International Journal of Computer Vision 53(3), 199–223, 2003

Human Body Model Acquisition and Tracking Using Voxel Data

IVANA MIKIĆ
Q3DM, Inc., 10110 Sorrento Valley Rd., Suite B, San Diego, CA 92121, USA
imikic@q3dm.com

MOHAN TRIVEDI
Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive 0434, La Jolla, CA 92093-0434, USA
trivedi@ece.ucsd.edu

EDWARD HUNTER
Q3DM, Inc., 10110 Sorrento Valley Rd., Suite B, San Diego, CA 92121, USA
ehunter@q3dm.com

PAMELA COSMAN
Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive 0407, La Jolla, CA 92093-0407, USA
pcosman@code.ucsd.edu

Received February 14, 2002; Revised November 5, 2002; Accepted January 15, 2003

Abstract. We present an integrated system for automatic acquisition of the human body model and motion tracking using input from multiple synchronized video streams. The video frames are segmented and the 3D voxel reconstructions of the human body shape in each frame are computed from the foreground silhouettes. These reconstructions are then used as input to the model acquisition and tracking algorithms.

The human body model consists of ellipsoids and cylinders and is described using the twists framework resulting in a non-redundant set of model parameters. Model acquisition starts with a simple body part localization procedure based on template fitting and growing, which uses prior knowledge of average body part shapes and dimensions. The initial model is then refined using a Bayesian network that imposes human body proportions onto the body part size estimates. The tracker is an extended Kalman filter that estimates model parameters based on the measurements made on the labeled voxel data. A voxel labeling procedure that handles large frame-to-frame displacements was designed resulting in very robust tracking performance.

Extensive evaluation shows that the system performs very reliably on sequences that include different types of motion such as walking, sitting, dancing, running and jumping and people of very different body sizes, from a nine year old girl to a tall adult male.

Keywords: human body model acquisition, motion capture, pose estimation

1. Introduction

Tracking of the human body, also called motion capture or posture estimation, is a problem of estimating the parameters of the human body model (such as joint angles) from the video data as the position and configuration of the tracked body change over time.

A reliable motion capture system would be valuable in many applications (Gavrila, 1999; Moeslund and Granum, 2001). One class of applications are those where the extracted body model parameters are used directly, for example to interact with a virtual world, drive an animated avatar in a video game or for computer graphics character animation. Another class of applications use extracted parameters to classify and recognize people, gestures or motions, such as surveillance systems, intelligent environments, or advanced user interfaces (sign language translation, gesture driven control, gait, or pose recognition). Finally, the motion parameters can be used for motion analysis in applications such as personalized sports training, choreography, or clinical studies of orthopedic patients.

Human body tracking algorithms usually assume that the body model of the tracked person is known and placed close to the true position in the beginning of the tracking process. These algorithms then estimate the model parameters in time to reflect the motion of the person. For a fully automated motion capture system, in addition to tracking, the model acquisition problem needs to be solved. The goal of the model acquisition is to estimate the parameters of the human body model that correspond to the specific shape and size of the tracked person and to place and configure the model to accurately reflect the position and configuration of the body in the beginning of the motion capture process.

In this paper we present a fully automated system for motion capture that includes both the model acquisition and the motion tracking. While most researchers have taken the approach of working directly with the image data, we use 3D voxel reconstructions of the human body shape at each frame as input to the model acquisition and tracking (Mikić et al., 2001). This approach leads to simple and robust algorithms that take advantage of the unique qualities of voxel data. The price is an additional preprocessing step where the 3D voxel reconstructions are computed from the image data—a process that can be performed in real-time using dedicated hardware.

Section 2 describes the related work and Section 3 gives an overview of the system. The algorithm for computing 3D voxel reconstructions is presented in Section 4, the human body model is described in Section 5 and the tracking in Section 6. Section 7 describes the model acquisition and Section 8 shows the results of the system evaluation. The conclusion follows in Section 9. Additional details can be found in Mikić (2002).

2. Related Work

Currently available commercial systems for motion capture require the subject to wear special markers, body suits or gloves. In the past few years, the problem of markerless, unconstrained posture estimation using only cameras has received much attention from computer vision researchers.

Many existing posture estimation systems require manual initialization of the model and then perform tracking. Very few approaches exist where the model is acquired automatically and in those cases, the person is usually required to perform a set of calibration movements that identify the body parts to the system (Kakadiaris and Metaxas, 1998; Cheung et al., 2000). Once the model is available in the first frame, a very common approach to tracking is to perform iterations of four steps until good agreement between the model and the data is achieved: prediction of the model position in the next frame, projection of the model to the image plane(s), comparison of the projection with the data in the new frame and adjustment of the model position based on this comparison.

Algorithms have been developed that take input from one (Rehg and Kanade, 1995; Hunter, 1999; Bregler, 1997; DiFranco et al., 2001; Howe et al., 1999; Wachter and Nagel, 1999; Sminchiescu and Triggs, 2001; Ioffe and Forsyth, 2001) or multiple cameras (Kakadiaris and Metaxas, 1996; Bregler and Malik, 1998; Gavrila and Davis, 1996; Delamarre and Faugeras, 2001; Yamamoto et al., 1998; Hilton, 1999; Jung and Wohn, 1997). Rehg and Kanade (1995) have developed a system for tracking a 3D articulated model of a hand based on a layered template representation of self-occlusions. Hunter (1999) and Hunter et al. (1997) developed an algorithm based on the Expectation Maximization (EM) procedure that assigns foreground pixels to body parts and then updates body part positions to explain the data. An extra processing step, based on virtual work static equilibrium conditions, integrates object kinematic structure into the

EM procedure, guaranteeing that the resulting posture is kinematically valid. An algorithm using products of exponential maps to relate the parameters of the human body model to the optical flow measurements was described by Bregler and Malik (1998). Wren (2000) designed the DYNA system, driven by 2D blob features from multiple cameras that are probabilistically integrated into a 3D human body model. Also, the system includes a feedback from the 3D body model to the 2D feature tracking by setting the appropriate prior probabilities using the extended Kalman filter. This framework accounts for 'behaviors'—the aspects of motion that cannot be explained by passive physics but represent purposeful human motion. Gavrila and Davis (1996) used a human body model composed of tapered super-quadrics to track (multiple) people in 3D. They use a constant acceleration kinematic model to predict positions of body parts in the next frame. Their locations are then adjusted using the undirected normalized chamfer distance between image contours and contours of the projected model (in multiple images). The search is decomposed in stages: they first adjust positions of the head and the torso, then arms and legs. Kakadiaris and Metaxas have developed a system for 3D human body model acquisition (1998) and tracking (1996) using three cameras placed in a mutually orthogonal configuration. The person under observation is requested to perform a set of movements according to a protocol that incrementally reveals the structure of the human body. Once the model has been acquired, the tracking is performed using the physics-based framework (Metaxas and Terzopoulos, 1993). Based on the expected body position, the difference between the predicted and actual images is used to calculate forces that are applied to the model. The dynamics are modeled using the extended Kalman filter. The tracking result, a new body pose, is a result of the applied forces acting on the physics-based model. The problem of occlusions is solved by choosing from the available cameras those that provide visibility of the part and observability of its motion, for every body part at every frame. Delamarre and Faugeras (2001) describe an algorithm that computes human body contours based on optical flow and intensity. Then, forces are applied that attempt to align the outline of the model to the contours extracted from the data. This procedure is repeated until a good agreement is achieved. Deutscher et al. (2000, 2001) developed a system based on the CONDENSATION algorithm (particle filter) (Isard and Blake, 1996). Deutscher introduced a modified particle filter to handle high dimensional configuration space of human motion capture. It uses a continuation principle, based on annealing, to gradually introduce the influence of narrow peaks in the fitness function. Two image features are used in combination: edges and foreground silhouettes. Good tracking results are achieved using this approach.

Promising results have been reported using the depth data obtained from stereo (Covell et al., 2000; Jojić et al., 1999; Plankers and Fua, 1999, 2001) for pose estimation. The first attempt at using voxel data obtained from multiple cameras to estimate body pose has been reported in Cheung et al. (2000). A simple six-part body model is fitted to the 3D voxel reconstruction. The tracking is performed by assigning the voxels in the new frame to the closest body part from the previous frame and by recomputing the new position of the body part based on the voxels assigned to it. This simple approach does not guarantee that two adjacent body parts would not drift apart and also can lose track easily for moderately fast motions.

In the algorithms that work with the data in the image planes, the 3D body model is repeatedly projected onto the image planes to be compared against the extracted image features. A simplified camera model is often used to enable efficient model projection (Hunter, 1999; Bregler, 1997). Another problem in working with the image plane data is that different body parts appear in different sizes and may be occluded depending on the relative position of the body to the camera and on the body pose.

When using 3D voxel reconstructions as input to the motion capture system, very detailed camera models can be used, since the computations that use the camera model can be done off-line and stored in lookup tables. The voxel data is in the same 3D space as the body model, therefore, there is no need for repeated projections of the model to the image space. Also, since the voxel reconstruction is of the same dimensions as the real person's body, the sizes of different body parts are stable and do not depend on the person's position and pose. This allows the design of simple algorithms that take advantage of our knowledge of average shapes and sizes of body parts.

3. System Overview

The system flowchart is shown in Fig. 1. The main components are the 3D voxel reconstruction, model

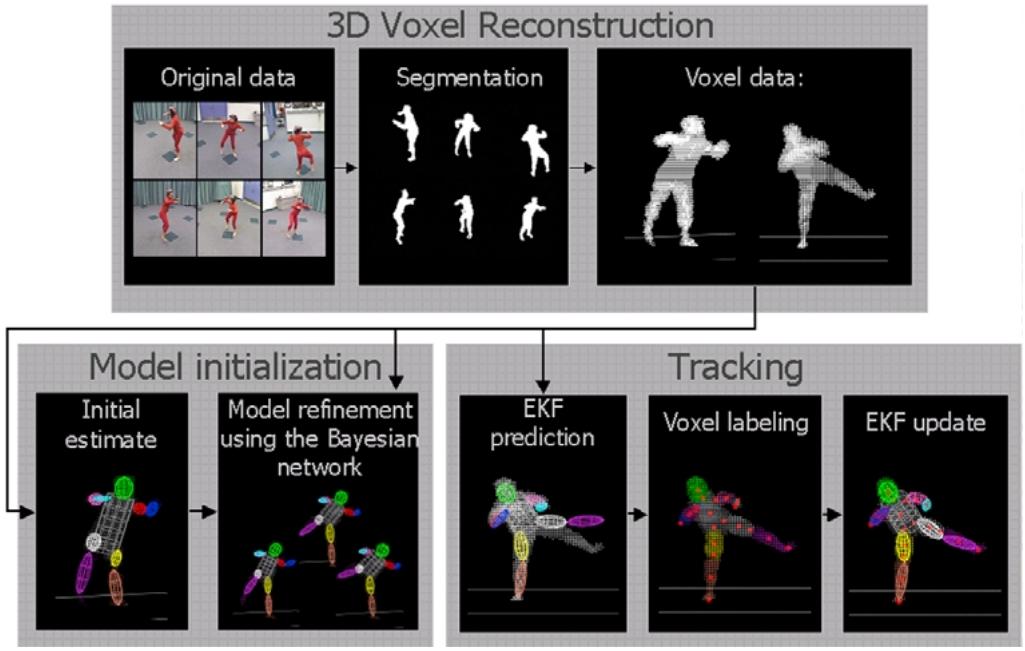

Figure 1. The system flowchart.

acquisition and motion tracking. The 3D voxel reconstruction takes multiple synchronized, segmented video streams and computes the reconstruction of the shape represented by the foreground pixels. The system computes the volume of interest from the foreground bounding boxes in each of the camera views. Then, for each candidate voxel in the volume of interest, it is checked whether its projections onto each of the image planes coincide with a foreground pixel. If yes, that voxel is assigned to the reconstruction, otherwise the voxel is set to zero. The cameras are calibrated, and calibration parameters are used by the system to compute the projections of the voxels onto the image planes.

Model acquisition is performed in two stages. In the first frame of the sequence, a simple template fitting and growing procedure is used to locate different body parts. This procedure takes advantage of the prior knowledge of average sizes and shapes of different body parts. This initial model is then refined using a Bayesian network that incorporates the knowledge of human body proportions into the estimates of body part sizes. This procedure converges rapidly.

Finally, the tracking procedure executes the predict-update cycle at each frame. Using the prediction of the model position and configuration from the previous frame, the voxels in the new frame are assigned to one of the body parts. Measurements of locations of specific points on the body are extracted from the labeled voxels and a Kalman filter is used to adjust the model position and configuration to best fit the extracted measurements. The voxel labeling procedure combines template fitting and distance minimizing approaches. This algorithm can handle large frame-to-frame displacements and results in robust tracking performance.

The human body model used in this project is described using the twists framework developed in the robotics community (Murray et al., 1993). In this formulation, the constraints that ensure kinematically valid postures allowed by the degrees of freedom in the joints and the connectedness between specific body parts are incorporated into the model. This results in the non-redundant set of model parameters and in the simple and stable tracking algorithm, since there is no need to impose these constraints during the tracking process.

4. 3D Voxel Reconstruction

To compute the voxel reconstruction, the camera images are first segmented using the algorithm described

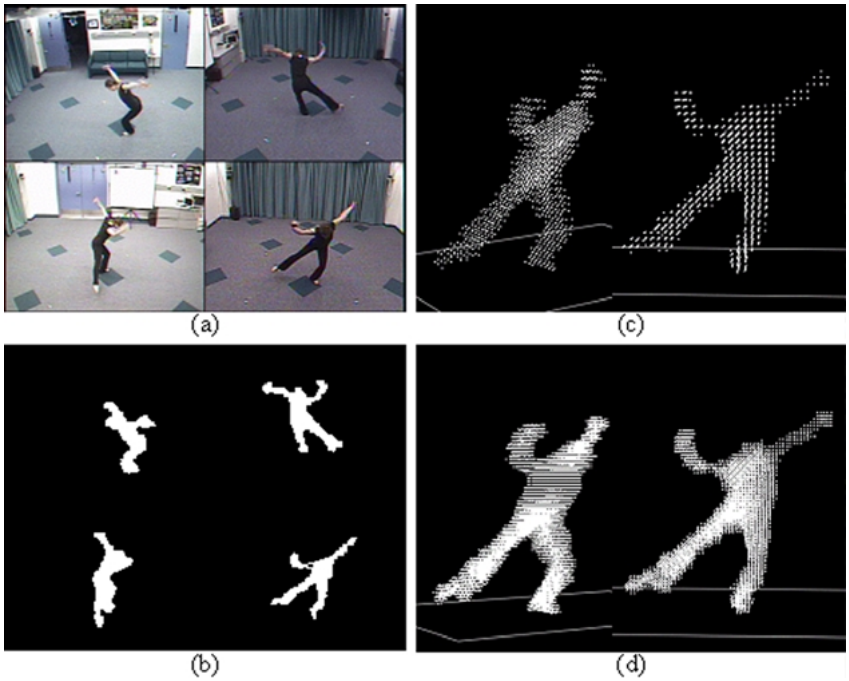

Figure 2. 3D voxel reconstruction: (a) original input images, (b) extracted 2D silhouettes, (c) two views of the resulting 3D reconstruction at voxel size of 50 mm and (d) two views of the 3D reconstruction at voxel size of 25 mm.

in Horprasert et al. (1999) which eliminates shadows and highlights and produces good quality silhouettes. Based on the centroids and bounding boxes of the 2D silhouettes, a bounding volume of the person is computed. Cameras are calibrated using Tsai's algorithm (Tsai, 1987).

Reconstructing a 3D shape using silhouettes from multiple images is called voxel carving or shape from silhouettes. Octree (Szeliski, 1993) is one of the best known approaches to voxel carving. The volume of interest is first represented by one cube, which is progressively subdivided into eight subcubes. Once it is determined that a subcube is entirely inside or entirely outside the 3D reconstruction, its subdivision is stopped. Cubes are organized in a tree, and once all the leaves stop dividing, the tree gives an efficient representation of the 3D shape. The more straightforward approach is to check for each voxel if it is consistent with all silhouettes. With several clever speed-up methods, this approach is described in Cheung et al. (2000). In this system, the person is known to always be inside a predetermined volume of interest. A projection of each voxel in that volume onto each of the image planes is precomputed and stored in a lookup table. Then, at runtime, the process of checking whether the voxel is consistent with a 2D silhouette is very fast since the use of the lookup table eliminates most of the necessary computations.

Our goal is to allow a person unconstrained movement in a large space by using multiple pan-tilt cameras. Therefore, designing a lookup table that maps voxels to pixels in each camera image is not practical. Instead, we pre-compute a lookup table that maps points from undistorted sensor-plane coordinates (divided by focal length and quantized) in Tsai's model to the image pixels. Then, the only computation that is performed at runtime is mapping from world coordinates to undistorted sensor-plane coordinates. A voxel is uniformly sampled and included in the 3D reconstruction if the majority of the sample points agree with all image silhouettes. Voxel size is chosen by the user. An example frame is shown in Fig. 2 with reconstruction at two resolutions.

The equations are given below, where $\mathbf{x}_w$ is the voxel's world coordinate, $\mathbf{x}_c$ is its coordinate in the camera coordinate system, X_u and Y_u are undistorted sensor-plane coordinates, X_d and Y_d are distorted (true) sensor-plane coordinates and X_f and Y_f are pixel coordinates. The lookup table is fixed for a camera regardless of its orientation (would work for a pan/tilt

camera also—only rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{T}$ change in this case).

Lookup table computations:

$$\left(\frac{X_u}{f}, \frac{Y_u}{f}\right) \rightarrow (X_f, Y_f)$$
$$X_u = X_d\left(1 + \kappa_1\left(X_d^2 + Y_d^2\right)\right)$$
$$Y_u = Y_d\left(1 + \kappa_1\left(X_d^2 + Y_d^2\right)\right)$$
$$X_f = d_x^{-1} X_d s_x + C_x$$
$$Y_f = d_y^{-1} Y_d + C_y$$

Run-time computations:

$$(x_w, y_w, z_w) \rightarrow \left(\frac{X_u}{f}, \frac{Y_u}{f}\right)$$
$$\mathbf{x}_c = \mathbf{R}\mathbf{x}_w + \mathbf{T}$$
$$\frac{X_u}{f} = \frac{x_c}{z_c}, \quad \frac{Y_u}{f} = \frac{y_c}{z_c}$$

5. Human Body Model

In designing a robust motion capture system that produces kinematically valid posture estimates, the choice of the human body modeling framework is critical. Valid configurations of the model are defined by a number of constraints that can be incorporated into the model itself or imposed during tracking. It is desirable to incorporate as many constraints as possible into the model, since that results in a more robust and stable tracking performance. However, it is also desirable that the relationship between the model parameters and locations of specific points on the body be simple, since it represents the measurement equation of the Kalman filter. The twists framework (Murray et al., 1993) for describing kinematic chains satisfies both requirements. The model is captured with a non-redundant set of parameters (i.e. with most of the constraints incorporated into the model), and the relationship between the model parameters and the locations of specific points on the body is simple.

5.1. *Twists and the Product of Exponentials Formula for Kinematic Chains*

Let us consider a rotation of a rigid object about a fixed axis. Let the unit vector along the axis of rotation be $\boldsymbol{\omega} \in \Re^3$ and $\mathbf{q} \in \Re^3$ be a point on the axis. Assuming that the object rotates with unit velocity, the velocity of a point $\mathbf{p}(t)$ on the object is:

$$\dot{\mathbf{p}}(t) = \boldsymbol{\omega} \times (\mathbf{p}(t) - \mathbf{q}) \tag{1}$$

This can be rewritten in homogeneous coordinates as:

$$\begin{bmatrix} \dot{\mathbf{p}} \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{\omega} & -\boldsymbol{\omega} \times \mathbf{q} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} = \hat{\xi} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}$$
$$\dot{\bar{\mathbf{p}}} = \hat{\xi}\bar{\mathbf{p}} \tag{2}$$

where $\bar{\mathbf{p}} = [\mathbf{p}\ \ 1]^T$ is a homogeneous coordinate of the point $\mathbf{p}$, and $\boldsymbol{\omega} \times \mathbf{x} = \hat{\boldsymbol{\omega}}\mathbf{x}$, $\forall \mathbf{x} \in \Re^3$, i.e.,

$$\hat{\boldsymbol{\omega}} = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \tag{3}$$

and

$$\hat{\xi} = \begin{bmatrix} \hat{\omega} & -\boldsymbol{\omega} \times \mathbf{q} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \hat{\omega} & \mathbf{v} \\ 0 & 0 \end{bmatrix} \tag{4}$$

is defined as a twist associated with the rotation about the axis defined by $\boldsymbol{\omega}$ and $\mathbf{q}$. The solution to the differential Eq. (1) is:

$$\bar{\mathbf{p}}(t) = \mathrm{e}^{\hat{\xi}t}\bar{\mathbf{p}}(0) \tag{5}$$

$\mathrm{e}^{\hat{\xi}t}$ is the mapping (the exponential map associated with the twist $\hat{\xi}$) from the initial location of a point $\mathbf{p}$ to its new location after rotating t radians about the axis defined by $\boldsymbol{\omega}$ and $\mathbf{q}$. It can be shown that

$$\mathbf{M} = \mathrm{e}^{\hat{\xi}\theta} = \begin{bmatrix} \mathrm{e}^{\hat{\omega}\theta} & (\mathbf{I} - \mathrm{e}^{\hat{\omega}\theta})(\boldsymbol{\omega} \times \mathbf{v}) + \boldsymbol{\omega}\boldsymbol{\omega}^T\mathbf{v}\theta \\ 0 & 1 \end{bmatrix} \tag{6}$$

where

$$\mathrm{e}^{\hat{\omega}\theta} = \mathbf{I} + \frac{\hat{\omega}}{\|\boldsymbol{\omega}\|}\sin(\|\boldsymbol{\omega}\|\theta) + \frac{\hat{\omega}^2}{\|\boldsymbol{\omega}\|^2}(1 - \cos(\|\boldsymbol{\omega}\|\theta)) \tag{7}$$

is a rotation matrix associated with the rotation of θ radians about an axis $\boldsymbol{\omega}$. If we have an open kinematic chain with n axes of rotation, it can be shown that:

$$g_P(\boldsymbol{\theta}) = \mathrm{e}^{\hat{\xi}_1\theta}\mathrm{e}^{\hat{\xi}_2\theta}\ldots\mathrm{e}^{\hat{\xi}_n\theta}g_P(0) \tag{8}$$

where $g_P(\boldsymbol{\theta})$ is the rigid body transformation between the base of the chain and a point on the last link of the chain, in the configuration described by the n angles $\boldsymbol{\theta} = [\theta_1\ \ \theta_2\ \ \cdots\ \ \theta_n]^T$; $g_P(0)$ represents the rigid body transformation between the same points for a reference configuration of the chain. Equation (8) is called the product of exponentials formula for an open kinematic chain and it can be shown that it is independent of the

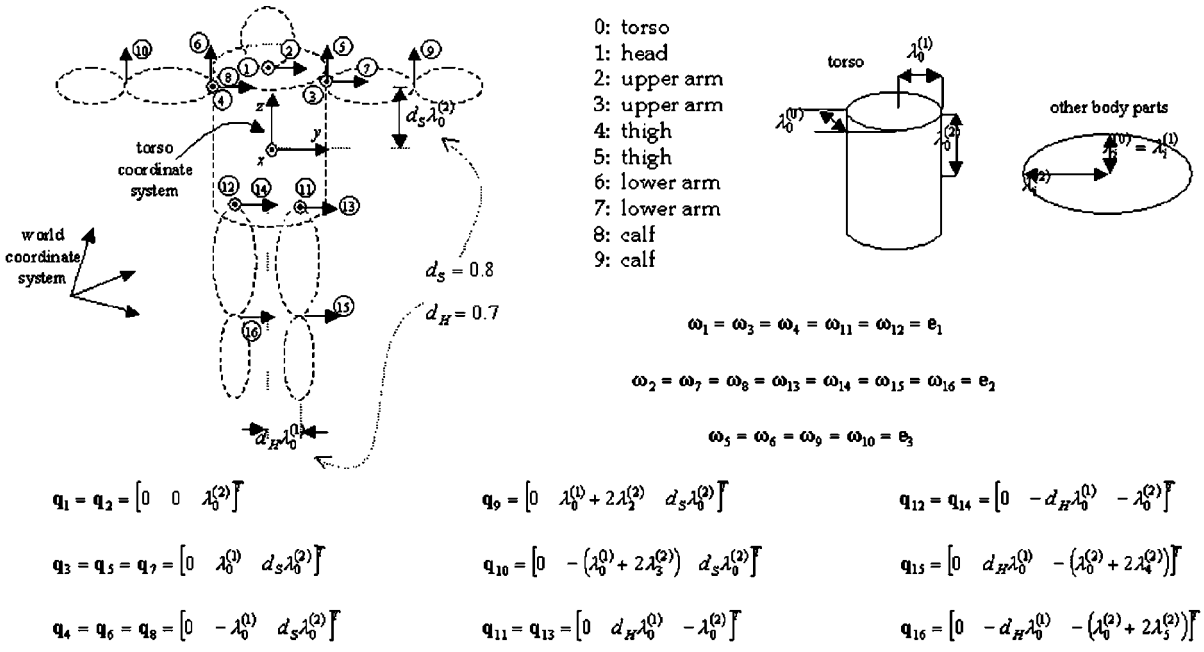

Figure 3. Articulated body model. Sixteen axes of rotation (marked by circled numbers) in body joints are modeled using twists relative to the torso-centered coordinate system. To describe an axis of rotation, a unit vector along the axis and a coordinate of a point on the axis in the "initial" position of the body are needed. As initial position, we chose the one where legs and arms are straight and arms are pointing away from the body as shown in the figure. Dimensions of body parts are determined in the initialization procedure and are held fixed thereafter. Body part dimensions are denoted by λ; subscript refers to the body part number and superscript to dimension order: 0 is for the smallest and 2 for the largest of the three. For all body parts except the torso, the two smaller dimensions are set to be equal.

order in which the rotations are performed. The angles are numbered going from the chain base toward the last link.

5.2. *Twist-Based Human Body Model*

The articulated body model we use is shown in Fig. 3. It consists of five open kinematic chains: torso-head, torso-left arm, torso-right arm, torso-left leg and torso-right leg. Sizes of body parts are denoted as $2\lambda_i^{(j)}$, where i is the body part index, and j is the dimension order—smallest dimension is 0 and largest is 2. For all parts except torso, the two smaller dimensions are set to be equal to the average of the two dimensions estimated during initialization. The positions of joints are fixed relative to the body part dimensions in the torso coordinate system (for example, the hip is at $[0 \;\; d_H\lambda_0^{(1)} \;\; -\lambda_0^{(2)}]^T$).

Sixteen axes of rotation are modeled in different joints. Two in the neck, three in each shoulder, two in each hip and one in each elbow and knee. We take the torso-centered coordinate system as the reference. The range of allowed values is set for each angle. For example, the rotation in the knee can go from 0 to 180 degrees—the knee cannot bend forward. The rotations about these axes (relative to the torso) are modeled using exponential maps, as described in the previous section.

Even though the axes of rotation change as the body moves, in the twists formulation the descriptions of the axes stay fixed and are determined in the initial body configuration. We chose the configuration with extended arms and legs, and with arms pointing to the side of the body (shown in Fig. 3) as the initial body configuration. In this configuration, all angles θ_i are equal to zero. The figure also shows values for the vectors ω_i and $\mathbf{q}_i$ for each axis. The location of a point on the body with respect to the torso is determined by its location in the initial configuration and the product of exponential maps for the axes that affect the position of that point.

Knowing the dimensions of body parts and using the body model shown in Fig. 3, the configuration of the body is completely captured with angles of rotation about each of the axes ($\theta_1 - \theta_{16}$) and the centroid ($\mathbf{t}_0$)

and orientation (rotation matrix $\mathbf{R}_0$) of the torso. Orientation of the torso is parameterized with a unit vector $\boldsymbol{\omega}_0 = [\omega_0 \ \ \omega_1 \ \ \omega_2]$ and the angle θ_0 (Eq. (9)). The position and orientation of the torso are captured using seven parameters—three coordinates for centroid location and four for the orientation. Therefore, the configuration of the described model is fully captured by 23 parameters: $\theta_1 - \theta_{16}$, $\mathbf{t}_0$, $\boldsymbol{\omega}_0$ and θ_0.

$$\mathbf{R}_0 = e^{\hat{\omega}_0\theta_0} = \mathbf{I} + \frac{\hat{\omega}_0}{\|\boldsymbol{\omega}_0\|}\sin(\|\boldsymbol{\omega}_0\|\theta) + \frac{\hat{\omega}_0^2}{\|\boldsymbol{\omega}_0\|^2}(1 - \cos(\|\boldsymbol{\omega}_0\|\theta_0)) \tag{9}$$

The exponential maps associated with each of the sixteen axes of rotation are easily computed using Eq. (6) and the vectors $\boldsymbol{\omega}_i$ and $\mathbf{q}_i$, for each $i = 1, \ldots, 16$ given in Fig. 3.

During the tracking process, locations of specific points on the body, such as the upper arm centroid or neck, are used to adjust the model configuration to the data. To formulate the tracker, it is necessary to derive the equations that relate locations of these points in the reference coordinate system to the parameters of the body model.

For a point $\mathbf{p}$, we define the significant rotations as those affecting the position of the point—if $\mathbf{p}$ is the wrist, there would be four: three in the shoulder and one in the elbow. The set of angles $\boldsymbol{\theta}_\mathbf{p}$ contains the angles associated with the significant rotations. The position of a point $\mathbf{p}_\mathbf{t}(\boldsymbol{\theta}_\mathbf{p})$ with respect to the torso is given by the product of exponential maps corresponding to the set of significant rotations and of the position of the point in the initial configuration $\mathbf{p}_\mathbf{t}(0)$ (in homogeneous coordinates):

$$\bar{\mathbf{p}}_\mathbf{t}(\boldsymbol{\theta}_\mathbf{p}) = \mathbf{M}_{i1}\mathbf{M}_{i2}\ldots\mathbf{M}_{in}\bar{\mathbf{p}}_\mathbf{t}(\mathbf{0}), \quad \text{where} \quad \boldsymbol{\theta}_\mathbf{p} = \{\theta_{i1}, \theta_{i2}, \ldots \theta_{in}\} \tag{10}$$

We denote with $\mathbf{M}_0$ the mapping that corresponds to the torso position and orientation:

$$\mathbf{M}_0 = \begin{bmatrix} \mathbf{R}_0 & \mathbf{t}_0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} e^{\hat{\omega}_0\theta_0} & \mathbf{t}_0 \\ 0 & 1 \end{bmatrix} \tag{11}$$

where $\mathbf{R}_0 = e^{\hat{\omega}_0\theta_0}$ and $\mathbf{t}_0$ is the torso centroid. The homogeneous coordinates of a point with respect to the world coordinate system can now be expressed as:

$$\bar{\mathbf{p}}_0(\boldsymbol{\theta}_\mathbf{p}, \boldsymbol{\omega}_0, \mathbf{t}_0) = \mathbf{M}_0\bar{\mathbf{p}}_\mathbf{t}(\boldsymbol{\theta}_\mathbf{p}) \tag{12}$$

It follows that the Cartesian coordinate of this point is:

$$\mathbf{p}_0(\boldsymbol{\theta}_\mathbf{p}) = \mathbf{R}_0(\mathbf{R}_{i1}(\mathbf{R}_{i2}(\ldots(\mathbf{R}_{in}\mathbf{p}_\mathbf{t}(\mathbf{0}) + \mathbf{t}_{in}) + \cdots) + \mathbf{t}_{i2}) + \mathbf{t}_{i1}) + \mathbf{t}_0 \tag{13}$$

6. Tracking

The algorithm for model acquisition, which estimates body part sizes and their locations in the beginning of the sequence, will be presented in the next section. For now, we will assume that the dimensions of all body parts and their approximate locations in the beginning of the sequence are known. For every new frame, the tracker updates the model position and configuration to reflect the motion of the tracked person. The algorithm flowchart is shown in Fig. 4.

The tracker is an extended Kalman filter (EKF) that estimates the parameters of the model given the measurements extracted from the data. For each new frame, the prediction of the model position and configuration produced from the previous frame is used to label the voxel data and compute the locations of the chosen measurement points. Those measurements are then used by the EKF to update the model parameters and to produce the prediction for the next frame. In this section, the extended Kalman filter is formulated and the voxel labeling algorithm is described.

6.1. The Extended Kalman Filter

The Kalman filter tracker for our problem is defined by:

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{F}[k]\mathbf{x}[k] + \mathbf{v}[k] \\ \mathbf{z}[k] &= \mathbf{h}[k, \mathbf{x}[k]] + \mathbf{w}[k] \end{aligned} \tag{14}$$

where $\mathbf{v}[k]$ and $\mathbf{w}[k]$ are sequences of zero-mean, white, Gaussian noise with covariance matrices $\mathbf{Q}[k]$ and $\mathbf{R}[k]$, respectively. The initial state $\mathbf{x}(0)$ is assumed to be Gaussian with mean $\hat{\mathbf{x}}[0/0]$ and covariance $\mathbf{P}[0/0]$. The 23 parameters of the human body model described in Section 5 constitute the Kalman filter state, $\mathbf{x}$ (Table 1):

$$\mathbf{x} = \begin{bmatrix} \mathbf{t}_0^T & \boldsymbol{\omega}_0^T & \theta_0 & \theta_1 & \cdots & \theta_{16} \end{bmatrix}^T \tag{15}$$

The state transition matrix is set to the identity matrix. For the measurements of the Kalman filter (contained in the vector $\mathbf{z}_k$) we chose 23 points on the human

Table 1. State variables.

$\mathbf{t}_0$	Torso centroid	$\theta_3, \theta_5, \theta_7$	Shoulder angles (L)	θ_{10}	Elbow (R)	θ_{15}	Knee (L)
ω_0, θ_0	Torso orientation	$\theta_4, \theta_6, \theta_8$	Shoulder angles (R)	θ_{11}, θ_{13}	Hip angles (L)	θ_{16}	Knee (R)
θ_1, θ_2	Neck angles	θ_9	Elbow (L)	θ_{12}, θ_{14}	Hip angles (R)		

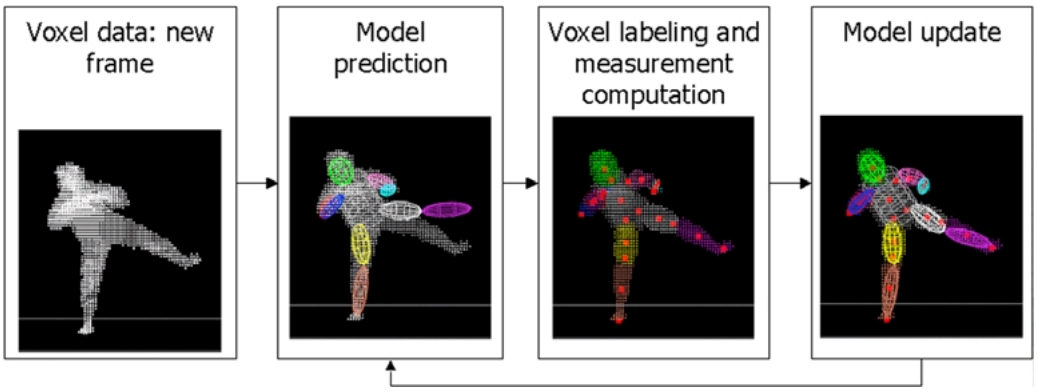

Figure 4. Flow chart of the tracking algorithm. For each new frame, the prediction of the model position is used to label the voxels and compute the locations of measurement points. The tracker then updates the model parameters to fit the new measurements.

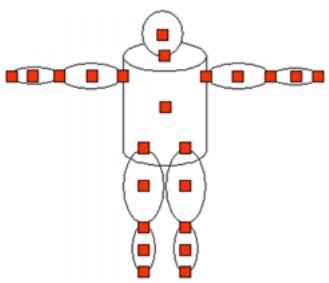

Figure 5. Measurement points.

body which include centroids and endpoints of different body parts (Fig. 5 and Table 2):

$$\mathbf{z} = \begin{bmatrix} \mathbf{p}_0^T & \cdots & \mathbf{p}_{22}^T \end{bmatrix}^T \tag{16}$$

The measurement equation is defined by Eq. (13). The measurement Jacobian $\mathbf{h}_\mathbf{x}[k+1] = [\nabla_\mathbf{x}\mathbf{h}^T[k+1, \mathbf{x}]]^T_{\mathbf{x}=\hat{\mathbf{x}}[k+1/k]}$ is easily computed. For example, the last row of the measurement Jacobian (for the right foot) is determined from:

$$\frac{\partial \mathbf{p}_{22}}{\partial \omega_0} = \frac{\partial \mathbf{R}_0}{\partial \omega_0}(\mathbf{R}_{12}(\mathbf{R}_{14}(\mathbf{R}_{16}\mathbf{p}_{22}(0) + \mathbf{t}_{16}) + \mathbf{t}_{14}) + \mathbf{t}_{12}) \tag{17}$$

$$\frac{\partial \mathbf{p}_{22}}{\partial \theta_0} = \frac{\partial \mathbf{R}_0}{\partial \theta_0}(\mathbf{R}_{12}(\mathbf{R}_{14}(\mathbf{R}_{16}\mathbf{p}_{22}(0) + \mathbf{t}_{16}) + \mathbf{t}_{14}) + \mathbf{t}_{12}) \tag{18}$$

$$\frac{\partial \mathbf{p}_{22}}{\partial \theta_{12}} = \mathbf{R}_0\left(\frac{\partial \mathbf{R}_{12}}{\partial \theta_{12}}(\mathbf{R}_{14}(\mathbf{R}_{16}\mathbf{p}_{22}(0) + \mathbf{t}_{16}) + \mathbf{t}_{14}) + \frac{\partial \mathbf{t}_{12}}{\partial \theta_{12}}\right) \tag{19}$$

$$\frac{\partial \mathbf{p}_{22}}{\partial \theta_{14}} = \mathbf{R}_0\mathbf{R}_{12}\left(\frac{\partial \mathbf{R}_{14}}{\partial \theta_{14}}(\mathbf{R}_{16}\mathbf{p}_{22}(0) + \mathbf{t}_{16}) + \frac{\partial \mathbf{t}_{14}}{\partial \theta_{14}}\right) \tag{20}$$

$$\frac{\partial \mathbf{p}_{22}}{\partial \theta_{16}} = \mathbf{R}_0\mathbf{R}_{12}\mathbf{R}_{14}\left(\frac{\partial \mathbf{R}_{16}}{\partial \theta_{16}}\mathbf{p}_{22}(0) + \frac{\partial \mathbf{t}_{16}}{\partial \theta_{16}}\right) \tag{21}$$

Usually several iterations of the extended Kalman filter algorithm are performed in each frame with the measurement Jacobian updated at every iteration.

6.2. *Voxel Labeling*

Initially, we labeled the voxels based on the Mahalanobis distance from the predicted positions of body parts (Mikić et al., 2001). However, in many cases, this led to loss of track. This was due to the fact that labeling

Table 2. Variables in the measurement vector.

$\mathbf{p}_0$	Torso centroid	$\mathbf{p}_8$	Calf cent. (L)	$\mathbf{p}_{16}$	Elbow (R)
$\mathbf{p}_1$	Head centroid	$\mathbf{p}_9$	Calf cent. (R)	$\mathbf{p}_{17}$	Fingertips (L)
$\mathbf{p}_2$	Upper arm cent. (L)	$\mathbf{p}_{10}$	Neck	$\mathbf{p}_{18}$	Fingertips (R)
$\mathbf{p}_3$	Upper arm cent. (R)	$\mathbf{p}_{11}$	Shoulder (L)	$\mathbf{p}_{19}$	Knee (L)
$\mathbf{p}_4$	Thigh cent. (L)	$\mathbf{p}_{12}$	Shoulder (R)	$\mathbf{p}_{20}$	Knee (R)
$\mathbf{p}_5$	Thigh cent. (R)	$\mathbf{p}_{13}$	Hip (L)	$\mathbf{p}_{21}$	Foot (L)
$\mathbf{p}_6$	Lower arm cent. (L)	$\mathbf{p}_{14}$	Hip (R)	$\mathbf{p}_{22}$	Foot (R)
$\mathbf{p}_7$	Lower arm cent. (R)	$\mathbf{p}_{15}$	Elbow (L)		

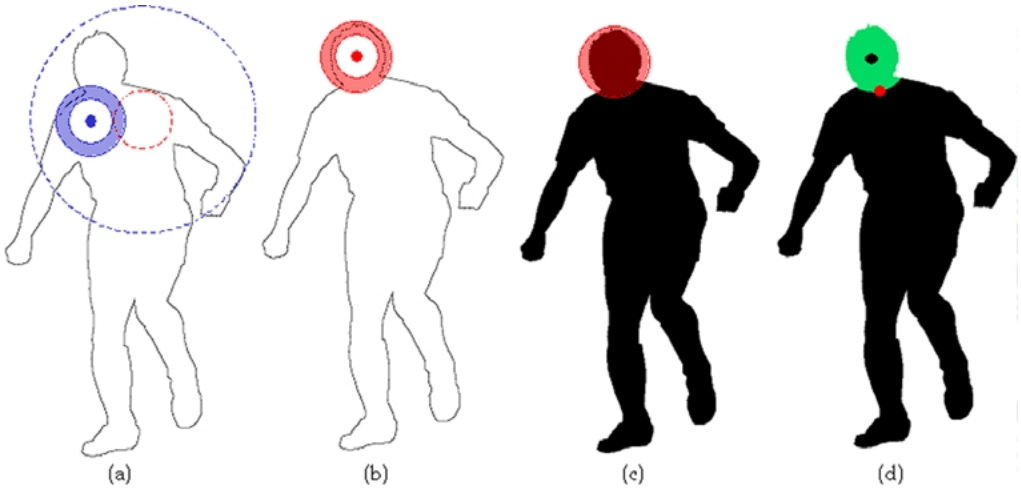

Figure 6. Head location procedure illustrated in a 2D cross-section. (a) Search for the location of the center of a spherical crust template that contains the maximum number of surface voxels. Small dashed circle is the prediction of the head pose from the previous frame. It determines the search area (large dashed circle) for the new head location; (b) the best location is found; (c) voxels that are inside the sphere of a larger diameter are labeled as belonging to the head; (d) head voxels, the head center and the neck.

based purely on distance cannot produce a good result when the model prediction is not very close to true positions of body parts. We have, therefore, designed an algorithm that takes advantage of the qualities of voxel data to perform reliable labeling even for very large frame-to-frame displacements. The head and torso are located first without relying on the distance from the prediction, but based on their unique shapes and sizes. Next, the predictions of limb locations are modified to preserve joint angles with respect to the new positions of the head and the torso. This is the key step that enables tracking for large displacements. The limb voxels are then labeled based on distance from the modified predictions. In the reminder of this section, the detailed description of the voxel labeling algorithm is given.

Due to its unique shape and size, the head is easiest to find and is located first (see Fig. 6). We create a spherical crust template whose inner and outer diameters correspond to the smallest and largest head dimensions. The template center location that maximizes the number of surface voxels that are inside the crust is chosen as the head center. Then, the voxels that are inside the sphere of the larger diameter, centered at the chosen head center are labeled as belonging to the head, and the true center and orientation of the head are recomputed from those voxels. The approximate location of the neck is found as an average over the head voxels with at least one neighbor a non-head body voxel. Since the prediction of the head location is available, the search for the head center can be limited to some neighborhood, which

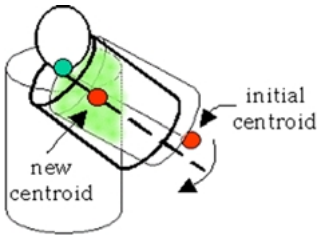

Figure 7. Fitting the torso. The torso template is placed so that its base is at the neck and its main axis passes through the centroid of non-head voxels. Voxels that are inside the template are used to calculate the new centroid and the template is rotated to align the main axis with the new centroid. The process is repeated until the template stops moving which happens when it is entirely inside the torso, or is well centered over it.

speeds up the search and reduces the likelihood of error.

The torso is located next. The template of the size of the person's torso (with circular cross-section of the radius equal to the larger of the two torso radii) is placed with its base at the neck and with its axis going through the centroid of non-head body voxels. The voxels inside the template are then used to recompute a new centroid, and the template is rotated so that its axis passes through it (Fig. 7). The template is anchored to the neck at the center of its base at all times. This procedure is repeated until the template stops moving, which is accomplished when it is entirely inside the torso or is well centered over it.

Next, the predictions for the four limbs are modified to maintain the predicted hip and shoulder angles with the new torso position, which usually moves them much closer to the true positions of the limbs. The remaining voxels are then assigned to the four limbs based on Mahalanobis distance from these modified positions. To locate upper arms and thighs, the same fitting procedure used for the torso is repeated, including only the appropriate limb voxels, with templates anchored at the shoulders/hips. When the voxels belonging to upper arms and thighs are labeled, the remaining voxels in each of the limbs are labeled as lower arms or calves.

Once all the voxels are labeled, the 23 measurement points are easily computed as centroids or endpoints of appropriate blobs. The extended Kalman filter tracker is then used to adjust the model to the measurements in the new frame and to produce the prediction for the next frame. Figure 8 illustrates the voxel labeling and tracking.

7. Model Acquisition

The human body model is chosen a priori and is the same for all humans. However, the actual sizes of body parts vary from person to person. Obviously, for each captured sequence, the initial locations of different body parts will vary also. Model acquisition, therefore, involves both locating the body parts and estimating their true sizes from the data in the beginning of a sequence. It is performed in two stages (Fig. 9). First, rough estimates of body part locations and sizes in the first frame are generated using a simple template fitting and growing algorithm. In the second stage, this estimate is refined over several subsequent frames using a Bayesian network that takes into account both the measured body dimensions and the known proportions of the human body. During this refinement process, the Bayesian network is inserted into the tracking loop, using the body part size measurements produced by the voxel labeling to modify the model, which is then adjusted to best fit the data using the extended Kalman filter. When the body part sizes stop changing, the Bayesian network is "turned off" and the regular tracking continues.

7.1. Initial Estimation of Body Part Locations and Sizes

This procedure is similar to the voxel labeling described in Section 6.2. However, the prediction from the previous frame does not exist (this is the first frame) and the sizes of body parts are not known. Therefore, several modifications and additional steps are needed.

The algorithm illustrated in Fig. 6 is still used to locate the head, however, the inner and outer diameters of the spherical crust template are now set to the smallest and largest head diameters we expect to see. Also, the whole volume has to be searched. Errors are more likely than during voxel labeling for tracking, but are still quite rare: in our experiments on 600 frames, this version located the head correctly in 95% of the frames.

To locate the torso, the same fitting procedure described for voxel labeling is used (Fig. 7), but with the template of an average sized torso. Then, the torso template is shrunk to a small predetermined size in its new location and grown in all dimensions until further growth starts including empty voxels. At every step of the growing, the torso is reoriented as shown in Fig. 7 to ensure that it is well centered during

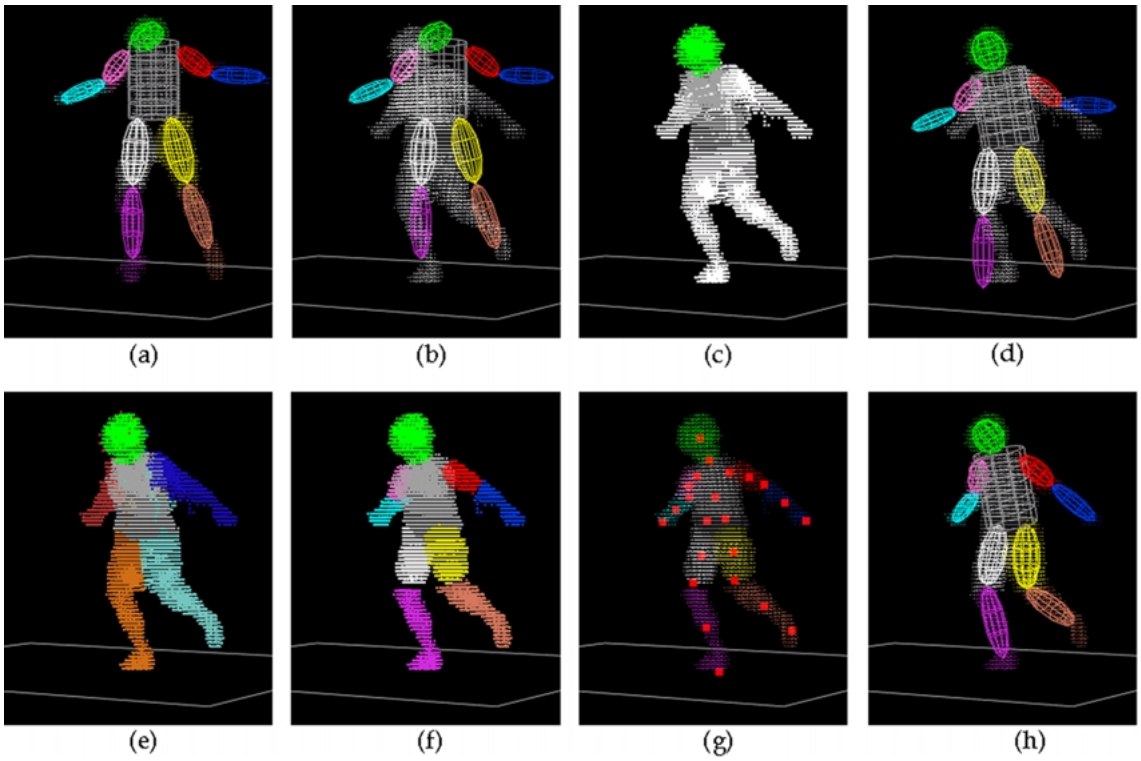

Figure 8. Voxel labeling and tracking. (a) Tracking result in the previous frame; (b) model prediction in the new frame; (c) head and torso located; (d) limbs moved to preserve the predicted hip and joint angles for the new torso position and orientation; (e) four limbs are labeled by minimizing the Mahalanobis distance from the limb positions shown in (d); (f) upper arms and thighs are labeled by fitting them inside the limbs, anchored at the shoulder/hip joints. The remaining limb voxels are labeled as lower arms and calves; (g) the measurement points are easily computed from the labeled voxels; (h) tracker adjusts the body model to fit the data in the new frame.

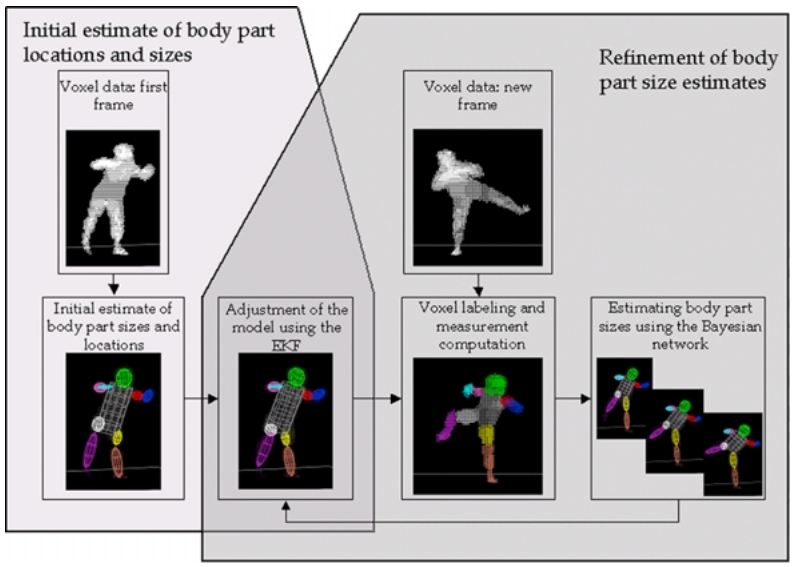

Figure 9. Flow chart of the model acquisition process.

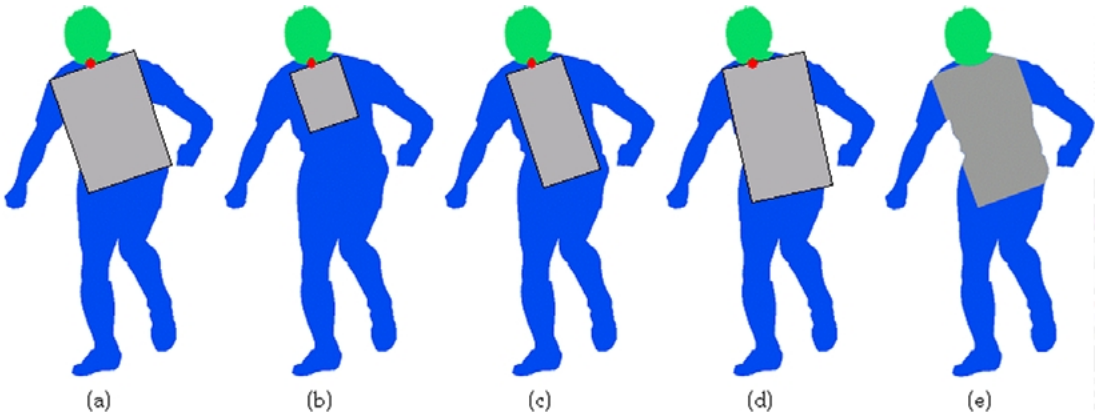

Figure 10. Torso locating procedure illustrated in a 2D cross-section. (a) Initial torso template is fitted to the data; (b) It is then replaced by a small template of predetermined size which is anchored at the same neck point and oriented the same way; (c) the template is then grown and reoriented at every step of growing to ensure the growth does not go in the wrong direction; (d) the growing is stopped when it starts including empty voxels; (e) voxels inside the final template are labeled as belonging to the torso.

growth. In the direction of the legs, the growing will stop at the place where legs part. The voxels inside this new template are labeled as belonging to the torso (Fig. 10).

Next, the four regions belonging to the limbs are found as the four largest connected regions of remaining voxels. The hip and shoulder joints are located as the centroids for voxels at the border of the torso and each of the limbs. Then, the same fitting and growing procedure described for the torso is repeated for thighs and upper arms. The lower arms and calves are found by locating connected components closest to the identified upper arms and thighs. Figure 11 shows the described initial body part localization on real voxel data.

7.2. *Model Refinement*

The estimates of body part sizes and locations in the first frame are produced using the algorithm described in the previous section. It performs robustly, but the sizes of the torso and the limbs are often very inaccurate and depend on the body pose in the first frame. For example, if the person is standing with legs straight and close together, the initial torso will be very long and include much of the legs. The estimates of the thigh and calf sizes will be very small. Obviously, an additional mechanism for estimating true body part sizes is needed.

In addition to the initial estimate of the body part sizes and of the person's height, a general knowledge of human body proportions is available. To take that important knowledge into account when reasoning about body part sizes, we are using Bayesian networks (BNs). A BN is inserted into the tracking loop (Fig. 12), modifying the estimates of body part lengths at each new frame. The EKF tracker adjusts the new model position and configuration to the data, the voxel labeling procedure provides the measurements in the following frame, which are then used by the BN to update the estimates of body part lengths. This procedure is repeated until the body part lengths stop changing, which is usually achieved in three to four frames.

The domain knowledge that is useful for designing the Bayesian network is: the human body is symmetric, i.e., the corresponding body parts on the left and the right sides are of the same dimensions; the lengths of the head, the torso, the thigh and the calf add up to the person's height; the proportions of the human body are known.

The measurements that can be made from the data are the sizes of all body parts and the person's height. The height of the person, the dimensions of the head and the two width dimensions for all other body parts are measured quite accurately. The lengths of different body parts are the ones that are inaccurately measured. This is due to the fact that the measured lengths depend on the borders between body parts, which are hard to locate accurately. For example, if the leg is extended, it is very hard to determine where the thigh ends and the calf begins, but the two width dimensions can be very accurately determined from the data.

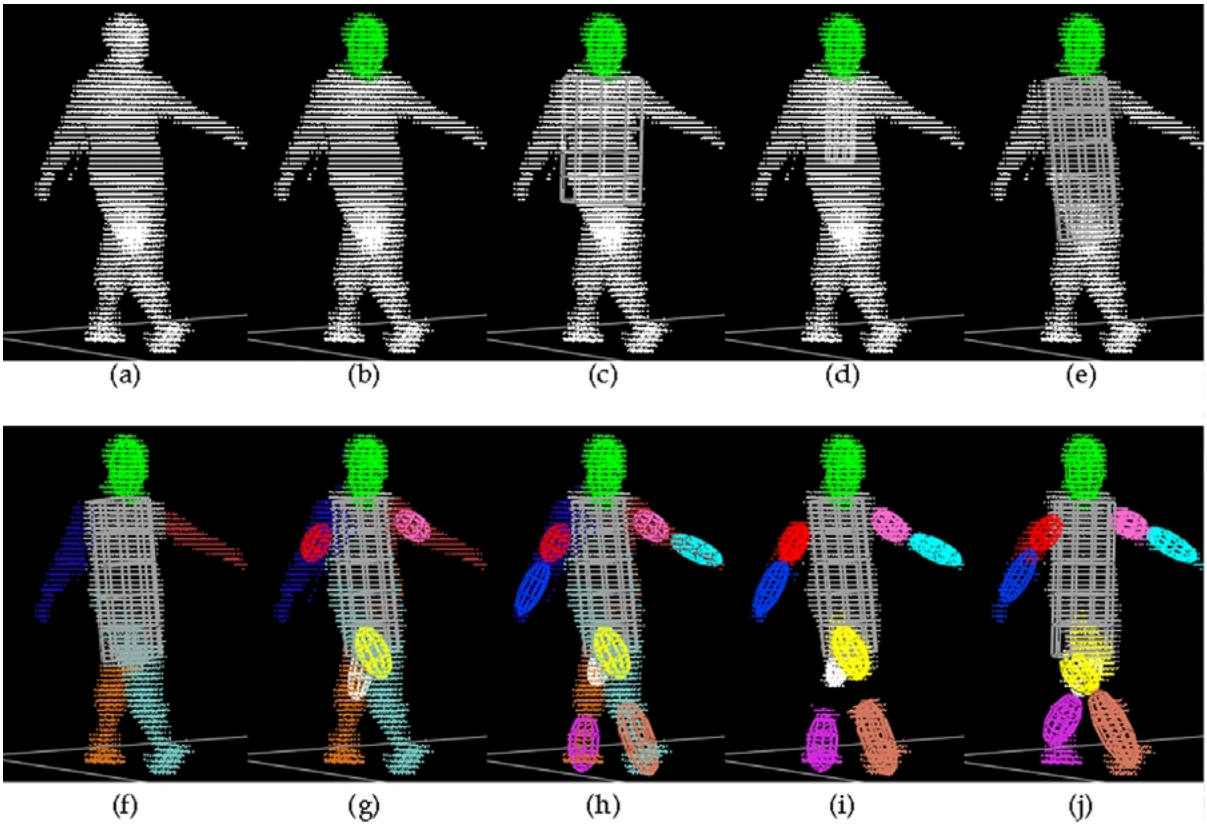

Figure 11. Initial body part localization. (a) 3D voxel reconstruction; (b) head located; (c) initial torso template anchored at the neck centered over the non-head voxels; (d) start of the torso growing; (e) final result of torso growing with torso voxels labeled; (f) four limbs labeled as four largest remaining connected components; (g) upper arms and thighs are grown anchored at the shoulders/hips with the same procedure used for torso; (h) lower arms and calves are fitted to the remaining voxels; (i) all voxels are labeled; (j) current model adjusted to the data using the EKF to ensure a kinematically valid posture estimate.

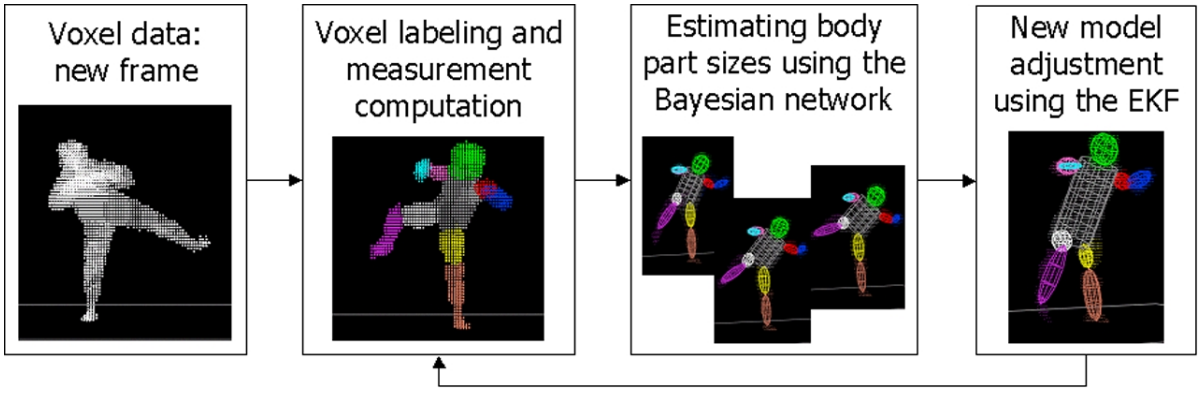

Figure 12. Body part size estimation.

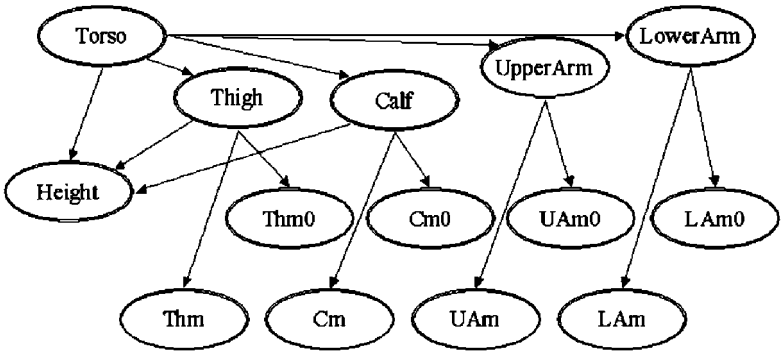

Figure 13. The Bayesian network for estimating body part lengths. Each node represents a length. The leaf nodes are measurements (Thm represents the new thigh measurement, Thm0 reflects the past measurements etc.). Nodes Torso, Thigh, Calf, UpperArm and LowerArm are random variables that represent true body part lengths.

Taking into account what is known about the human body and what can be measured from the data, we can conclude that there is no need to refine our estimates of the head dimensions or the width dimensions of other body parts since they can be accurately estimated from the data, and our knowledge of body proportions would not be of much help in these cases anyway. However, for body part lengths, the refinement is necessary and the available prior knowledge is very useful. Therefore, we have built a Bayesian network shown in Fig. 13 that estimates the lengths of body parts and that takes into account what is known and what can be measured.

Each node represents a continuous random variable. Leaf nodes Thm, Cm, UAm and LAm are the measurements of the lengths of the thigh, calf and upper and lower arm in the current frame. Leaf node Height is the measurement of the person's height (minus head length) computed in the first frame. If the person's height is significantly smaller than the sum of measured lengths of appropriate body parts, we take that sum as the true height—in case the person is not standing up. Leaf nodes Thm0, Cm0, UAm0 and LAm0 are used to increase the influence of past measurements and speed up the convergence. Each of these nodes is updated with the mean of the marginal distribution of its parent from the previous frame. Other nodes (Torso, Thigh, Calf, UpperArm and LowerArm) are random variables that represent true body part lengths. Due to the body symmetry, we include only one node for each of the lengths of the limb body parts and update the corresponding measurement node with the average of the measurements from the left and right sides. The measurement of the torso length is not used because the voxel labeling procedure just fits the known torso to the data, therefore the torso length measurement is essentially the same as the torso length in the model from the previous frame.

All variables are Gaussian and the distribution of a node Y with continuous parents $\mathbf{Z}$ is of the form:

$$p(Y/\mathbf{Z} = \mathbf{z}) = \mathcal{N}(\alpha + \boldsymbol{\beta}^T \mathbf{z}, \sigma^2) \qquad (22)$$

Therefore, for each node with n parents, a set of n weights $\boldsymbol{\beta} = [\beta_1 \ \ldots \ \beta_n]^T$, a standard deviation σ and possibly a constant α are the parameters that need to be chosen. These parameters have clear physical interpretation and are quite easy to select. The selected parameters for the network in Fig. 13 are shown in Table 3. Nodes Thigh, Calf, UpperArm and LowerArm have each only one parent (Torso) and the weight parameters represent known body proportions. Node Height has a Gaussian distribution with the mean equal to the sum of thigh, calf and torso lengths (hence all three weights are equal to 1). Each of the nodes Thm0, Cm0, UAm0 and LAm0 is updated with the mean of the marginal distribution of its parent in the previous frame—hence the weight of 1.

7.3. Determining Body Orientation

The initial body part localization procedure does not include determining which side of the body is left and which is right. This is important to do because it affects the angle range for each joint. For example, if the y-axis of the torso points to the person's left, the angle

Table 3. The parameters used in the Bayesian network shown in Fig. 13. Torso is the only parent node, and therefore the only node that needs an a priori distribution. Nodes Thigh, Calf, UpperArm and LowerArm have each only one parent. Torso and their probabilities conditioned on the torso length are Gaussian with the mean that linearly depends on the torso length. The weight parameters represent known body proportions. Node Height has a Gaussian distribution with the mean equal to the sum of thigh, calf and torso lengths. Each of the nodes Thm0, Cm0, UAm0 and LAm0 is updated with the mean of the marginal distribution of its parent in the previous frame—hence the weight of 1.

Node	Parameter values (mm) β, σ	Node	Parameter values (mm) β, σ	Node	Parameter values (mm) β, σ
Thigh	[0.9], 100	Thm	[1], 200	Thm0	[1], 100
Calf	[1], 100	Cm	[1], 200	Cm0	[1], 100
UpperArm	[0.55], 100	UAm	[1], 200	UAm0	[1], 100
LowerArm	[0.75], 100	LAm	[1], 200	LAm0	[1], 100
Torso	$(\mu, \sigma) = (500, 300)$			Height	$[1, 1, 1]^{\mathrm{T}}$, 100

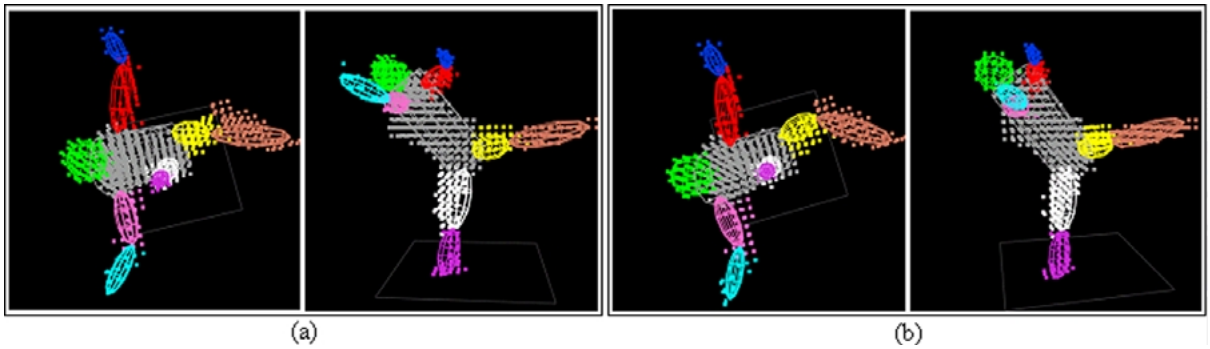

Figure 14. Determining the body orientation. Left: top-down view, Right: side view. (a) In the first frame, using the erroneous orientation, EKF cannot properly fit one leg because it is trying to fit the left leg to the measurements from the right leg; (b) using the other orientation, the fit is much closer and this orientation is accepted as the correct one.

range for either knee is [0, 180] degrees. If, however, it points to the right, the knee range is [−180, 0]. To determine this overall body orientation, we first try to adjust the model to the body part positions that are initially estimated using the EKF described in the previous section using an arbitrary body orientation. Then, we switch the body orientation (i.e., the angle limits) and repeat the adjustment. If a significant difference in the quality of fit exists, the orientation that produces a smaller error is chosen. The quality of the fit is quantified by the sum of Euclidean distances between the measurement points and the corresponding points in the model. Otherwise, the decision is deferred to the next frame. The quality of the fit is quantified by the sum of Euclidean distances between the measurement points and the corresponding points in the model.

This is illustrated in Fig. 14. Figure 14(a) shows the model adjustment with erroneous orientation. Figure 14(b) shows the correct orientation with knees appropriately bent.

8. Results

The system presented in this paper was evaluated in a set of experiments with sequences containing people of different heights and body types, from a nine-year-old girl to a tall adult male, and different types of motions such as walking, dancing, jumping, sitting, stair climbing, etc. (Fig. 15). For each sequence, six synchronized full frame (640 × 480 pixels) video streams were captured at approximately 10 Hz. The voxel size was set to 25 × 25 × 25 mm.

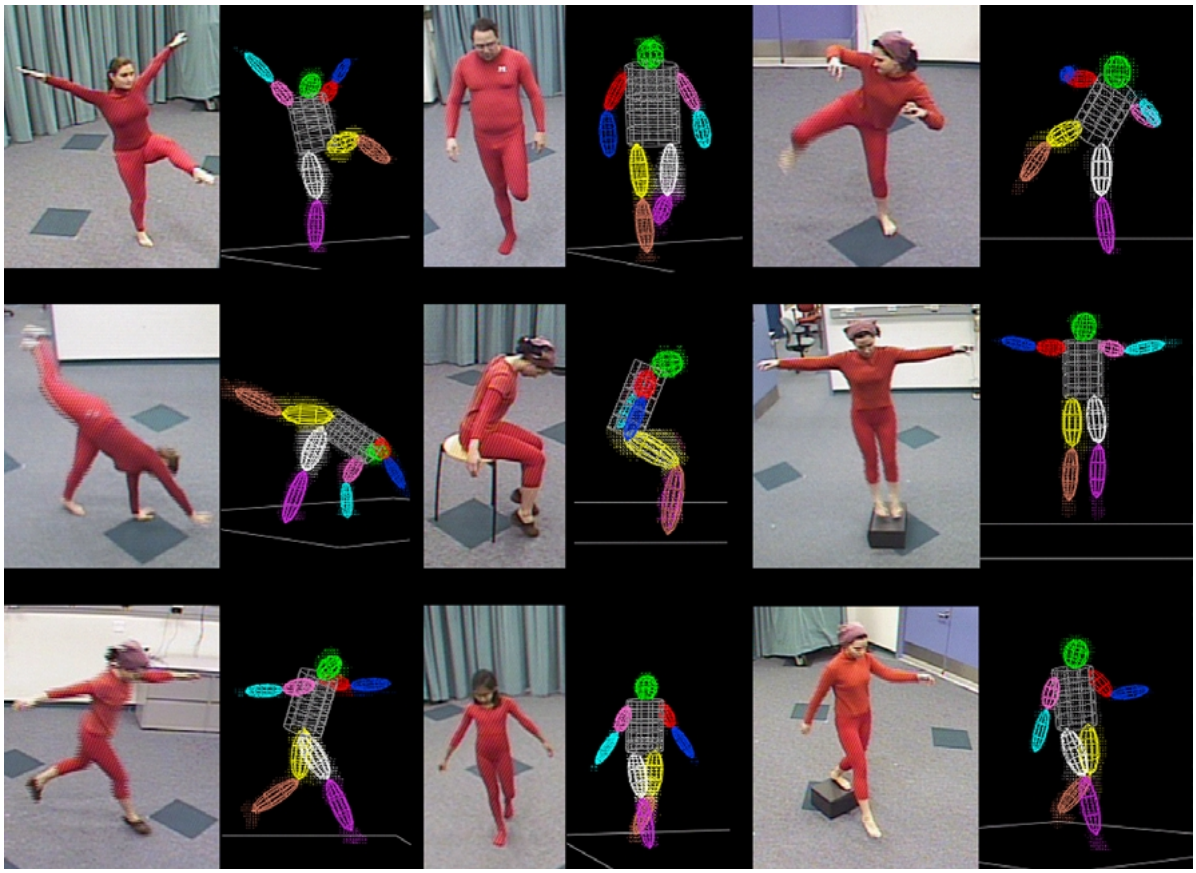

Figure 15. The performance of the system was evaluated on sequences with people of different body sizes and with different motions.

In this section, we present some of the results of these experiments. However, the quality of the results this system produces can be fully appreciated only by viewing the movies that show the model overlaid over the 3D voxel reconstructions, which can be found at http://cvrr.ucsd.edu/∼ivana/projects.htm. The details of the system evaluation can be found in Mikić (2002).

8.1. Model Acquisition

The iterations of the model refinement are stopped when the sum of body part length changes falls below 3 mm, which usually happens in three to four frames. Figure 16 shows the model acquisition process in a sequence where the person is climbing a step that is behind her. In the first frame, the initial model has a very long torso and short thighs and calves. The model refinement converges in three frames producing a very good estimate of body part lengths.

Figure 17 shows the body part size estimates in six sequences recorded with the same person and the true body part sizes measured on the person. Some variability is present, but it is quite small—about 3–4 voxel lengths, i.e. 75–100 mm.

Figure 18 shows the original camera views of five people and Fig. 19 shows the corresponding acquired models. The models successfully capture the main features of these very different human bodies. All five models were acquired using the same algorithm and the same Bayesian network with fixed parameters.

8.2. Tracking

The ground truth, to which the tracking results should be compared, is very difficult to obtain. We have taken

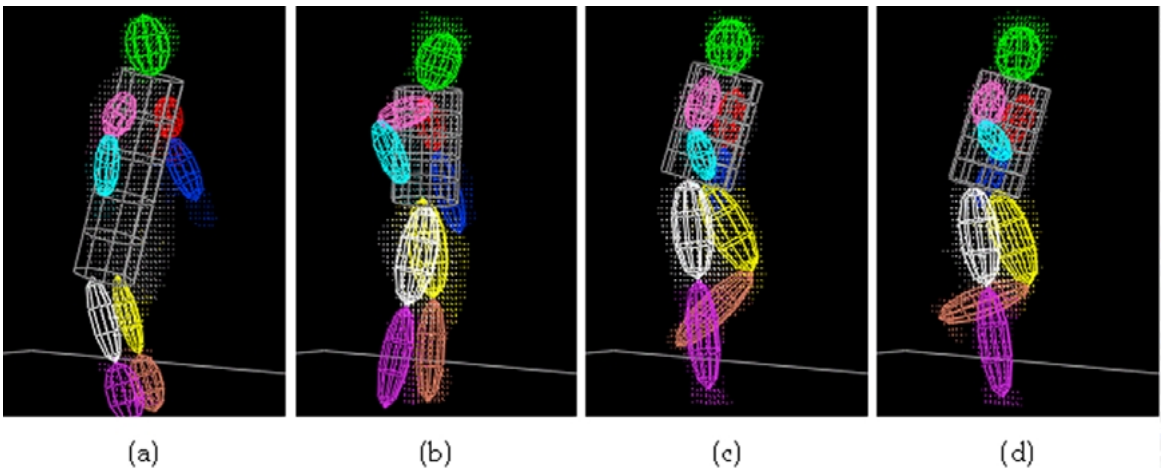

Figure 16. Model refinement. The person is climbing the step that is behind her. For this sequence the procedure converged in three frames. The person is tracked while the model is acquired. (a) Initial estimate of body part sizes and locations in the first frame; (b)–(d) model refinement.

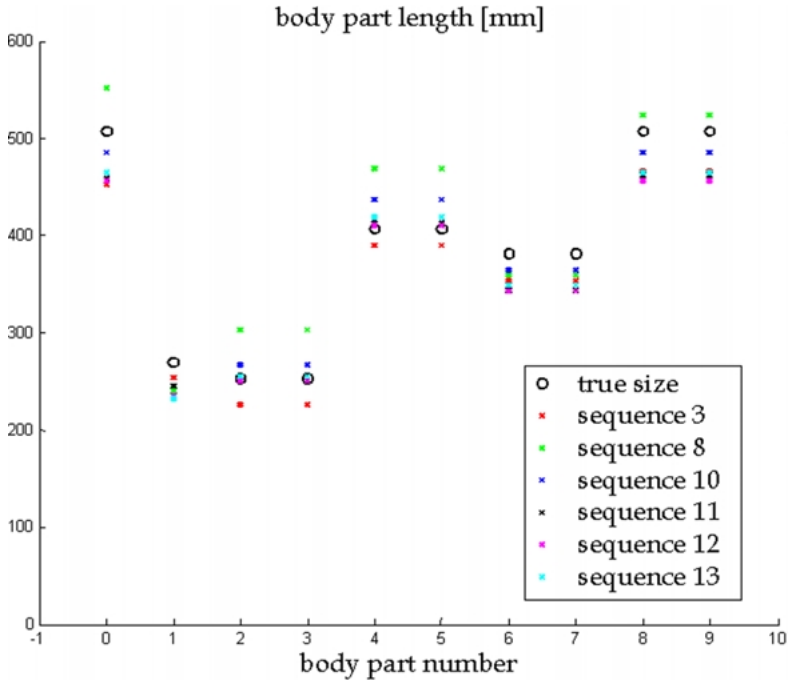

Figure 17. Body part size estimates for six sequences showing the same person, and the true body part sizes. Body part numbers are: 0—torso, 1—head, 2—upper arm (L), 3—upper arm (R), 4—thigh (L), 5—thigh (R), 6—lower arm (L), 7—lower arm (R), 8—calf (L), 9—calf (R).

the approach, as in Hunter (1999), of verifying the tracking accuracy by subjective evaluation—careful visual inspection of result movies. Once we are convinced that the results are accurate, i.e. unbiased, a smooth curve fitted to the data is taken as the substitute for the ground truth and the precision of the tracking output is evaluated using quantitative measures. Savitzky-Golay filter (Press et al., 1993) is used

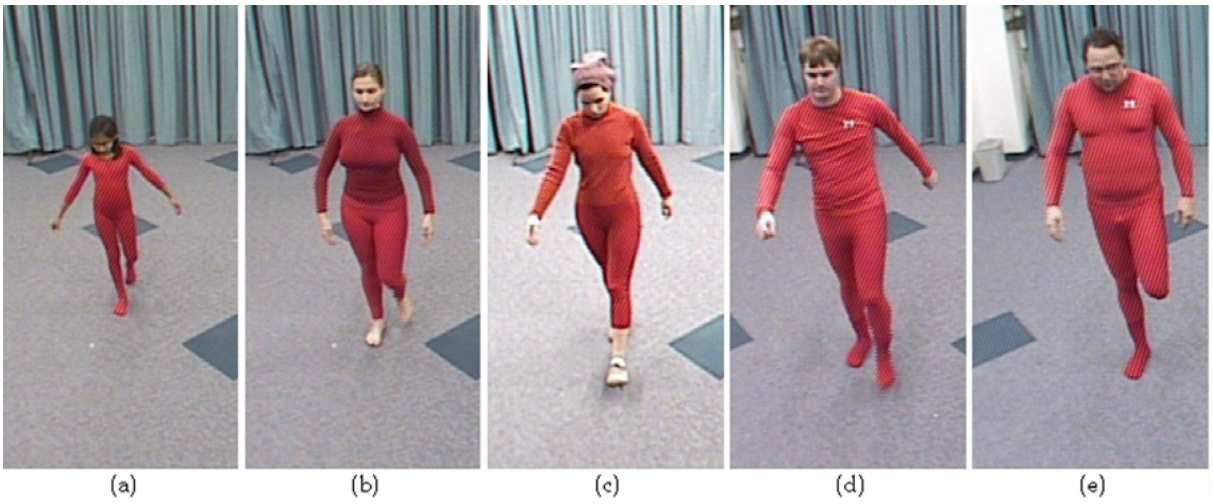

Figure 18. Original views of the five people shown in Fig. 19. (a) Aditi (height: 1295.4 mm); (b) Natalie (height: 1619.25 mm); (c) Ivana (height: 1651 mm); (d) Andrew (height: 1816.1 mm); (e) Brett (height: 1879 mm).

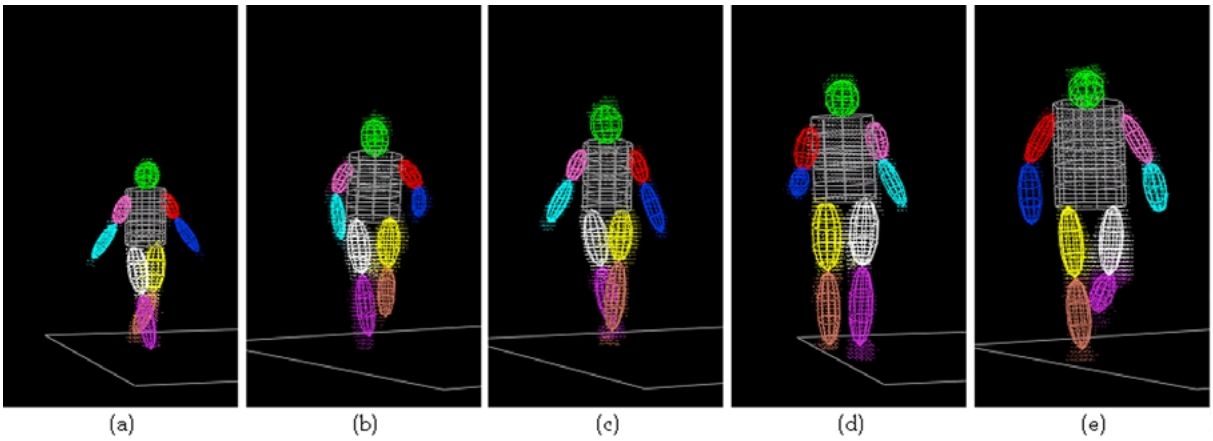

Figure 19. Estimated models for the five people that participated in the experiments. The models are viewed from similar viewpoints—the size differences are due to the true differences in body sizes. (a) Aditi (height: 1295 mm); (b) Natalie (height: 1619 mm); (c) Ivana (height: 1651 mm); (d) Andrew (height: 1816 mm); (e) Brett (height: 1879 mm).

to compute the smooth curve fit. Its output at each point is equal to the value of a polynomial of order M, fitted (using least squares) to points in a $(2n + 1)$ window centered at the point that is currently evaluated. It achieves smoothing without much attenuation of the important data features. The values of M and n are chosen by the user who subjectively optimizes between the level of smoothing and the preservation of important data features. In our experiments, we used $M = 3$ and $n = 7$.

The tracking results look very good by visual inspection. We show sample frames from three sequences in this section, and invite the reader to view the result movies at http://cvrr.ucsd.edu/~ivana/projects.htm. The precision analysis showed an average absolute error for joint angles of three to four degrees. Shoulder angles θ_7 and θ_8 that capture the rotation about the main axis of the upper arm contribute the most to the average error. When the arm is not bent at the elbow, that angle is not defined and is difficult to estimate. However, in

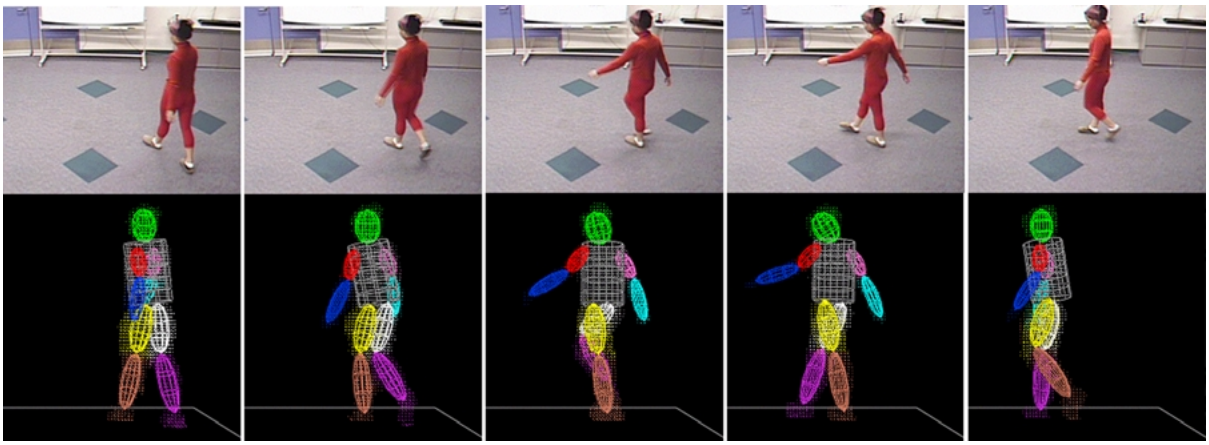

Figure 20. Tracking results for the walking sequence and one of the six original camera views.

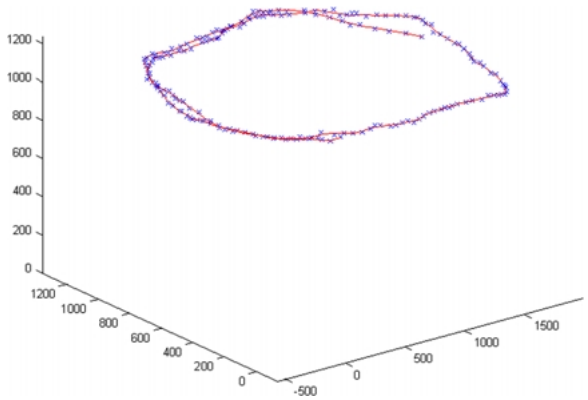

Figure 21. Walking: trajectory of the torso centroid.

these cases, the deviation from the "ground truth" is not really an error. The average error drops below three degrees when these two angles are not considered. The detailed precision figures can be found in Mikić (2002).

Figure 20 shows five sample frames of a walking sequence. Figure 21 shows the trajectory of the torso centroid as the person walked around the room. Figure 22 shows plots of the hip and knee angles as functions of time. This sequence contained 19 steps, 10 by the left and 9 by the right leg, which can easily be correlated with the shown plots.

Figure 23 shows six sample frames for a dance sequence. The results for a running and jumping sequence are shown in Fig. 24.

Our experiments also reveal some limitations of the current version of the model. Figure 25 shows a case where the rotation in the waist that is not captured in our

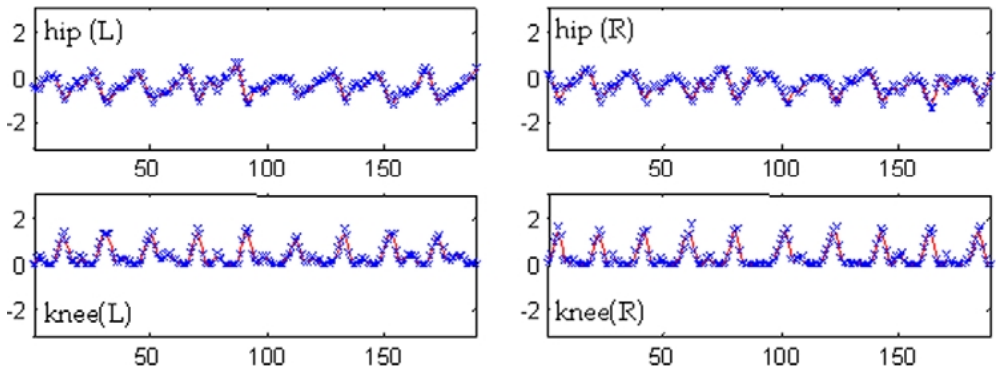

Figure 22. Hip and knee angles [rad] as functions of frame number for the walking sequence.

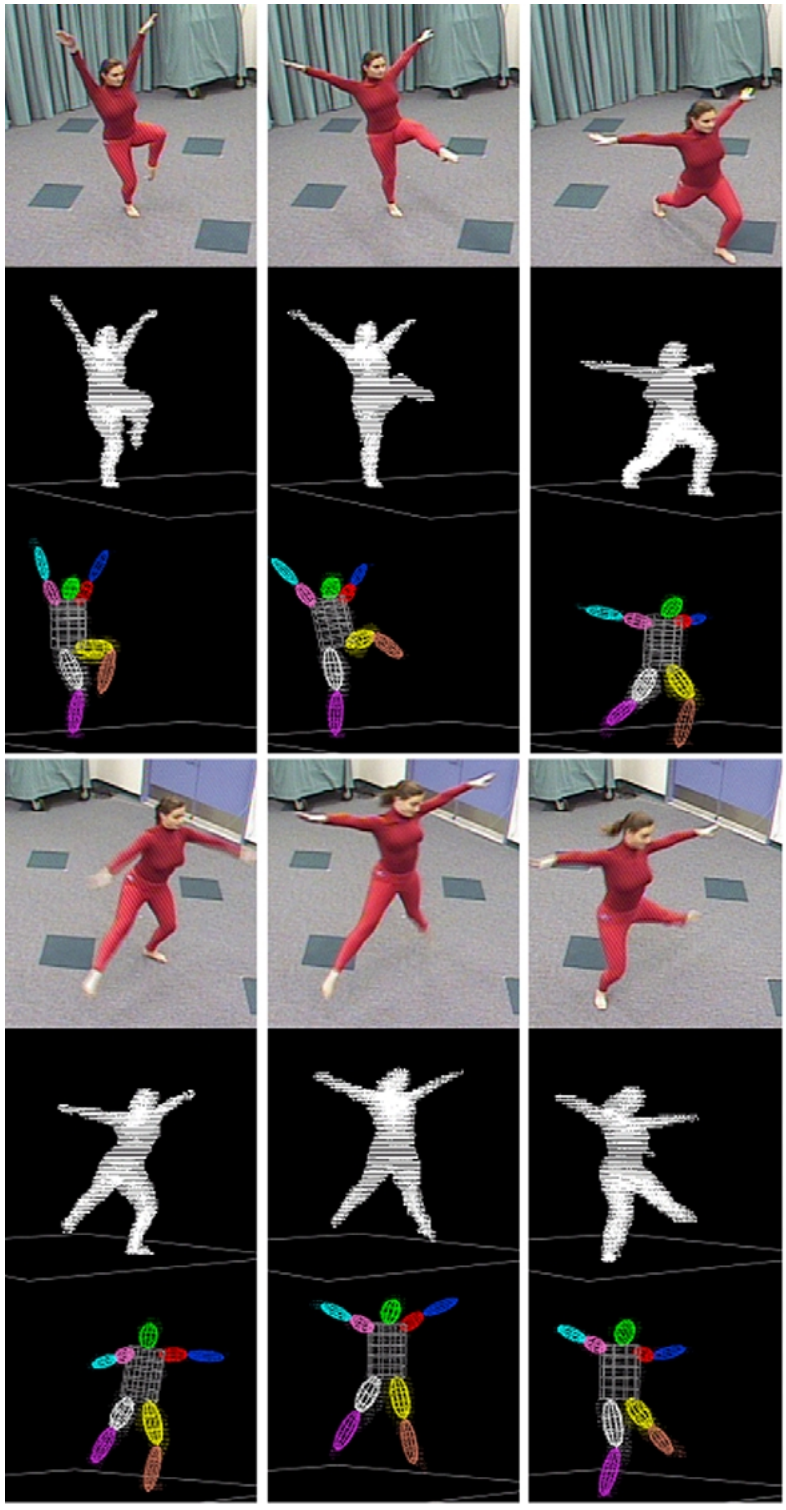

Figure 23. Dance sequence: one of the original camera views, the voxel reconstructions and the tracking results for six sample frames.

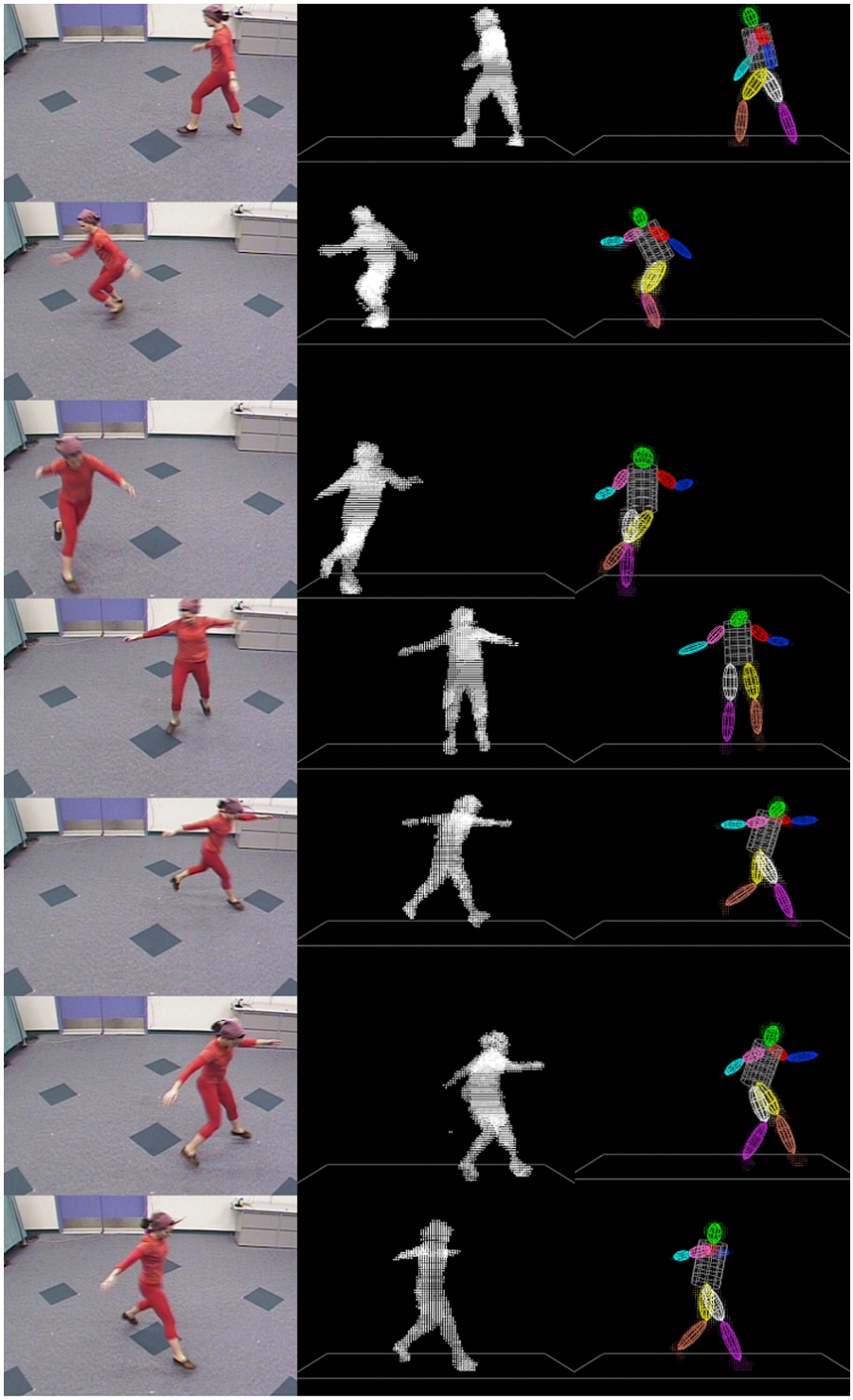

Figure 24. Running and jumping sequence: one of the original camera views, the voxel reconstructions and the tracking results for six sample frames.

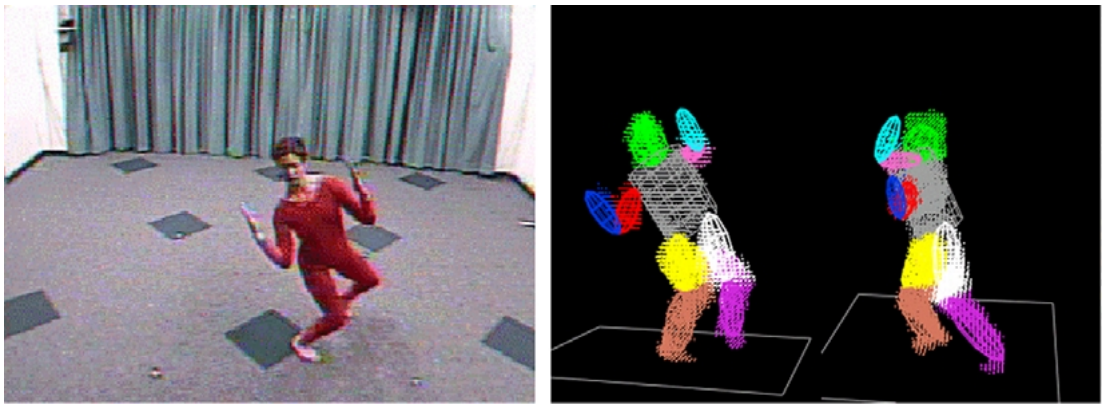

Figure 25. Tracking error due to the rotation in the waist that is not modeled in our body model. Left: original view. Right: the tracking result shown from two different angles.

model causes a tracking error. To improve the results in such cases, we can split the torso in two parts and add one rotation (about the z axis) between them. Such refinements can be easily incorporated in the proposed framework.

9. Conclusion

We have presented a fully automated system for human body model acquisition and tracking using multiple cameras. The system does not perform the tracking directly on the image data, but on the 3D voxel reconstructions computed from the 2D foreground silhouettes. This approach removes all the computations related to the transition between the image planes and the 3D space from the tracking and model acquisition algorithms, making them simple and robust.

The advantages of this approach are twofold. First, the transition from the image space to the 3D space that the real person and the model inhabit is performed once, during the preprocessing stage when the 3D voxel reconstructions are computed. The approaches where the image data is directly used for tracking require repeated projections of the model onto the image planes, often performed several times per frame. Second, analysis of the voxel data is in many ways simpler than the analysis of the image data. Voxel data is in the same 3D space as the model and the real person. Therefore, the measurements made on the voxel data are very easily related to the parameters of the human body model, which makes the tracking algorithm simple and stable. Also, the dimensions and shapes of different body parts are the same in the data as in the real world, which leads to simple algorithms for locating body parts that rely on knowledge of their average shapes and sizes. For example, finding the head in the voxel data is an easy task, since the head has a unique and stable shape and size. However, this is a much harder problem when image data is directly analyzed. The head may be occluded in some views and its size will depend on the relative distance to each of the cameras.

We are using the twists framework to describe the human body model. The constraints that ensure physically valid body configurations are inherent to the model, resulting in a non-redundant set of parameters. Only the constraints for joint angle limits have to be imposed during the tracking. The body is described in a reference configuration, where axes of rotation in different joints coincide with basis vectors of the world coordinate system. The coordinates of any point on the body are easily computed from the model parameters using this framework, which results in a simple tracker formulation that uses locations of different points as measurements to which the model is adjusted.

We have developed an automated model acquisition procedure that does not require any special movements by the tracked person. Model acquisition starts by initial estimation of body part sizes and locations using a template fitting and growing procedure that takes advantage of our knowledge of average shapes and sizes of body parts. Those estimates are then refined using a system whose main component is a Bayesian network, which incorporates the knowledge of human body proportions. The Bayesian network is inserted into the tracking loop, modifying the model as the tracking is performed.

The tracker relies on a hybrid voxel labeling procedure to obtain quality measurements. It combines the minimization of distance from the model prediction and template fitting to produce reliable labeling results, even for large frame to frame displacements.

We have conducted an extensive set of experiments, involving multiple people of heights ranging from 1.3 to 1.9 m, and complex motions ranging from sitting and walking to dancing and jumping. The system performs very reliably in capturing these different types of motion and in acquiring models for different people.

References

Bregler, C. 1997. Learning and recognizing human dynamics in video sequences, *IEEE International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico.

Bregler, C. and Malik, J. 1998. Tracking people with twists and exponential maps, *IEEE International Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA.

Cheung, G., Kanade, T., Bouguet, J., and Holler, M. 2000. A real time system for robust 3D voxel reconstruction of human motions. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, USA, vol. 2, pp. 714–720.

Covell, M., Rahimi, A., Harville, M., and Darrell, T. 2000. Articulated-pose estimation using brightness- and depth-constancy constraints. In *IEEE Int. Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, pp. 438–445.

Delamarre, Q. and Faugeras, O. 2001. 3D articulated models and multi-view tracking with physical forces, The special issue of the *CVIU* journal on modeling people, 81(3):328–357.

Deutscher, J., Blake, A., and Reid, I. 2000. Articulated body motion capture by annealed particle filtering, *IEEE Int. Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC.

Deutscher, J., Davison, A., and Reid, I. 2001. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture, *IEEE Int. Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii.

DiFranco, D., Cham, T., and Rehg, J. 2001. Reconstruction of 3D figure motion from 2D correspondences. In *IEEE Int. Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii.

Gavrila, D. 1999. Visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Gavrila, D. and Davis, L. 1996. 3D model-based tracking of humans in action: A multi-view approach. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 73–80.

Hilton, A. 1999. Towards model-based capture of persons shape, appearance and motion. In *International Workshop on Modeling People at ICCV'99*, Corfu, Greece.

Horprasert, T., Harwood, D., and Davis, L.S. 1999. A statistical approach for real-time robust background subtraction and shadow detection. In *Proc. IEEE ICCV'99 FRAME-RATE Workshop*, Kerkyra, Greece.

Howe, N., Leventon, M., and Freeman, W. 1999. Bayesian reconstruction of 3D human motion from single-camera video. In *Neural Information Processing Systems*, Denver, Colorado.

Hunter, E. 1999. Visual estimation of articulated motion using the expectation-constrained maximization algorithm, Ph.D. Dissertation, University of California, San Diego.

Hunter, E., Kelly, P., and Jain, R. 1997. Estimation of articulated motion using kinematically constrained mixture densities. In *IEEE Nonrigid and Articulated Motion Workshop*, San Juan, Puerto Rico.

Ioffe, S. and Forsyth, D. 2001. Human tracking with mixtures of trees. In *IEEE International Conference on Computer Vision*, Vancouver, Canada.

Isard, M. and Blake, A. 1996. Visual tracking by stochastic propagation of conditional density. In *Proc. 4th European Conference on Computer Vision*, Cambridge, England.

Jojić, N., Turk, M., and Huang, T. 1999. Tracking self-occluding articulated objects in dense disparity maps. In *IEEE Int. Conference on Computer Vision*. Corfu, Greece.

Jung, S. and Wohn, K. 1997. Tracking and motion estimation of the articulated object: A hierarchical Kalman filter approach, *Real-Time Imaging*, 3:415–432.

Kakadiaris, I. and Metaxas, D. 1996. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, San Francisco, CA.

Kakadiaris, I. and Metaxas, D. 1998. Three-dimensional human body model acquisition from multiple views, *International Journal of Computer Vision*, 30(3):191–218.

Metaxas, D. and Terzopoulos, D. 1993. Shape and nonrigid motion estimation through physics-based synthesis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):580–591.

Mikić, I. 2002. Human body model acquisition and tracking using multi-camera voxel data, Ph.D. Dissertation, University of California, San Diego.

Mikić, I., Trivedi, M., Hunter, E., and Cosman, P. 2001. Articulated body posture estimation from multi-camera voxel data. In *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii.

Moeslund, T. and Granum, E. 2001. A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding*, 81:231–268.

Murray, R., Li, Z., and Sastry, S. 1993. A mathematical introduction to robotic manipulation, CRC Press.

Plankers, R. and Fua, P. 1999. Articulated soft objects for video-based body modeling. In *International Workshop on Modeling People at ICCV'99*, Corfu, Greece.

Plankers, R. and Fua, P. 2001. Tracking and modeling people in video sequences, *Computer Vision and Image Understanding*, 81:285–302.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. 1993. *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press.

Rehg, J. and Kanade, T. 1995. Model-based tracking of self-occluding articulated objects. In *IEEE International Conference on Computer Vision*, Cambridge.

Sminchiescu, C. and Triggs, B. 2001. Covariance scaled sampling for monocular 3D body tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii.

Szeliski, R. 1993. Rapid octree construction from image sequences, *CVGIP: Image Understanding*, 58(1):23–32.

Tsai, R. 1987. 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*, RA-3(4):323–344.

Wachter, S. and Nagel, H. 1999. Tracking persons in monocular image sequences, *Computer Vision and Image Understanding*, 74(3):174–192.

Wren, C. 2000. Understanding expressive action, Ph.D. Dissertation, Massachusetts Institute of Technology.

Yamamoto, M., Sato, A., Kawada, S., Kondo, T., and Osaki, Y. 1998. Incremental tracking of human actions from multiple views, *IEEE International Conference on Computer Vision and Pattern Recognition*.