Hierarchical 3D Pose Estimation for Articulated Human Body Models from a Sequence of Volume Data

Sebastian Weik and C.-E. Liedtke

Institut für Theoretische Nachrichtentechnik und
Informationsverarbeitung, University of Hanover, Germany
weik/liedtke@tnt.uni-hannover.de

Abstract. This contribution describes a camera-based approach to fully automatically extract the 3D motion parameters of persons using a model based strategy. In a first step a 3D body model of the person to be tracked is constructed automatically using a calibrated setup of sixteen digital cameras and a monochromatic background. From the silhouette images the 3D shape of the person is determined using the shape-from-silhouette approach. This model is segmented into rigid body parts and a dynamic skeleton structure is fit. In the second step the resulting movable, personalized body template is exploited to estimate the 3D motion parameters of the person in arbitrary poses. Using the same camera setup and the shape-from-silhouette approach a sequence of volume data is captured to which the movable body template is fit. Using a modified ICP algorithm the fitting is performed in a hierarchical manner along the the kinematic chains of the body model. The resulting sequence of motion parameters for the articulated body model can be used for gesture recognition, control of virtual characters or robot manipulators.

1 Introduction

In recent time emphasis has been put on the extraction of human body shape and motion parameters from videosequences. Application areas appear in the TV and film production where virtual actors have to be taught to exhibit human behaviour like human facial expressions and human gestures. Another area of application is the control of remote systems from the passive observation of body motions. Examples are remote control of avatars in multi-player games or the remote control of robots which may act in hazardeous and dangerous environments. The creation of those models consists mainly of two parts: firstly the extraction of the shape and texture of the real person and secondly the automatic adaptation and fitting of an interior skeleton structure to extract motion.

In this paper an approach for motion estimation using a hierarchic ICP algorithm is presented. This is illustrated in Fig. 1. From a real person an initial

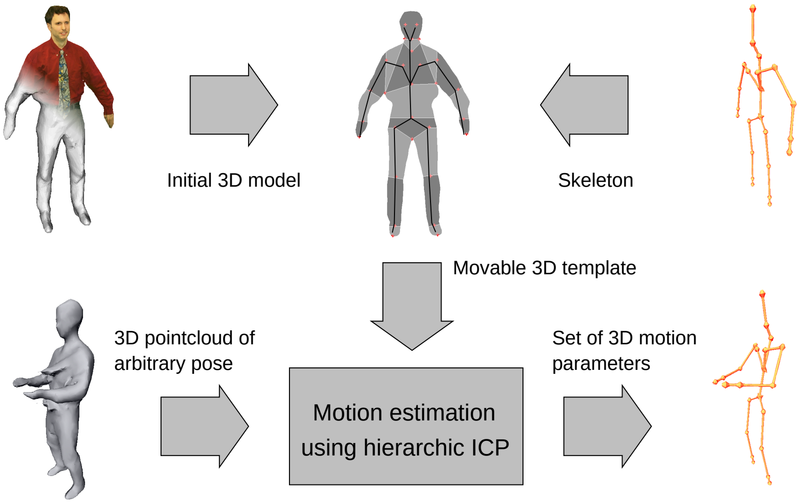
Initial 3D model
Skeleton
Movable 3D template
3D pointcloud of arbitrary pose
Motion estimation using hierarchic ICP
Set of 3D motion parameters

Fig. 1. System Overview for model based 3D estimation of arbitrary human poses

3D surface model is obtained. The motion of the person and its model can be described by the motion of a skeleton, which is fitted to and connected to the model surface. When an arbitrary pose of the same person represented again as surface model is obtained the parameters of this pose can be determined by a model based approach. As model serves the movable 3D template which has been obtained in the first step. The result of the analysis is a set of 3D motion parameters, which describes for a sequence of poses the motion of the gesture.

2 Body Modeling

As shown in Fig.1 the model based motion estimation requires two steps: the creation of a 3D segmented movable model of the person and the fitting of that model to a 3D measurement of the same person in an arbitrary pose. The first step is performed with a camera based passive full body scanner.

2.1 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 [4]. In this section the reconstruction technique is described briefly.

To capture the body model and to extract the sequence of volume data for motion estimation a special setup of sixteen digital cameras has been constructed which is suitable for using the shape-from-silhouette approach (Fig. 2). The person is situated in front of a monochromatic coated background which is used later on for silhouette extraction. The combination of background and camera

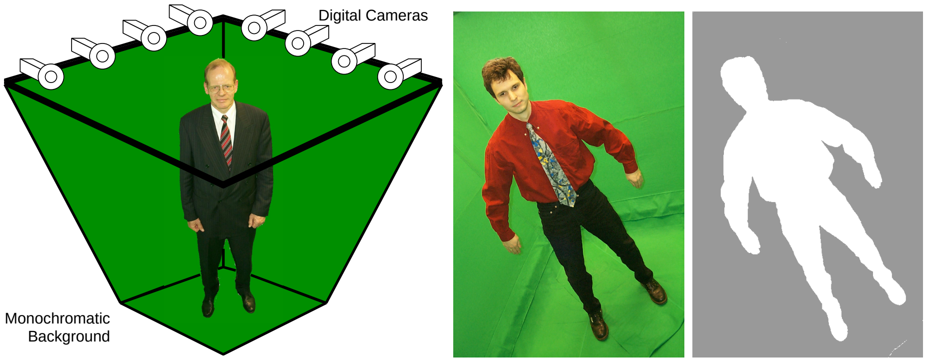

Fig. 2. Principal measurement setup (left) and input image and segmented foreground (center and right)

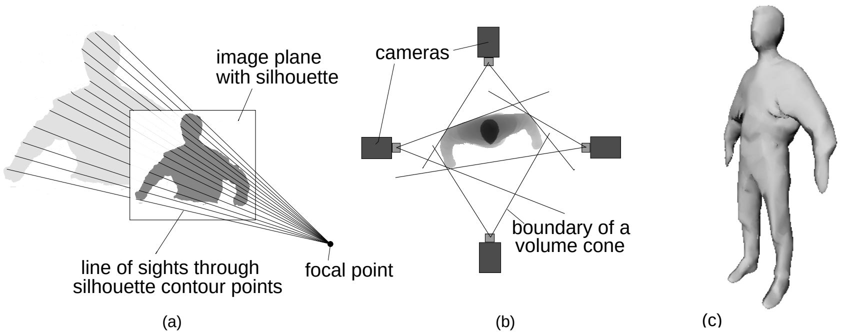

Fig. 3. Volume reconstruction: (a) Construction of a volumetric cone, (b) Top-view of the cone intersection, (c) 3D modeling result

positions need to fulfil mainly two important constraints: firstly, *all* cameras must see the complete person in front of the monochromatic background and secondly no camera should be visible from any other camera.

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 Fig. 2. In the proposed environment the segmentation of the person against the background is facilitated by using the monochromatic 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 Fig. 3a. The convex hull of the cone is formed by the lines of sight from the camera focal point through all contour 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 Fig. 3b a two dimensional top view of the intersection of the cones is shown. In Fig. 3c a triangulated 3D point cloud representing the volume model surface is shown.

After the reconstruction of the geometry the model can automatically be textured using the original camera images giving a highly realistic impression [1].

2.2 Skeleton fitting

To extract the 3D motion parameters of a moving person a template based approach has been used. Therefore an internal skeleton structure is needed which controls the model movements. In order to find the correct set of motion parameters the body model has to be adapted to the specific person that is to be tracked later. Normally this requires a tedious manual positioning of the joint positions within the model. In order to reduce the costs of model creation it is desirable to automate this process. As opposed to other algorithms that use the thinning of 3D data[2][3] we propose to find the skeleton as a multistep process based on re-projected images of the voxel model of the person.

In a first step a principal axis analysis is performed to transform the model into a defined position and orientation. Using a virtual camera – not to be confound with one of the real cameras – a synthetic silhouette from a frontal viewpoint is calculated. The outer contour of this image is used to extract certain feature points like the bounding box, the position of the neck, the hands and so on as can be seen in Fig. 4 on the left. In the last step the 2D joint positions of the desired skeleton are derived directly from the detected feature points using certain ratios (Fig. 4, left). Using the real model and the virtual camera these 2D joint positions are extended to their real 3D positions. In Fig. 4 on the right the skeleton has been used to segment the model into different body parts.

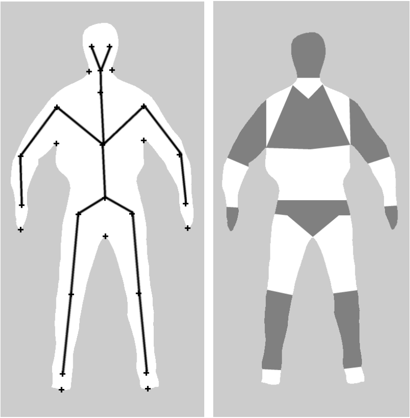

Fig. 4. Extracted features and calculated skeleton (left) - Automatic segmentation in body parts (right)

The resulting reference model describes the relation between the elements of the skeleton and the surface points of the model and it contains parameters like the bone length which are assumed to remain constant during the following motion analysis. The reference model has been derived from a special pose, which exhibits most distinctly the elements of the model skeleton, like the neck, head,

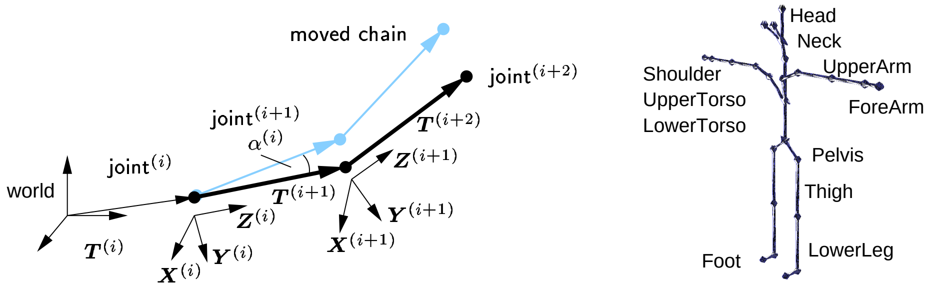

Fig. 5. The used coordinate systems within the kinematic chain(left) and the skeleton structure in neutral position (right)

the torso, the limbs and their parts. During pose analysis this reference model serves as a movable 3D template in order to estimate the free parameters of the internal skeleton from the observation of the recorded surface points of the particular pose under investigation.

3 3D motion analysis

From the body scanner mentioned above a cloud of 3D surface points is obtained for each pose from a real person. The task of the motion analysis is to estimate for each pose the free parameters of the underlying skeleton, i.e. the location of the skeleton and the angular positions. An overview in the area of visual analysis of human movement can be found in [7].

The known approaches for this motion estimation problem can be divided into two different types. The first kind, which is often based on optical flow, tries to register the differential motion of an object between subsequent frames like for instance in [8]. This approach lacks the possibility to find an appropriate motion for sequences of arbitrary length because estimation errors from frame to frame add up until the tracking is lost. The second approach which is proposed here normally requires some kind of a (3D) model. It dispenses with the information yielded from prior processing stages and thus avoids the mentioned problem. Because of the larger movements between the initial pose of the model and the pose to be estimated finding the motion is more difficult. The proposed approach tries to eliminate manifolds in the solution by exploiting the motion hierarchy of the model.

3.1 Skeleton Hierarchy

The internal skeleton structure as shown in Fig. 5 on the right is organized in form of a kinematic chain as shown in Fig. 5 on the left. Each body part is described as a bone of a certain length and is connected to a parent and child part respectively through a joint. Each joint is equipped with a set of two local coordinate systems. The first gives the transformation of the parent to the child in its neutral position and the second coordinate system describes the actual movement around a joint. This approach has been chosen to be able to control maximum rotation angles around the axes of the fixed coordinate system. In addition the fixed local coordinate systems are oriented within the skeleton such that the $\boldsymbol{Z}$-axis always runs along the longitudinal orientation of the body part and the $\boldsymbol{X}$-axis is oriented along the viewing direction of the person. The $\boldsymbol{Y}$-axis follows from a right handed coordinate system. This makes sure that for instance the maximum twist of a body part can always be controlled employing a minimum and maximum rotation angle around the $\boldsymbol{Z}$-axis.

To transform the coordinates of a locally given point or into the global world coordinate system the following operation has to be performed:

$$\boldsymbol{P}_G = \prod_{i=1}^{n} \mathbf{M}_R^{(i)} \cdot \mathbf{M}_B^{(i)}(\alpha, \beta, \gamma) \cdot \boldsymbol{P}_L, \tag{1}$$

where the homogenous matrix $\mathbf{M}_R^{(i)}$ is constructed from the directions of the coordinate axes $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ and the position $\boldsymbol{T}$ of the fixed coordinate system:

$$\mathbf{M}_R^{(i)} = \begin{pmatrix} \boldsymbol{X}^{(i)} & \boldsymbol{Y}^{(i)} & \boldsymbol{Z}^{(i)} & \boldsymbol{T}^{(i)} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

The homogeneous matrix $\mathbf{M}_B^{(i)}(\alpha, \beta, \gamma)$ is a rotation matrix

$$\mathbf{M}_B^{(i)}(\alpha, \beta, \gamma) = \begin{pmatrix} & & & 0 \\ & \mathbf{R}_{X,Y,Z}^{(i)}(\alpha, \beta, \gamma) & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3}$$

whose coefficients contain the product of the rotations around the $\boldsymbol{X}$-, $\boldsymbol{Y}$- and $\boldsymbol{Z}$-axis with the values α, β and γ respectively.

Each joint carries two coordinate systems that are responsible for the overall motion. $\mathbf{M}_R^{(i)}$ is the transformation of the fixed coordinate system and gives the neutral position as shown in Fig. 5 on the right whereas $\mathbf{M}_B^{(i)}$ gives the actual motion depending on the angles α, β and γ respectively.

The task of motion estimation is to find the matrices $\mathbf{M}_B^{(i)}$ such that the deformed template fits into the 3D measured point cloud of an arbitrary pose. From the matrices the values α, β and γ are derived and can be used to animate computer graphic models or robot manipulators.

3.2 Hierarchical ICP

The body part which exhibits within several poses of a gesture the smallest motion is the lower torso. Therefore the lower torso serves as root of the motion hierarchy in Fig.6 and its motion is investigated first. The ICP (Iterative Closest Point)-algorithm[5] is used to calculate the translation and rotation parameters of the "closest points" from the measured surface data. The translational and rotational parameters of the lower torso represent the body position and orientation of the pose under investigation. In the next step the rotation of the lower torso of the reference model is adapted according to the previous measurements. The motion parameters of the child node in Fig.6 (left), here the upper torso, are calculated using a modified version of the ICP algorithm which only calculates the rotational parameters. The center point for this calculation is given by the joint position of that particular body part which has been determined already through the motion parameters of the hierarchically higher body part (in this case the lower torso). The five independent kinematic chains from Fig. 6 are calculated in the described hierarchic manner from the root (lower torso) to the respective end effectors. Measured pose points, which have served for the previous adaptation are eliminated from the data set in order to prevent manifold assignments of points for the following estimation steps of the hierarchy.

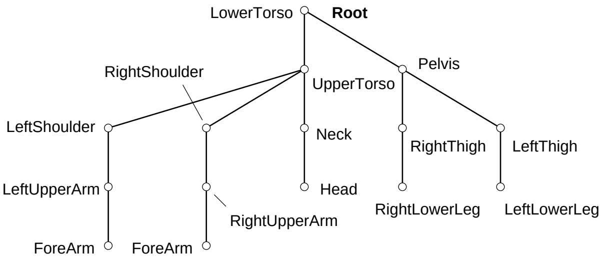

Fig. 6. Estimation hierarchy

In order to consider the differing degrees of freedom for the different joints along the kinematic chains from the root to the end effector a post processing step follows which shifts additional degrees of freedom (DOF) between joints. E.g. the additional DOFs in the elbow joint (the algorithm estimates 3 rotational DOFs) are shifted to the motion in the shoulder joint which must be responsible for the additional motion since the elbow is only equipped with a single DOF.

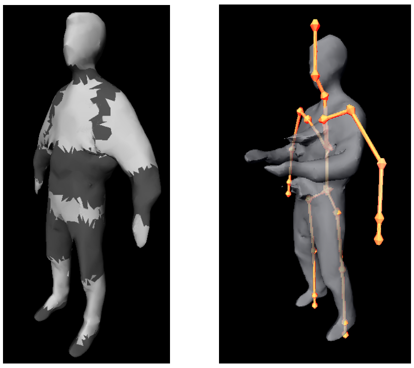

Fig. 7. 3D segmented body model (left) and initial pose of skeleton within the pose “carry”

4 Results

Figure 7 shows the pose under investigation overlaid by the skeleton of the reference model. The differences in the arm and leg positions are obvious. Fig. 8 shows the results of an automated pose estimation as described in this paper. Since the pose estimation is done in three dimensions, the fitting of the skeleton to the cloud of surface points in the pose under investigation is illustrated by views from four different spatial positions. It can be seen, that in this case the skeleton has been adapted to the pose almost perfectly.

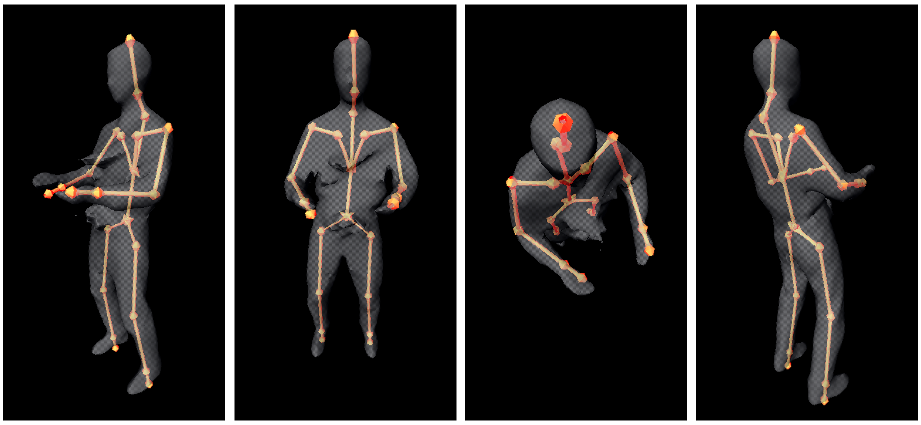

Fig. 8. Estimation results: The initial body skeleton has been fit to the pose "carry"

Right now each pose is treated independently from the others by match against the same, locally adapted reference model. For future applications where body parts merge, for instance in a pose, where an arm is pressed against the torso, it might be necessary to process poses in their natural sequence in order to track the skeleton. However, this endangers the accumulation of estimation errors, which is prevented by our present approach.

References

1. 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.
2. C. Pedney, "Distance-ordered homotopic thinning: a skeletonization algorithm for 3D digital images", Comput. Vis. Image Underst., vol.72, no.3, p404-13, 1998
3. L. Dekker, I. Douros, B.F. Buston, P. Treleaven, "Building symbolic information for 3D human body modeling from range data", Second International Conference on 3D Digital Imaging and Modeling, Ottawa, Ont., Canada, 4-8 Oct. 1999
4. S. Weik, J. Wingbermuehle, W. Niem, "Creation of flexible anthropomorphic models for 3D videoconferencing using shape from silhouettes", Journal of Visualization and Computer Animation 2000, 11, pp 145-154
5. P.J. Besl, N.D. McKay, "A Method for Registration of 3D Shapes", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No.2, 1992
6. D.A. Simon, M. Hebert, T. Kanade, "Real-time 3D Pose Estimation Using a High-Speed Range Sensor", Proc. IEEE International Conference on Robotics and Automation, Vol. 3, pp 2235-2240, 1994
7. D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey", Computer Vision and Image Understanding, Vol. 73, No. 1, pp. 82-98, 1999
8. S. Weik, O. Niemeyer, "Three-dimensional Motion Estimation for Articulated human templates using a sequence of stereoscopic image pairs", Proceedings of Visual Communications and Image Processing, VCIP99, Proceedings of SPIE, Volume 3653, 1999