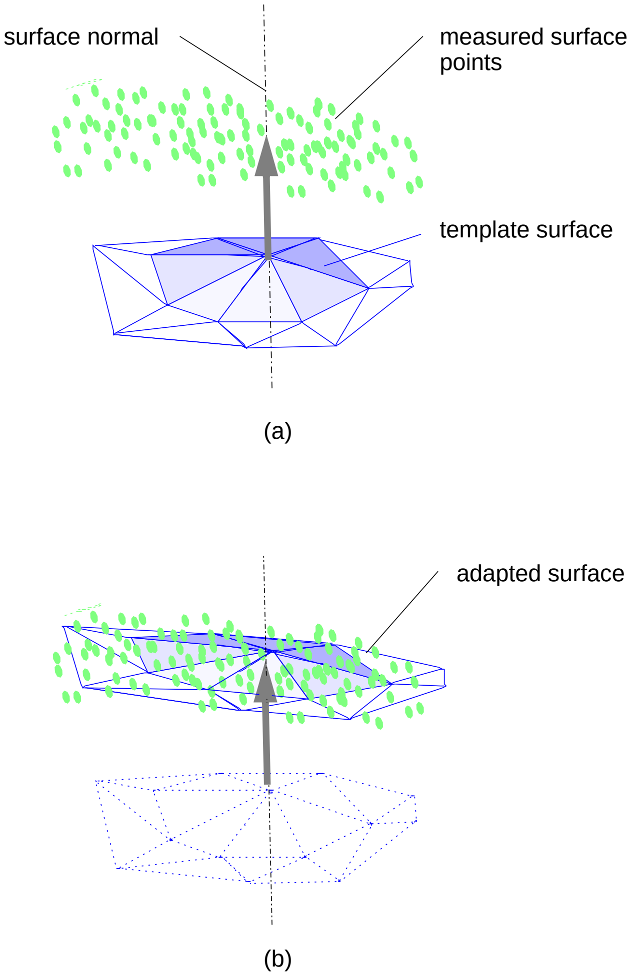

Figure 10. Shape adaptation

5. Model Adaptation

This section describes how the model-template is adapted to the 3D data of an individual conferee obtained by a shape from silhouette approach. The adaptation process can be divided into three steps:

- The skeleton's size and pose are fit to the observed conferee. This is currently done manually using a 3D tool to set the 3D joint positions. We have, however, some experimental results on automatic extraction and dynamic refinement of 3D joint positions for persons. These are discussed in detail in [15] but require the person to perform a sequence of certain gestures in front of a stereoscopic camera. We are currently investigating means of integrating these techniques into the *shape from silhouettes approach.*
- Knowing the 3D joint positions the individual skele-

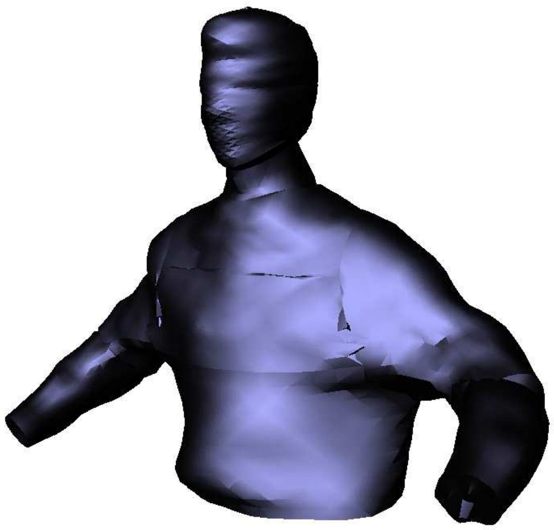

Figure 11. 360 degree shape adapted template

ton measures as well as the current joint angles can be calculated. The template can then be sized and moved according to the pose of the conferee. This allows for the surface of the template to be *inflated* to the conferee's shape.

- Finally the resulting model is textured from the same images used for the shape from silhoutte approach [11].

5.1. Surface Mesh adaptation

The previously created mesh template which has been sized and moved into the current pose of the conferee is inflated to the 3D data, obtained by the shape from silhouettes approach (see section 3). For each vertex of the mesh the surface normal is determined and the vertex is moved in the normal's direction until it fits to the measured data (fig. 10). In order to allow local expansion as well as local shrinkage both positive (outward) and negative (inward) shifts along the normal are possible. The direction of the surface normal at a vertex is approximated by the mean normal of its surrounding triangles. The deformation model still works properly after adapting the model because this operation does not affect the topological structure of the mesh, i.e. the adapted vertices are still organized in ring-form along the bones. Figure 11 depicts the inflated template which fits the measured 3D data determined with the shape from silhouettes approach.

6. Animation Results

We tried the approach for three different individuals. Each person was scanned using the shape from silhouette algorithm as described in section 3. In order to reduce occlusions special care had to be taken for the appropiate scanning position. Within the scanned geometry a skeleton structure has been positioned manually. Using this internal bone structure the initial template was created and automatically inflated to fit the real measured 3D data of the person.

The used antropomorphic template consists of approx. 5000 vertices connected by 10000 triangles. In areas where a higher resolution is desirable (e.g. the head) more triangles are spent. Since we use a shape from silhouette approach for surface scanning both 3D information and respective texture is available. Thus the inflated template can be textured by remapping the original image information onto the surface [11]. This leads to a highly realistic, flexible model that can be easily animated through a simple set of motion parameters. On an SGI Octane sytem the models can be animated with up to five frames per second. The upper part of figure 12 shows an animation of one of the authors following a synthetic sequence of motion parameters, whereas the lower part depicts a close wireframe view at the slice controled deformation model. In figure 13 the authors are gathered in a virtual conference room discussing their paper ...

7. Conclusions

An efficient and promising approach has been presented for anthropomorphic modeling. Employing a shape adaptation and texturing from real images highly realistic models of videoconferencing participants could be created. The 3D data of the real persons to be modeled is obtained using a shape from silhouette approach.

We proposed a simple ring-oriented deformation algorithm obeying an interior skeleton structure. Errors resulting from overlapping rings are sufficiently covered by the real texture. On the other hand the simple model enabled us to meet the real-time constraint for animation, with up to several frames per second. Real time performance is indispensible for communication applications.

The presented approach guides the way to a highly automatic kind of creating antropomorphic models as needed in 3D telecommunication applications where the timeconsuming manual creation of of a large number of different models is inappropiate. We introduced a chain of algorithms for the automatic adaptation of the initial template to different speaker persons. Experiments were carried out and we expect the approach to be equally suitable for both male and female persons. Currently the 3D joint positions were manually determined within the measured 3D data of a speaker.

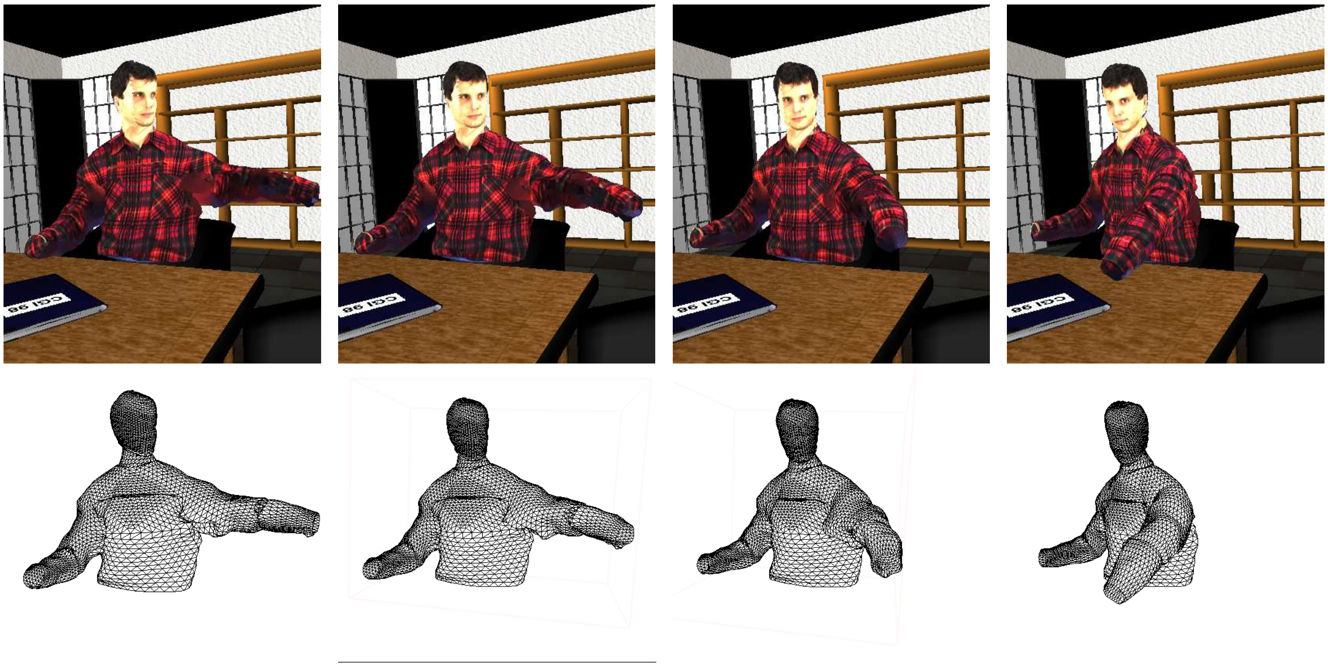

Figure 12. Synthetically animated sequence (textured and wireframe representation)

We were able to adapt the 3D shape and to texture the surface from real images automatically.

Future work aims at finding the 3D joint positions within the measured data without user interaction and automatic tracking of the conferee's motion over the image sequence. In addition a specialised head and hand template is desirable.

8. Acknowledgements

The authors would like to thank the Deutsche Telekom Berkom (especially Andreas Kopernik), without whose funding part of this work would have been impossible.

References

[1] N. Badler, C. B. Philips and B. L. Webber, "Simulating Humans", Oxford University Press, Oxford/UK, 1993.

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

[3] 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.

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

[5] L.S. Chen, T.S. Huang and J.Ostermann, "Animated Talking Head with Personalized 3D Head Model," Proc. 1997 IEEE 1st Workshop on Multimedia Signal Processing, Princeton NJ, June 1997, pp. 274–279.

[6] R. Boulic, "Human Avatar and Agent Animation for Virtual Reality", 50p, Computer Science Postgraduate Course in Virtual Reality and Animation, Module M4, EPFL DI-LIG, September 1996.

[7] K. Nagao and A. Takeuchi, "Speech Dialogue With Facial Displays: Multimodal Human-Computer Conversation", Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 102-109, 1994

[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] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," in CVGIP: Image Understanding, vol. 59, no. 1, pp. 94–115, Jan. 1994.

Figure 13. The authors discussing their CGI submission ...

[10] R. P. Paul, “Robot Manipulators: Mathematics, Programming and Control”, The MIT Press, Cambridge, Mass., 1981.

[11] W. Niem, H. Broszio, “Mapping Texture from Multiple Camera Views onto 3D-Object Models for Computer Animation,” in Proceedings of the International Workshop on Stereoscopic and Three Dimensional Imaging, Santorini, Greece, 1995.

[12] A. Tesei, G. L. Foresti and C. S. Regazzoni, “Human Body Modeling for People Localization and Tracking from Real Image Sequences,” Image Processing and it’s Applications, Conference Publication No. 410, July 1995, IEEE Press, 1995, pp. 806–809.

[13] D. M. Gavrila and L. S. Davis, “3D Model-Based Tracking of Humans in Action, A Multi View Approach,” Proc. IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recogn., pp. 73-80, 1996.

[14] J. Wingbermuehle, S. Weik and A. Kopernik, “Highly Realistic Modeling of Persons for 3D Videoconferencing Systems,” Proc. 1997 IEEE 1st Workshop on Multimedia Signal Processing, Princeton NJ, June 1997, pp. 286–291.

[15] J. Wingbermuehle, S. Weik, “Towards Automatic Creation of Realistic Antropomorphic Models for Realtime 3D Telecommunication”, to appear in: Journal of VLSI Signal Processing Systems, Special Issue on Multimedia Signal Processing, 1998.

[16] H. Busch, “Automatic Modelling of Rigid 3D Objects Using an Analysis by Synthesis System” , SPIE Proceedings Visual Communication and Image Processing IV, Philadelphia, 1989, pp. 356-364.

[17] W. Niem, “Robust and Fast Modelling of 3D Natural Objects from Multiple Views”, SPIE Proceedings Image and Video Processing II, Vol. 2182, 1994, pp. 388-397.

[18] R. Szeliski, “Rapid Octree Construction from Image Sequences”, Graphical Models and Image Processing: Image Understanding, Vol. 58, No 1, July, pp. 23-32, 1993.

[19] R. Y. Tsai, “A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses”, IEEE Journal of Robotics and Automation, Vol. RA-3. No.4, August 1987, pp. 323-344.