# 3D Motion Estimation for Articulated Human Templates Using a Sequence of Stereoscopic Image Pairs

Sebastian Weik and Oliver Niemeyer

Affiliation: University of Hannover
Institut fuer Theoretische Nachrichtentechnik und Informationsverarbeitung

## ABSTRACT

This contribution describes an approach towards 3D teleconferencing. Textured, 3D antropomorphic models are used in a virtual environment to give the impression of physical closeness. The requirements for such a conferencing system are on the one hand textured, articulated 3D models of the conferees. For high realism a flexible deformation model has been integrated in the 3D models. On the other hand these models have to be animated in the virtual meeting room according to the motion parameters of the real conferees. Therefore motion estimation has to be performed. To avoid wiring of the persons this has to be done optically. In this approach a gradient based motion tracker has been implemented. No markers or optical tracking points are needed to extract the hierarchic motion parameters of the conferee. It works on a stereoscopic image sequence and employs the flexible, articulated antropomorphic model of the conferee. The motion hierarchy of the articulated model is used to reduce the degrees of freedom and to make the estimation more robust.

**Keywords:** Flexible antropomorphic models, 3D motion estimation, 3D Teleconferencing, image sequence analysis

## 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 receivers 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. The upper picture of Figure 1 shows such a virtual conference room.

Due to the 3D character of the scene description each client is free to display the scene from an arbitrary viewpoint. Normally this should be the viewpoint of the respective participant to create an impression of physical presence at the conference site. If the receiving conferee moves this motion is beeing tracked and distributed via the central server to the other conference participants. On the other hand this information is used to render his view into the virtual scene description accordingly to his new position. This enables the conferee to freely walk within the scene and thus getting a realistic feeling of the real conference scene. Figure 1 depicts two viewpoints for two different conferees.

Further author information: (Send correspondence to Sebastian Weik)
Sebastian Weik: E-mail: weik@tnt.uni-hannover.de
Oliver Niemeyer: E-mail: niemeyer@tnt.uni-hannover.de

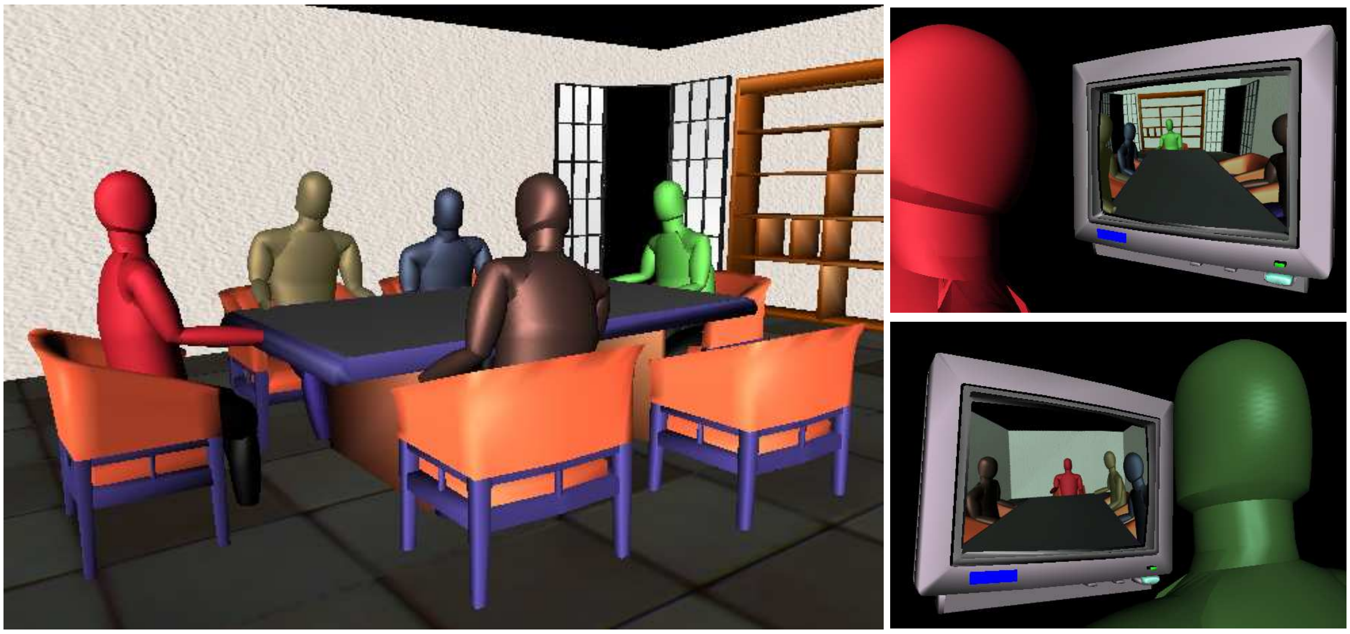

**Figure 1.** Different views of the same virtual cenference scene for different speakers

Figure 2 depicts the main elements that are necessary for the transceiver of such a 3D virtual teleconferencing system. The main functionality is covered by the movable 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 an stereoscopic 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.

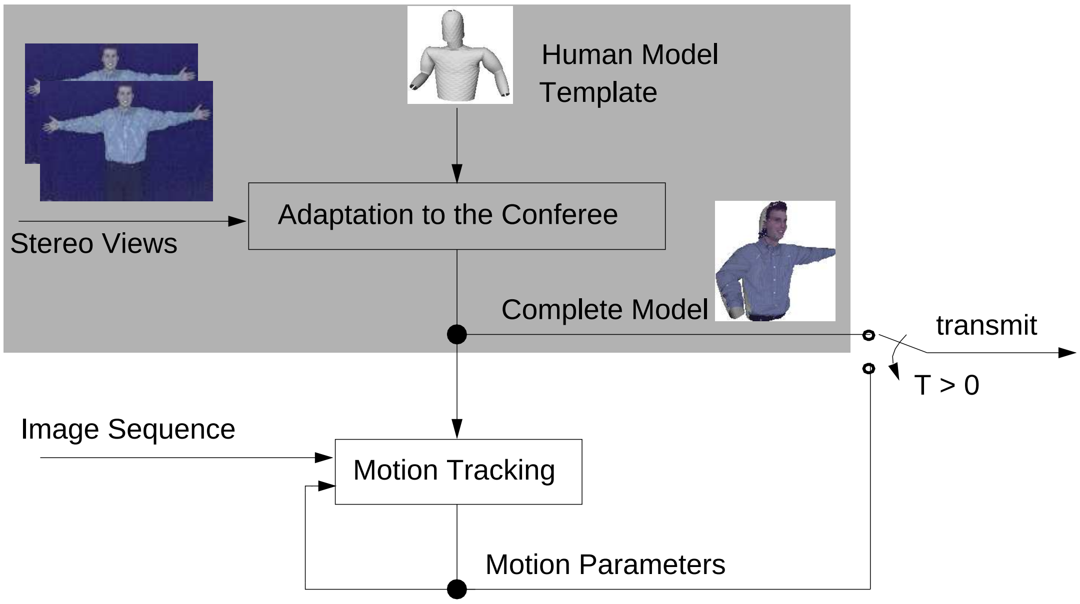


**Figure 2.** Overview of the transceiving part of the conferencing system

This paper mainly deals with the motion estimation of such models and is organized as follows: In section 2 the initial structure of the human body template, consisting of the internal skeleton and the covering triangle mesh together with its adaptation to an individual conferee is shortly described. The motion estimation from a stereoscopic image sequence based on a gradient approach is discussed in section 3. The paper finishes with some experimental results we achieved with the proposed approach.

## 2. ARTICULATED HUMAN BODY TEMPLATE

As indicated in Figure 2 the 3D videoconferencing tool requires two main parts: the generation of anthropomorphic models and the determination of motion parameters from video sequences. The following sections roughly describe the generation of those models. More details can be found in.[6]

### 2.1. Initial human template

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 simplifed version of the real human interior structure. Secondly of a flexible trianglesurface mesh that is linked to the skeleton structure and its movement.

#### 2.1.1. Internal skeleton structure

The used model currently consists of a hierarchy of skeleton parts as shown in Figure 3 (left). The root of the hierarchy is situated in the lower torso. Each skeleton part is connected to its 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. 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.

#### 2.1.2. 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 its 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 3 (right). The propagation range varies for different joints and may even

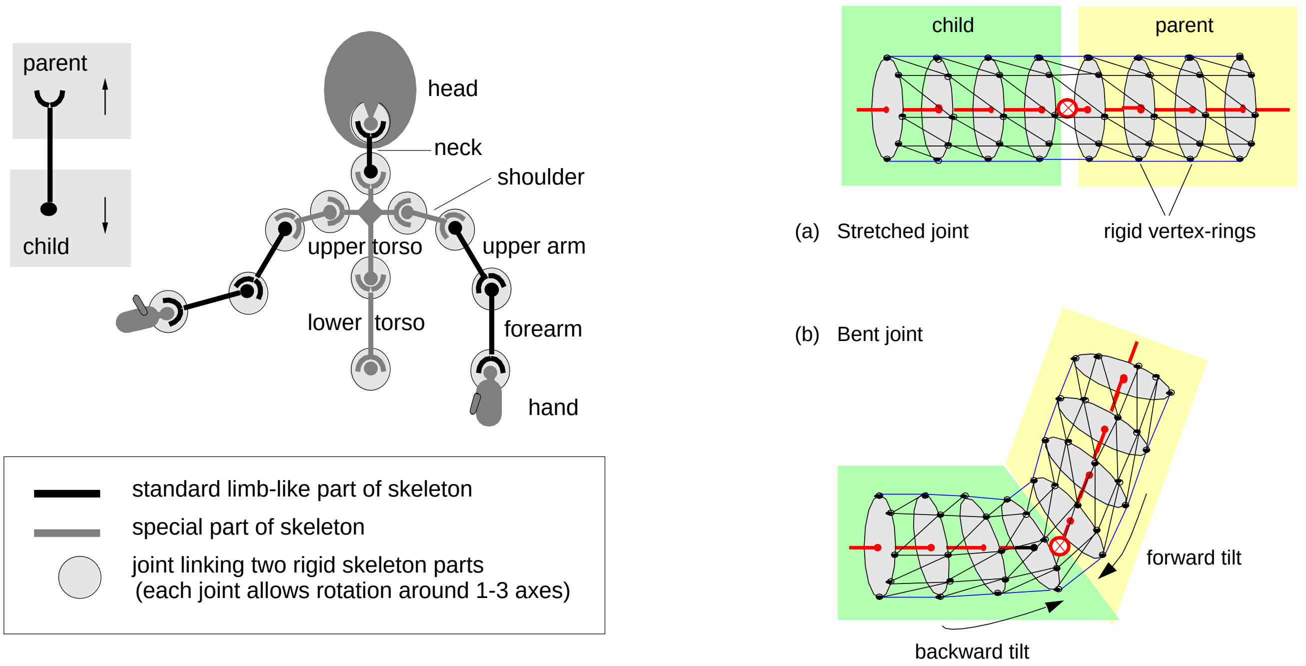


**Figure 3.** Model of human bone structure (left), ring-oriented deformation model (right)

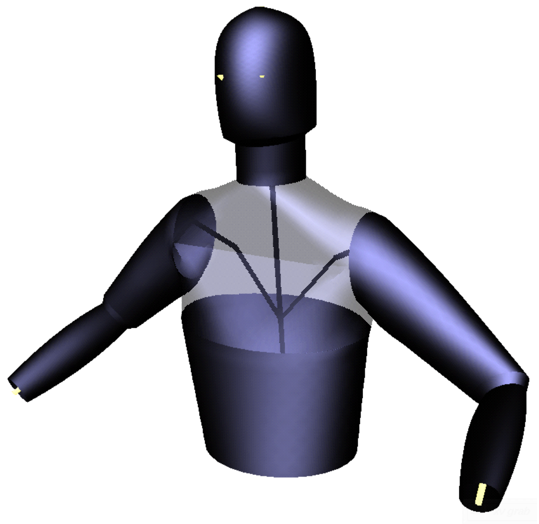

**Figure 4.** Initial template with underlying bone structure

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 4 the initial template consisting of the internal bone structure and surrounding triangle mesh is depicted.

### 2.2. Template adaptation to measured 3D data

This section describes how the model-template is adapted to the 3D data of an individual conferee obtained by a stereoscopic 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[6] but require the person to perform a sequence of certain gestures in front of a stereoscopic camera.
- Knowing the 3D joint positions the individual skeleton 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. The 3D surface description in this experiment

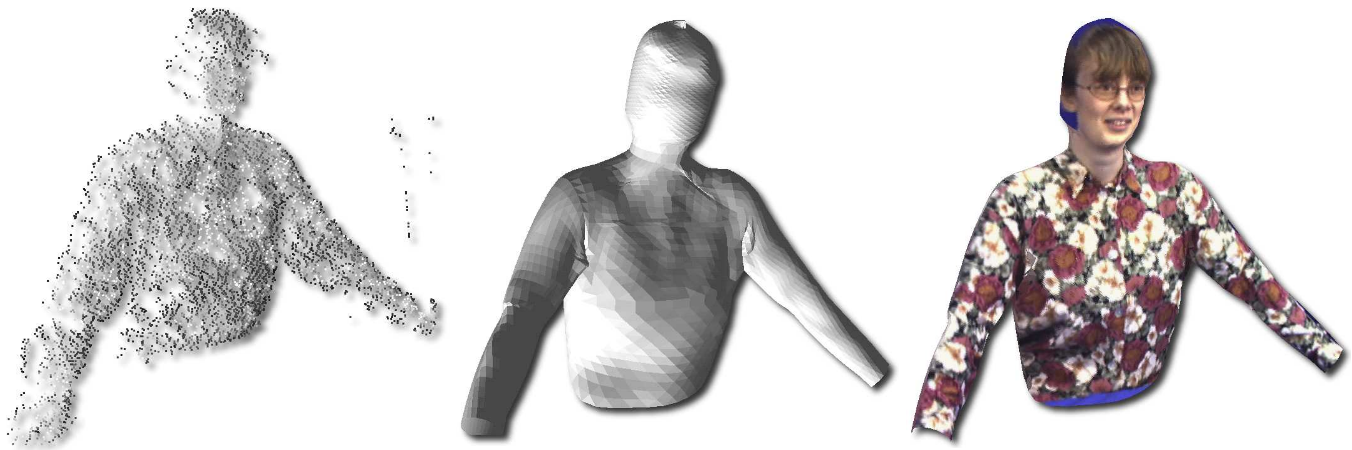

**Figure 5.** Point cloud from stereo (left), shape adapted model (middle), textured model (right)

has been extracted using a stereoscipoc approach from a frontal view only. However, the proposed algorithm works as well with 360 degree data sets e.g. from a laser scanner or shape from silhouette.[7]

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 3D point cloud (Figure 5 (left)). 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 5 (middle) depicts the inflated template which fits the measured 3D data determined with the stereoscopic approach.

- Finally the resulting model is textured using the original camera images giving the model a realistic appearance[2] (Figure 5 (right)).

## 3. HIERARCHICAL 3D MOTION ESTIMATION

In standard applications for 3D motion estimation 6 motion parameters for each independently moving object have to be estimated. On the other hand independent 6-degree-of-freedom motion for each part is not desirable in applications where human body motion is estimated. Inaccuracies and measurement errors would lead to a kind of "disintegration" of the body structure over long video sequences.

In the case of an articulated object like the human template as described above, which is a tree-like structure of rigid objects connected by joints, only up to three degrees of freedom have to be estimated for each body part. In this contribution we suggest a skeleton structure for a typical videokonferencing situatiuon (a person sitting behind a table) consisting of 12 body components. Starting from the root of the skeleton, the lower torso, motion is estimated hierarchically down to the "leafs", which are the head and the hands of the model.

To estimate 3D motion from videosequences two main directions can be found in literature: Methods exploiting feature correspondences on the on hand or using an optical flow approach on the other hand. Most other publications focus on the first method. Gavrila, for example, proposes to match the projections of a model of the articulated object with the real n-view image sequence.[4] Similar methods are described in[8] and in.[9] The method we are proposing is based on the second approach. In the next section motion estimation for rigid objects using a gradient based approach is described, whereas section 3.2 focuses on the implemented specializations for articulated, hierarchical objects.

### 3.1. GRADIENT BASED MOTION ESTIMATION FOR RIGID OBJECTS

If we assume constant illumination and diffuse reflecting object surfaces, the changes of the intensity $I(x,y,t)$ in the image sequence only depend on the movement of the projected objects.[10] This leads to the equation of the continuity of the optic flow (with the velocity vector $\vec{u} = [\frac{dx}{dt} \quad \frac{dy}{dt}]$)

$$dI(x,y,t) \quad = \quad \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt \quad = \quad \frac{\partial I}{\partial t} + \vec{u}\cdot\nabla I \quad = \quad 0. \tag{1}$$

In order to iteratively solve this nonlinear equation it is approximated with the linear term of a Taylor series. In addition we state, that the images of the sequence are taken periodically ($\triangle t$) replacing the continuous value $t$ by the discrete value $t_n$. The resulting linear equation reads as follows

$$\frac{\triangle I}{\triangle t} + \begin{bmatrix} \frac{\triangle x}{\triangle t} & \frac{\triangle y}{\triangle t} \end{bmatrix} \cdot \vec{G} \quad = \quad 0, \tag{2}$$

where $\triangle I$ is the intensity difference of a pixel in subsequent images at time instances $t_n$ and $t_{n+1}$.

The vector $\vec{G}$ is the intensity image gradient. After eleminating $\triangle t$ and replacing the picture coordinates $x$ and $y$ by the image vector $\vec{B}$ we achieve the basis for the gradient based motion estimation,

$$\triangle I + (\vec{B}' - \vec{B})\cdot\vec{G} \quad = \quad 0 \tag{3}$$

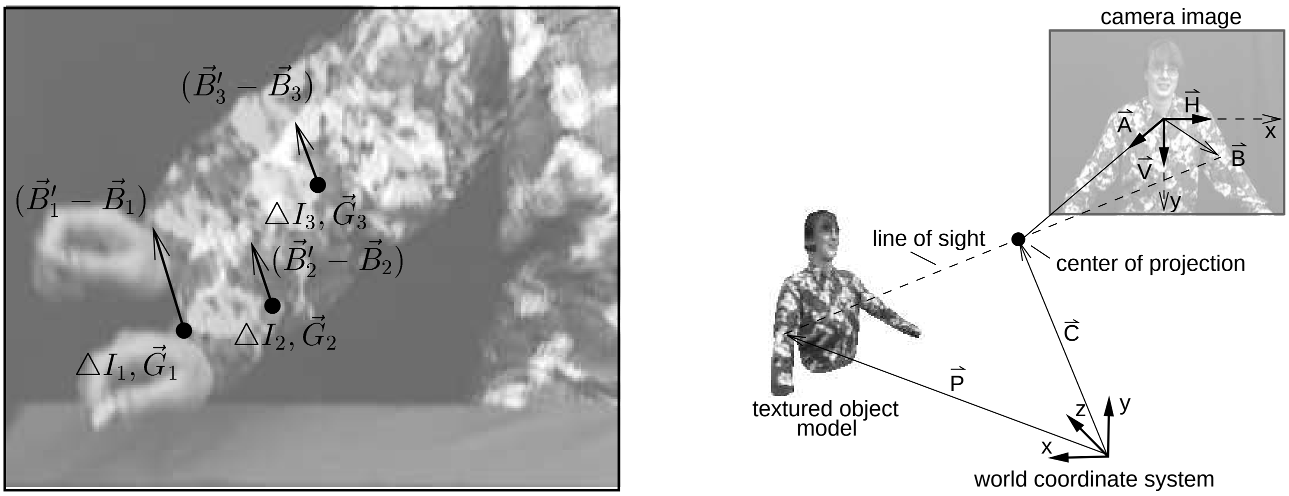


**Figure 6.** Gradient based motion estimation (left), the CAHV camera model (right)

with $\vec{B}'$ the image vector at the time instance $t_{n+1}$ and $\vec{B}$ the image vector at the time $t_n$ of the same projected 3D point. Figure 6 (left) illustrates this equation.

The image vectors in equation 3 can be replaced by their corresponding 3D points in the real world $\vec{P}$ and $\vec{P}'$ respectively. Therefor a camera model and a projection equation is necessary. We use the CAHV-camera model (pinhole camera model),[11] with $\vec{C}$ describing the center of projection. The vector $\vec{A}$ determines the viewing direction of the camera whereas $\vec{H}$ and $\vec{V}$ span the actual image plane. The geometry of the camera model is illustrated in figure 6 on the right side.

The projection equations for this camera model are

$$B_x = \frac{(\vec{P} - \vec{C}) \cdot \vec{H}}{(\vec{P} - \vec{C}) \cdot \vec{A}} \tag{4}$$

$$B_y = \frac{(\vec{P} - \vec{C}) \cdot \vec{V}}{(\vec{P} - \vec{C}) \cdot \vec{A}} \tag{5}$$

Replacing the image vectors $\vec{B}'$ and $\vec{B}$ in equation 3 leads to the following expression

$$\begin{aligned} \triangle I \cdot (\vec{P} - \vec{C}) \cdot \vec{A} &= (\vec{P}' - \vec{C}) \cdot \left( G_x \vec{H} + G_y \vec{V} - \frac{(\vec{P} - \vec{C}) \cdot (G_x \vec{H} + G_y \vec{V})}{(\vec{P} - \vec{C}) \cdot \vec{A}} \cdot \vec{A} \right) \\ &\quad - \triangle I \cdot \vec{A} \cdot (\vec{P}' - \vec{P}). \end{aligned} \tag{6}$$

In order to estimate the six motion parameters of the object (three rotational and three translational degrees of freedom) the motion $(\vec{P} - \vec{P}')$ has to be expressed using a motion model (eq. 10). To transform equation 6 into a suitable form an additional constraint, $\triangle I = 0$ if the points $\vec{P}'$ and $\vec{P}$ are equal, is necessary. The insertion in equation 6 results in 7, which in turn can be inserted into equation 6.

$$\begin{aligned} \vec{C} \cdot \left( G_x \vec{H} + G_y \vec{V} - \frac{(\vec{P} - \vec{C}) \cdot (G_x \vec{H} + G_y \vec{V})}{(\vec{P} - \vec{C}) \cdot \vec{A}} \cdot \vec{A} \right) &= \\ \vec{P} \cdot \left( G_x \vec{H} + G_y \vec{V} - \frac{(\vec{P} - \vec{C}) \cdot (G_x \vec{H} + G_y \vec{V})}{(\vec{P} - \vec{C}) \cdot \vec{A}} \cdot \vec{A} \right) & \end{aligned} \tag{7}$$

This leads to the base equation in which the motion $(\vec{P} - \vec{P}')$ is separated.

$$\triangle I = \vec{K} \cdot (\vec{P}' - \vec{P}) \tag{8}$$

with

$$\vec{K} = \frac{G_x \vec{H} + G_y \vec{V} - \left( \frac{(\vec{P}-\vec{C})\cdot(G_x\vec{H}+G_y\vec{V})}{(\vec{P}-\vec{C})\cdot\vec{A}} + \triangle I \right) \cdot \vec{A}}{(\vec{P} - \vec{C}) \cdot \vec{A}} \tag{9}$$

The motion $(\vec{P} - \vec{P}')$ is then substituted by a rigid motion model with three rotational and three translational degrees of freedom:

$$\vec{P}' = \underline{R} \cdot (\vec{P} - \vec{S}) + \vec{T} + \vec{S} \tag{10}$$

Equation 10 describes the motion of $\vec{P}$ to $\vec{P}'$ with the translation $\vec{T}$ and the linearized rotation matrix

$$\underline{R} = \begin{bmatrix} 1 & -R_z & R_y \\ R_z & 1 & -R_x \\ -R_y & R_x & 1 \end{bmatrix} \tag{11}$$

describing a rotation around the center of rotation $\vec{S}$. The equation is used to replace $\vec{P}'$ in equation 8. The resulting vector equation reads as follows

$$\triangle I = \vec{L} \cdot \vec{\phi} \tag{12}$$

with

$$\vec{L} = [R_x \;\; R_y \;\; R_z \;\; T_x \;\; T_y \;\; T_z] \tag{13}$$

$$\tag{14}$$

the desired motion parameters and

$$\vec{\phi} = \begin{bmatrix} K_z \cdot (P_y - S_y) - K_y \cdot (P_z - S_z) \\ K_x \cdot (P_z - S_z) - K_z \cdot (P_x - S_x) \\ K_y \cdot (P_x - S_x) - K_x \cdot (P_y - S_y) \\ K_x \\ K_y \\ K_z \end{bmatrix} \tag{15}$$

containing the measured parameters. For a number of selected points $\vec{P}$ (the criterion is a high image gradient) the equation 12 is set up which leads to an overdetermined linear equation system that can be solved using a least squares approach

$$\underline{\phi}^T \cdot \underline{\phi} \cdot \vec{L} = \underline{\phi}^T \cdot \vec{\triangle I} \tag{16}$$

with

$$\underline{\phi}^T = \begin{bmatrix} \vec{\phi}_1^T & \vec{\phi}_2^T & \dots & \vec{\phi}_n^T \end{bmatrix}^T \tag{17}$$

and

$$\vec{\triangle I} = [\triangle I_1 \quad \triangle I_2 \quad \dots \quad \triangle I_n]^T . \tag{18}$$

Since we want to estimate real 3D motion using imprecise models the exploitation of only one camera showed insufficient results. To increase the accuracy we therefore integrated additional functionality to be able to use a calibrated stereoscopic camera. Each point $\vec{P}$ of the virtual object model is projected in both image planes yielding an expression according to equation 12 for both, the left and the right image. Using the same 3D model the solution of the equation system, the actual motion of the body parts, simultaneously fulfills the constraints imposed by the left and right video sequence.

Due to the linearizations of the image signal (eq. 2) and the rotation matrtix (eq. 11) the final motion parameters have to be determined in an iterative process. More details on this analysis through synthetis approach can be found in Koch.[1] The image information of the images at the time instance $t_n$ is projected onto the model. The part of the model for which the motion is to be estimated is then iteratively moved until the synthetically rendered image best resembles the images at the time instance $t_{n+1}$. The criterion used is the sum of squared intensity differences of the rendered image and the image at the time instance $t_{n+1}$.

### 3.2. GRADIENT BASED MOTION ESTIMATION FOR ARTICULATED OBJECTS

As described each part of the articulated object is treated separately. Except the lower torso (the root of our hierachical model template), which is able to perform all 6 degrees of freedom in its motion, all other parts of the object are only able to perform a rotational motion relatively to their hierachical higher neighbour. Therefore the motion model has been adapted to these demands. In the case of an articulated object the motion between subsequent time instances $\triangle t$ can be desrcibed as follows:

$$\vec{P}' = \underline{R} \cdot (\vec{P} - \vec{S}) + \vec{S}. \tag{19}$$

Equation 19 is inserted in eq. 8 instead of eq. 10. This again leads to a vector equation similar to equation 12. The further practice to achieve the motion parameters is the same as formerly described.

Nevertheless, the image area covered by small body parts like shoulders or the neck is too small to be estimated directly. Therefore we added special means to estimate such bodyparts jointly with their hierarchical child (e.g. neck / head). Once more we took the 6 degree of freedom motion estimation as basis. But now the translational parameters are not set to zero, we rather replaced the translation $\vec{T}$ (see eq. 10) by a function of the rotational parameters of the hierachically higher, small body part. Figure 7 illustrates this approach.

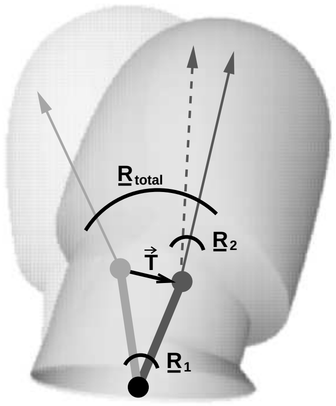


**Figure 7.** Combined motion estimation of small body parts

The resulting motion equation reads as follows:

$$\vec{P}' = \underline{R}_{total} \cdot (\vec{P} - \vec{S}) + \vec{S} + f(\underline{R}_1). \tag{20}$$

This forces the translation vector $\vec{T} = f(\underline{R}_1)$ to fulfill the constraint of constant bone-length between the two joints. The further processing to achieve the motion parameters from eq. 20 is the same as with the formerly described motion models. In addition the rotation of the hierarchical child ($\underline{R}_2$) has to be calculated from the total rotation of both body parts ($\underline{R}_{total}$) and the separate rotation of the hierarchically higher body part ($\underline{R}_1$) to extract the motion for both parts. Further details on this exceed the scope of this contribution.

## 4. RESULTS

We tested the described approach both with synthetic and real data. In the case of synthetic data the motion estimation performed without noticable errors.

The test on real data employed a stereoscopic videosequence containing 180 frames, i.e. of approx. 7 seconds duration. The image size was 768x576. A person sitting behind a table is observed by a calibrated stereoscopic camera. During the sequence the person first looks to the left, to the right and then raises its hand (Figure 8). Starting from a known position of the 3D model for the first frame, the algorithm was able to successfully track the person's motion over the full sequence. The estimated motion parameters have been used to animate the model, which consists of approx. 10000 triangles and is currently only adapted and textured from a frontal viewpoint. The model animation can be done with up to five frames per second on an SGI Indigo2 while the analysis takes up to two minutes for each frame. Future work therefore aims at reducing the computational time since realtime is a crucial prerequisit for the desired 3D videoconferencing application. In addition it is desirable to Kalman-filter the resulting motion parameters to reduce noise artefacts.

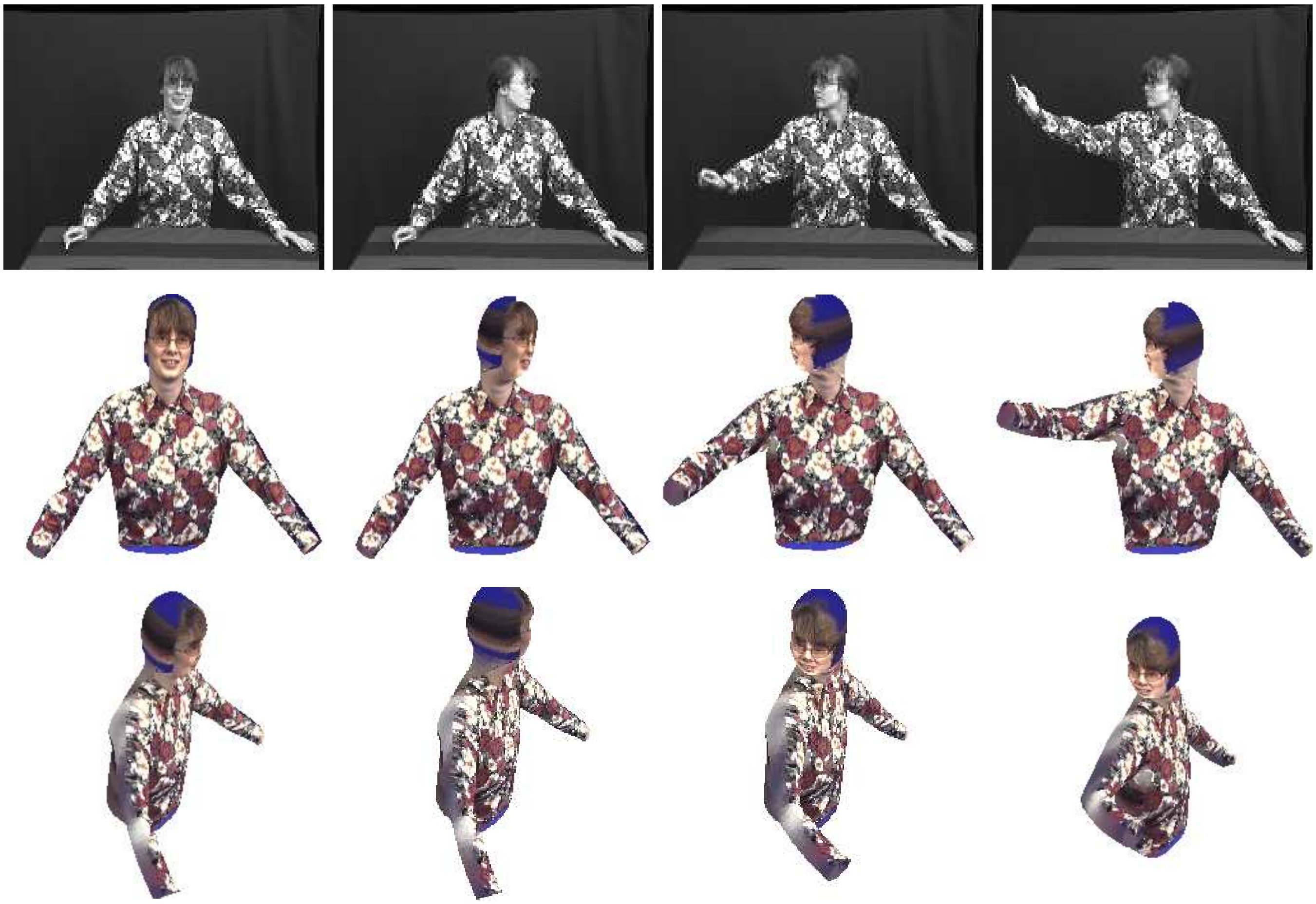

**Figure 8.** Real image sequence and textured 3D model animated according to the estimated motion parameters (front and side view for frames 0, 60 140 and 180)

## REFERENCES


1. R. Koch: "Dynamic 3-D scene analysis using synthesis feedback control", IEEE PAMI, Vol. 15(6), pp. 556-68, June 1993
2. 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.
3. 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.
4. 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.
5. 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.
6. J. Wingbermuehle, S. Weik, "Towards Automatic Creation of Realistic Antropomorphic Models for Realtime 3D Telecommunication", Journal of VLSI Signal Processing Systems, Special Issue on Multimedia Signal Processing, 1998.
7. S.Weik, J.Wingbermuehle, W. Niem, "Automatic Creation of Flexible Antropomorphic Models for 3D Videoconferncing", Proc. IEEE Comput. Soc. Conf. Computer Graphics International (CGI), pp. 520-527, Hannover, 1998
8. K.Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences", CVGIP: Image Understanding, Vol. 59, No. 1, January, pp. 94-115, 1994
9. F. J. Perales, J. Torres, "A System for Human Motion Matching between Synthetic and Real Images Based on a Biomechanic Model", Proc. of the Workshop on Motion of non-rigid and articulated Objects, IEEE, 1994
10. B. Jhne, "Digitale Bildverarbeitung", Springer, Hamburg, 1993
11. 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, Aug. 1987, pp 323-344