Markerless Motion Capture using Multiple Cameras

Aravind Sundaresan and Rama Chellappa*
Center for Automation Research
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742-3275, USA

Abstract

Motion capture has important applications in different areas such as biomechanics, computer animation, and human-computer interaction. Current motion capture methods use passive markers that are attached to different body parts of the subject and are therefore intrusive in nature. In applications such as pathological human movement analysis, these markers may introduce an unknown artifact in the motion, and are, in general, cumbersome. We present computer vision based methods for performing markerless human motion capture. We model the human body as a set of super-quadrics connected in an articulated structure and propose algorithms to estimate the parameters of the model from video sequences. We compute a volume data (voxel) representation from the images and combine bottom-up approach with top down approach guided by our knowledge of the model. We propose a tracking algorithm that uses this model to track human pose. The tracker uses an iterative framework akin to an Iterated Extended Kalman Filter to estimate articulated human motion using multiple cues that combine both spatial and temporal information in a novel manner. We provide preliminary results using data collected from 8-16 cameras. The emphasis of our work is on models and algorithms that are able to scale with respect to the requirement for accuracy. Our ultimate objective is to build an end-to-end system that can integrate the above mentioned components into a completely automated markerless motion capture system.

1: Introduction

Motion capture for humans describes the activity of analysing and expressing human motion in mathematical terms. The task of motion capture can be divided into a number of systematically distinct groups, initialisation, tracking, pose estimation and gesture recognition. Motion capture is typically accomplished by one of three technologies: optical, magnetic and electro-mechanical, all of which involve markers or devices attached to the subject. Markerless motion capture is a method for motion capture that does not use such markers but uses images obtained from multiple cameras placed around the subject to estimate the pose of the subject. There exist a number of algorithms to estimate the pose from images captured from a single camera, a task that is extremely difficult and ambiguous. Segmentation of the image into different, possibly self-occluding, body parts and tracking them is an inherently difficult problem. It is, therefore, necessary to use multiple cameras to deal with occlusion and kinematic singularities. In general, most computer vision algorithms target applications where only an approximate estimate of the pose is required. They also assume that human body model parameters are available. The complex articulated structure of human beings makes accurately tracking articulated human motion a difficult task. We need mathematical body models to deal with the large number of body segments and to guide the tracking and pose estimation processes.

*This research was funded in part by NSF ITR 0325715.

0-7695-2524-5/05 $20.00 © 2005 IEEE

1.1: Overview

We propose algorithms for automatic human body model acquisition, pose initialisation, and tracking. It is our objective to integrate the elements proposed above into an end-to-end system that performs completely automated markerless motion capture.

In our work, we use parametric shape representations such as modified super-quadrics to represent body segments that are connected in an articulated chain to represent human body structure. We describe the models that we use in Section 2. The following steps are the key elements of the markerless motion capture process and we propose algorithms to solve each of these steps.

- Acquire the model of human subject.
- Estimate the initial pose of the subject.
- Track the pose in subsequent frames using different cues.

The first two steps listed above are closely inter-twined. In the algorithm that we propose, we use an estimate of the pose in order to refine the estimate of the human body model parameters and *vice-versa*. The acquisition-initialisation algorithm is presented in Section 3. The tracking algorithm is described in Section 4. We describe our experiments and present the summary in Section 5. We present a novel method to combine spatial cues such as silhouettes and motion residues. Although we do not use edges, it is also possible to incorporate edges in our method. We also do not constrain the motion or the pose parameters for specific types of motion, such as walking or running.

1.2: Prior work

Most of the existing tracking algorithms use either motion information or spatial information to perform tracking. The use of spatial cues lead to inaccuracies in the pose estimation process while they generally perform well when approximate pose is required. Using only motion cues for tracking leads to drift in the tracking process but is more accurate between two frames. We address the problem of markerless motion capture using multiple cameras. This involves the estimation of pose and the human body model parameters as well as tracking the pose. We propose an algorithm [42] to automatically acquire the human body model using prior knowledge of the articulated structure of a human body in a systematic manner. We also propose an algorithm [43] to track the pose using an articulated model using multiple cameras and multiple cues.

Badler et al. [4] suggest several methods to represent human subjects in terms of their shape as well as the articulated structure. We find that using modified super-quadrics to represent shapes [19] is reasonably accurate for our purposes, though our model can be extended in the same framework to use more sophisticated mesh-models if the data is accurate enough and if the application demands it. Gavrila and Davis [18], Aggarwal and Cai [1], Moeslund and Granum [29], and, more recently, Wang et al. [25] provide surveys of human motion tracking and analysis methods. Cedras and Shah [8] provide a survey of motion-based recognition methods which require the use of motion data that the markerless motion capture methods can provide. There are several applications for markerless motion capture in animation, human-computer interaction and notably in biomechanical and clinical applications where the capture of human motion enables the understanding of normal and pathological human movement [16]. There has also been work on markerless motion capture using articulated Iterative Closest Point algorithm [16, 17] and also estimating models from 3D range data as well as shape-from-silhouette algorithms [3, 15]. We present the prior work in model acquisition in Section 1.2.1 and the prior work in tracking of human pose using both single camera and multiple cameras in Section 1.2.2.

1.2.1 Model acquisition and pose initialisation

Rohr [34] performs automated initialisation of the pose for single camera motion. The model is trained using several human shape models and the assumptions are that the motion is parallel to image plane and is that of gait or cycling. Ramanan and Forsyth [31] also suggest an algorithm that performs rough pose estimation and can be used in an initialisation step. Mikic et al. [27] obtain the human body model using voxels, though their model acquisition algorithm starts with a simple body part localisation procedure based on template fitting and growing, which uses prior knowledge of average body part shapes and dimensions. Kakadiaris and Metaxas [22] present a Human Body Part Identification Strategy (HBPIS) that recovers all the body parts of a moving human based on the spatio-temporal analysis of its deforming silhouette using input from three mutually orthogonal views. However, they specify a protocol of movements that the subject is required to go through. Krahnstoever [24] addresses the issue of acquiring articulated models directly from monocular video. Structure, shape and appearance of articulated models are estimated, but this method is limited in its application of a complete human body model and the fact that it uses a single camera.

Chu et al. [10] describes a method for estimating pose using isomaps [46]. They use isomaps to transform the voxel body to its pose-invariant intrinsic space representation and obtain a skeleton representation. Belkin and Niyogi [5] describe the construction of a representation for data (voxels) lying in a low dimensional manifold embedded in a high dimensional space. We use Laplacian eigenmaps as proposed in [5] in order to simultaneously segment and extract the one-dimensional structure of the voxels. We obtain much better segmentation and explicitly compute the position of the point along this one-dimensional chain and use it to acquire the shape and joint model. Elad and Kimmel [14] describe a method to reduce articulated objects to pose invariant structure. Belkin and Niyogi analyse the connection of Locally Linear Embedding algorithm proposed by Roweis and Saul [35] to the Laplacian. There also exist other methods for dimensionality reduction such as Kernel Eigenvalue analysis [36] and charting a manifold [6]. However, we find the Laplacian Eigenmaps to be intuitively satisfying and effective. Anguelov et al. [2] describe an algorithm that automatically decomposes an object into approximately rigid parts, their location, and the underlying articulated structure given a set of meshes describing the object in different poses. They use an unsupervised non-rigid technique to register the meshes and perform segmentation using the EM algorithm.

1.2.2 Tracking of articulated motion

There are several algorithms to track the pose that use either motion or use silhouettes or voxels, but few combine both motion and static cues as we propose. We look at some existing methods that use either motion-based methods or silhouette or edge based methods to perform tracking. Some of these algorithms use monocular videos and propose algorithms to remove the kinematic ambiguity in the estimation process.

We first look at the methods that use multiple cameras. Yamamoto and Koshikawa [49] analyse human motion based on a robot model and Yamamoto et al. [50] track human motion using multiple cameras. Gavrila and Davis [19] discuss a multi-view approach for 3D model-based tracking of humans in action. They use a generate-and-test algorithm in which they search for poses in a parameter space and match them using a variant of Chamfer matching. Bregler and Malik [7] use an orthographic camera model and integrate a mathematical technique based on the product of exponential maps and twist motions, with differential motion estimation. Kakadiaris and Metaxas [21] use a 3D, model-based, motion estimation method based on the spatio-temporal analysis of the subject's silhouette. Plaenkers and Fua [30] use articulated soft objects with an articulated underlying skeleton as a model, and silhouette data for shape and motion recovery from stereo and trinocular image sequences. Theobalt et al. [47] project the texture of the model obtained from silhouette-based methods and refine the pose using the flow field. Delamarre and Faugeras [11] use 3D articulated models for tracking with silhouettes. They use silhouette contours and apply forces to the contours obtained from the projection of the 3D model so that they move towards the silhouette contours obtained from multiple images. Cheung et al. [23] extend shapes-from-silhouette methods to articulated objects. Given silhouettes of a moving articulated object, they propose an iterative algorithm to solve the simultaneous assignment of silhouette points to a body part and alignment of the body part.

There are also methods that attempt to estimate the pose from a monocular video sequence. They propose different techniques to resolve kinematic ambiguities faced in the monocular pose estimation problem. Ju et al. [20] use planar patches to model body segments. The motion of each patch is defined by eight parameters. For each frame the eight parameters are estimated by applying the optical flow constraint on all pixels in the predicted patches. Sidenbladh et al. [37] provide a framework to track 3D human figures in monocular image sequences using 2D image motion and particle filters with a constrained motion model that restricts the kinds of motions that can be tracked. Wachter and Nagel [48] track persons in monocular image sequences. They use an IEKF with a constant motion model and use edges to region information in the pose update step in their work. Moeslund and Granum[28] use multiple cues for model-based human motion capture and use kinematic constraints to estimate pose of a human arm. The multiple cues are depth (obtained from a stereo rig) and the extracted silhouette, whereas the kinematic constraints are applied in order to restrict the parameter space in terms of impossible poses. Sigal et al. [39, 38] use non-parametric belief propagation to track in a multi view set up. Lan and Huttenlocher [26] use a Bayesian framework to combine pictorial structure spatial models with hidden Markov temporal models. DeMirdjian et al.[12] constrain pose vectors based on kinematic models using SVMs.

Rehg and Morris [32] and Rehg et al. [33] describe ambiguities and singularities in tracking of articulated objects and propose a 2D scaled prismatic model. Sminchisescu and Triggs present a method for recovering 3D human body motion from monocular video sequences using robust image matching, joint limits and non-self-intersection constraints, and a new sample-and-refine search strategy [41]. They also try to remove kinematic ambiguities in monocular pose estimation by using simple kinematic reasoning to enumerate the tree of possible forwards/ backwards flips, thus greatly speeding the search within each linked group of minima [40]. Cham and Rehg [9] describe a probabilistic multiple-hypothesis framework for tracking highly articulated objects using a monocular video. Deutscher et al. [13] introduce a modified particle filter for search in high dimensional configuration spaces that uses a continuation principle, based on annealing, to introduce the influence of narrow peaks in the fitness function, gradually, and is capable of recovering full articulated body motion efficiently.

2: Human body model

The use of human body models greatly simplifies the pose estimation problem and also makes the pose estimation more robust and accurate. As described in Section 3, we can construct voxel-based models of the human body, but this representation is meaningless in terms of our understanding of what the pose of the subject is. We therefore use a human body model consisting of several segments to describe the shape of different parts of the body. These segments are connected in a kinematic tree structure, and the relative positions of these segments with respect to their neighbours determines the pose of the subject. The trade-off in using the model is that it introduces another set of parameters to measure or estimate, namely the body model parameters. However, the level of accuracy we target in our work necessitates the use of such elaborate models. The human body model we use is a set of body segments, modelled in our work as tapered super-quadrics, connected in an articulated model. The parameters of the model are of two kinds: shape parameters and kinematic chain parameters. The shape parameters are the parameters of the super-quadrics representing the human body segments and the kinematic parameters are the positions of the joints in the articulated body. We describe in detail the human body model that we use in our work in Section 2.1. We also describe the capture environment (Keck laboratory) where the sequences used in the experiments were captured, in Section 2.2.

2.1: Human body model

We model the human body as being composed of rigid body segments connected at joints and free to rotate about the joint connecting two segments. Badler et al. [4] in their book describe various types of models that can be used. While their work is mainly intended for the human factors engineers, the issues addressed with respect to the appearance and motion of human beings and their relation to the body structure and various joints is very relevant to our work. The human body model that we use is illustrated in Figure 1 with the different body segments as well as some joints labelled. Each of these body segments have a coordinate frame attached to them. The segment can be described by an arbitrary shape in terms of the coordinates of this frame, and in our case is modelled using a tapered super-quadric. The trunk is the base and the neck and head segments as well as the four limbs form kinematic chains originating from the trunk.

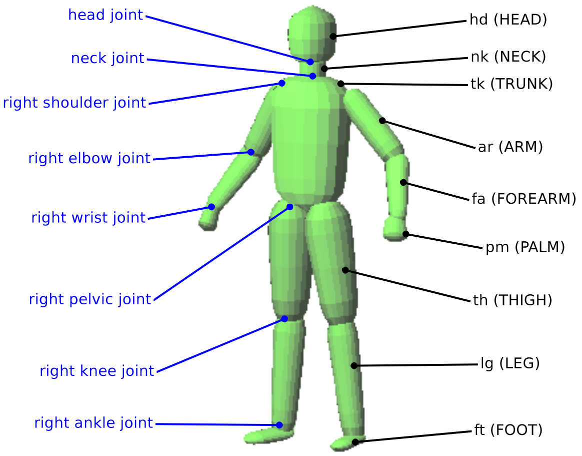


Figure 1. Super-quadric based human body model with the body segments and some joint locations labelled. The joints not labelled are the joints corresponding to the left side of the body.

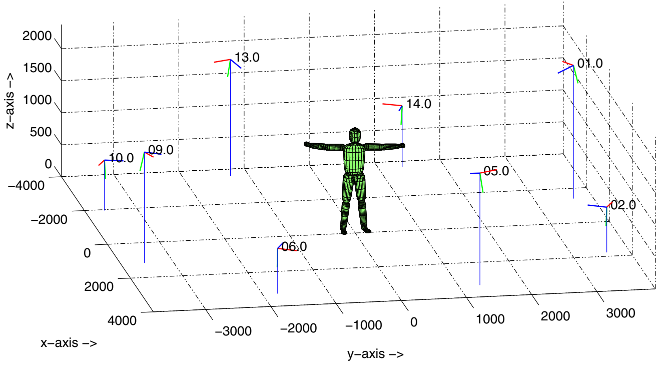
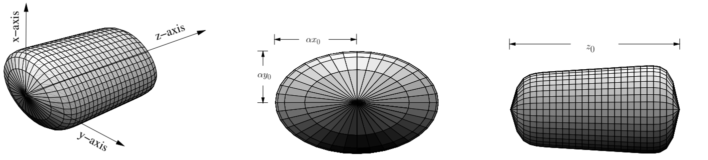


Figure 2. Camera configuration (mm units)



(a) Super-quadric (b) xy plane cross-section (c) Profile along z-axis

Figure 3. Modified super-quadric parameters

We note that the human body model introduced above allows us to represent the human body in a large set of postures and is yet simple enough to describe in terms of relatively few parameters. We use tapered super-quadrics in order to represent the different body segments for their simplicity and flexibility. They capture the shape of different body segments using just five intuitive parameters in our case. The compact representation we use makes it easy to estimate the parameters of the super-quadrics. It is our objective to eventually replace this model with rigid triangular mesh models for each body segment. Estimation of such mesh objects is generally difficult unless accurate 3-D scans are available. The tapered super-quadric is given by

$$\left(\frac{x}{x_0}\right)^2 + \left(\frac{y}{y_0}\right)^2 = \left(1 + s\frac{z}{z_0}\right)\left(1 - \left(1 - 2\frac{z}{z_0}\right)^d\right),$$

where z takes values in $[0, z_0]$, and is characterised by the parameters x_0, y_0, z_0, d, and s. The meaning of some of the parameters is illustrated in Figure 3. If sliced in a plane parallel to the xy plane, the cross section is an ellipse with parameters αx_0 and αy_0 as shown in Figure 3 (b), where α is a scalar. The length of the segment is z_0 as shown in Figure 3 (c). The scale parameter, s, denotes the amount of taper, and the exponential parameter, d, denotes the curvature of the radius profile along the z-axis. For e.g., $d = 2, s = 0$, is an ellipsoid, $d = \infty, s = 0$ is a right-elliptical cylinder and $d = \infty, s = -1$ is a right-elliptical cone.

A joint between two body segments is described as a vector in the coordinate frame of the parent body segment, connecting the origin of the parent segment coordinate frame to the origin of the child segment. The pose of the child segment is described in terms of the rotational parameters between the child coordinate frame and the parent coordinate frame. The trunk, which forms the base of the kinematic chain, has 6 degree-of-freedom (DoF) motion, both translational and rotational. The body model includes the locations of the joints of the different body segments.

2.2: Capture environment

We use multiple calibrated cameras in our system. We position cameras all around the subject and pointing towards the centre of the capture space. However, due to the complex structure of the human body, there is bound to be some occlusion unless there are cameras that are positioned all around and pointing at the subject. Our system consists of cameras positioned around the room as illustrated in Figure 2 so as to obtain images of the subject from different angles. The specifications of the capture in the Keck laboratory, where we captured most of the data that was used in the experiments, are as follows.

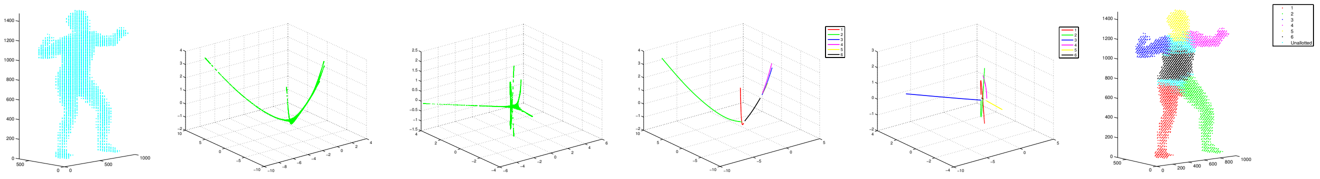
- The number of cameras used ranges from 8-16.
- The images are 484×648 grey-level with 8-bit depth.
- The frequency of the capture is 30 frames per second.

The cameras are calibrated using Tomas Svoboda's algorithm [45] that provides the intrinsic and extrinsic calibration parameters that are accurate to a scale. We then use a calibration device of known dimensions and use images from two cameras to obtain the scale parameter and a world reference frame. The camera (lenses) used in the Keck laboratory possess negligible radial distortion and we ignore the radial distortion parameters. If the radial distortion is not negligible, we undistort the images using the estimated radial distortion parameters.

3: Model acquisition

We present an algorithm that builds a complete articulated human body model using video obtained from multiple calibrated cameras. We use a bottom-up approach in order to build a parametric representation of a general articulated body from the voxel data. We, then, register the parametric representation with the known human body model, and estimate the parameters of the human body model.

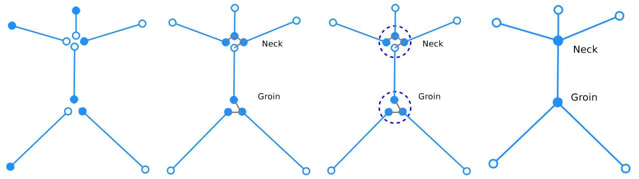
We base our algorithm on the observation that the human body consists of a base body (trunk) with articulated chains originating from it, such as the neck-head chain, arm-forearm-palm chain and the thigh-leg-foot chain (Refer Figure 1.) We build a voxel representation at each instance of time using the foreground silhouettes computed at that time instant. Our key observation is that the human body can be visualised as consisting of 1-D segments (or articulated chains) embedded in three-dimensional space. We note this in Figure 4 (a), where we can make out the five articulated chains, the head and four limbs, attached to the trunk. We would like to extract the 1-D structure of the voxel data, as well as successfully segment them into different articulated chains. The articulated nature of these chains, however, make it difficult to segment them in normal 3-D space. We adapt the method proposed by Belkin and Niyogi [5] to extract the geometric structure of the underlying 1-D manifold. In the first part of the algorithm, we segment the voxels into different articulated chains and also obtain the "position" of the voxel along the articulated chain. We transform the voxels into the Laplacian Eigenspace of six dimensions (Figure 4 (b-c)), where we can segment the voxels by fitting splines to the voxels in eigenspace the results of which are shown in Figure 4 (d-f). We are able to obtain for each voxel, a parameter describing its location on the spline representing



(a) 3D space (b) Eigenspace (dim. 1-3) (c) (dim. 4-6) (d) Eigenspace (dim. 1-3) (e) (dim. 4-6) (f) 3-D space

Figure 4. Segmentation in Eigenspace: Splines are colour coded according to index.

that articulated segment. This voxel registration is a key step and used at various stages in the algorithm. Once we have obtained the six spline segments representing the six articulated chains, we can visualise each segment as an edge connecting two nodes. We thus have twelve nodes and six edges (Figure 5 (a)). We connect "close" nodes (Figure 5 (b)), merge them (Figure 5 (c)), and register to the graph structure of our human body model shown in Figure 5 (d).



(a) Unconnected (b) Add edges (c) Merge nodes (d) Human body

Figure 5. Computing body segment graph

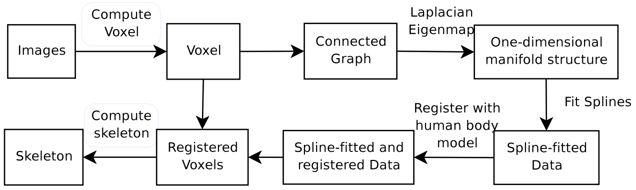


Figure 6. Segmentation of voxels into articulated chains and computation of skeleton

We, then, construct a skeletal representation as illustrated in the flow-chart in Figure 6. We may not be able to successfully register the voxels to the articulated chains at all time instances in case of error in the voxel estimation or in the case of a difficult pose. Therefore, while the method may not be useful in performing pose estimation at every frame, we are able to acquire the model and simultaneously estimate the pose of the human subject in a few key frames where the registration is successful.

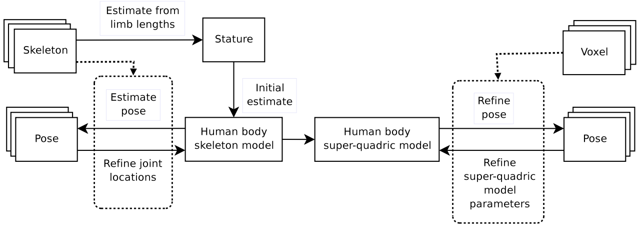

Skeleton
Estimate from limb lengths
Stature
Estimate pose
Initial estimate
Pose
Human body skeleton model
Refine joint locations
Human body super-quadric model
Voxel
Refine pose
Pose
Refine super-quadric model parameters

Figure 7. Flow-chart for estimating the model parameters from a set of key frames of computed skeletons and registered voxels.

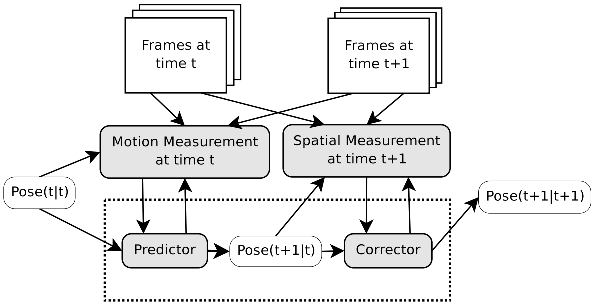


Figure 8. Overview of tracking algorithm

In the second part of our model acquisition algorithm, we obtain the model parameters for the subject progressing from a simple skeleton model at first to the complete super-quadric model. We obtain an initial estimate of the human body "skeleton" from the key frame skeleton and we optimise the human body model parameters and the pose of the key frames to minimise the model fitting error with the computed skeleton as illustrated in Figure 7. The human body model parameters are the joint locations and the shape (super-quadric) parameters of the different body segments. The important steps are as follows.

- Estimate the stature of the person from limb lengths, build a basic body skeleton for different values of the

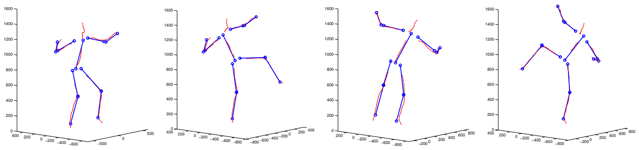
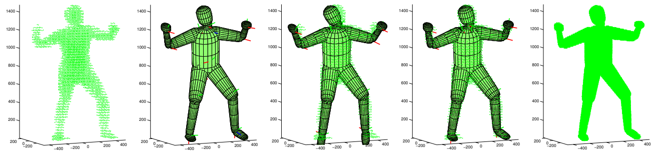


Figure 9. Optimised model superimposed on computed skeleton



(a) Voxels (b) Model (c) Init. pose (d) Fin. pose(e) Syn. Voxel

Figure 10. Estimated model and pose refinement

stature and find the model that best fits the key frame skeletons.

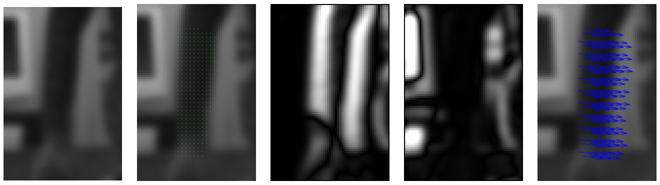
- Estimate the pose at each key frame that minimises the skeleton-model fitting error. Estimate the body model parameters (joint locations at this stage) that minimise the skeleton model fitting error. Iterate until the values converge numerically. (Figure 9)
- Estimate the parameters of the super-quadric body segments based on the currently existing body model parameters.
- Estimate the pose at each key frame that minimises the voxel-model fitting error. Estimate the body model parameters (super-quadric parameters) that minimise the voxel-model fitting error. Iterate until the values converge numerically. (Figure 10) The model (Figure 10 (b)) constructed from initial estimate of the quadratic parameters compared with the voxels (Figure 10 (a)), and super-imposed with voxels before pose refinement (Figure 10 (c)) and after (Figure 10 (d)). Figure 10 (e) is the model in voxel representation.

It is our objective to use our knowledge of the general structure of the human body in order to estimate the parameters of the human body model. In our model acquisition procedure we simultaneously estimate the parameters of the model as well as the pose. We could, of course, use this algorithm to estimate the pose at each frame, but, as noted earlier, the registration to the articulated chains at the voxel level may not succeed for all frames. However, we could estimate the pose at regular intervals and use tracking methods described in the next section to estimate the pose in the intermediate frames.

4: Tracking using multiple cues

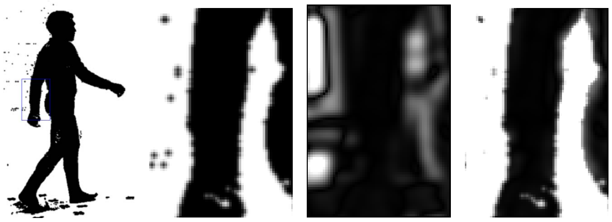
The algorithm we propose in [43] combines multiple cues, such as pixel displacements, silhouettes and "motion residues" to track the pose. The objective is to estimate the pose at time $t+1$ given the pose at time t, using the images at time t and $t+1$. The pose at $t+1$ is estimated in two steps, the prediction step and the correction step as illustrated in Figure 8. The motion information between time t and $t+1$ is used to predict the pose at time $t+1$ in the first step, while the spatial cues at time $t+1$ are used to correct the estimated pose. The steps required to estimate the pose at time $t+1$ are listed below.

- We register pixels to body segments and obtain the 3D coordinates at time t using the known pose at t. We convert each body segment into a triangular mesh and project it onto each image. We thus obtain the 3-D coordinates of some the pixels in the image corresponding to the mesh vertices. We can compute the 3-D coordinates and of all the pixels belonging to that segment by interpolation.
- Estimate pixel displacement between time t and time $t+1$ for each all pixels in the mask for each body segment and each image. (Figure 11)
- Predict pose at time $t+1$ using pixel displacement of pixels belonging to all body segments and in all images [44].
- Combine silhouette and "motion residue" (Figure 11 (d)) for each body segment into an "energy image" for each body segment and each image. (Figure 12)



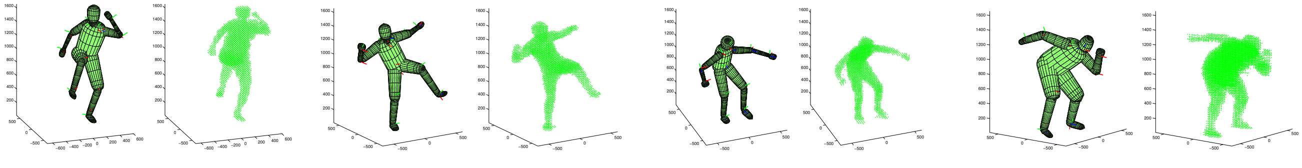
(a) Image (b) Mask (c) Image diff. (d) MR (e) Flow

Figure 11. (a) Smoothed image, (b) Mask, (c) Difference between images at time t and $t + 1$ (d) Motion residue (MR) for computed pixel motion (e) Estimated Pixel displacement



(a) Silhouette (b) Forearm (c) Motion Residue (d) Energy

Figure 12. Obtaining unified energy image for the forearm: (b), (c) and (d) represent the magnified box in (a).



A: 184cm, 72kg B: 165cm, 64kg C: 172cm, 52kg D: 178cm, 115kg

Figure 13. Estimated models and corresponding voxels for different subjects.

- Correct the predicted pose at time $t+1$ using the "energy image" obtained in previous step using optimisation.

5: Experimental results and summary

In the experiments we used grey scale images from 8-16 cameras. The foreground silhouette is obtained using a simple background subtraction algorithm. We present the results of our experiments using the proposed algorithms for human body model parameter estimation and tracking in Section 5.1 and our tracking results in Section 5.1. We finally present the summary in Section 5.3.

5.1: Human body model parameter estimation

We used 16 calibrated cameras in our experiments. The background subtraction algorithm does not perform very well on grey-scale images and as a result the voxel reconstruction is of poor quality at times. The algorithm is fairly robust to such errors and rejects frames where registration fails due to missing body segments or when the pose is not suitable.

We conducted experiments on four male subjects with different body mass, stature and BMI (body mass index). The same algorithm parameters were used in all the cases. Twenty key frames (where registration was successful) were used to estimate the model parameters as well as the pose at each time instant. The results are illustrated in Figure 13. We constructed a synthetic voxel image for each of the key frames using the estimated model and pose. We use the synthetic voxels (illustrated in Figure 10 (e)) in order to evaluate the algorithm with respect to data voxels (Figure 10 (a)). We also acquire the model parameters from the synthetic voxels, so that we can compare the original and estimated poses. The pose errors were computed at 24 major joint angles as the absolute differences between the original and estimated values for all the key frames used in the model estimation algorithm. The mean and median errors in degrees were 6.9° and 2.1° respectively.

5.2: Pose Tracking

In the experiments performed, we used grey-scale images, with a spatial resolution of 648×484, from eight cameras. We present the results of the experiments that were conducted using different sequences. The subject performs motions that exercise several joint angles in the body. The initial pose was set manually. Our results show that using only motion cues for tracking causes the pose estimator to lose track eventually, as we are estimating only the *difference* in the pose and therefore the error accumulates.

This underlines the need for "correcting" the pose estimated using motion cues. We see that the "correction" step of the algorithm prevents drift in tracking. In Figure 14, we present results in which we have superimposed the images with the model assuming the estimated pose over the images obtained from two cameras. The length of the first sequence is 10 seconds (300 frames), during which there is considerable movement and bending of the arms and occlusion at various times in different cameras. We also provide results for a walking sequence. Experiments using other sequences also show successful tracking of all the body segments inasmuch as a visual inspection of the model on images from all cameras reveals.

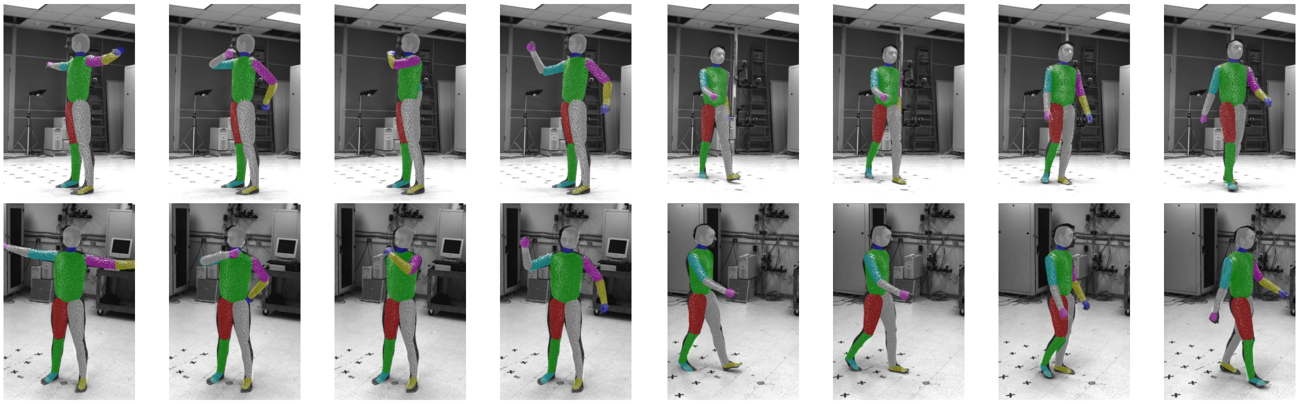


Figure 14. Tracking results using both motion and spatial cues. The images on the left belong to the first sequence and the images on the right belong to the second sequence.

5.3: Summary

We have addressed the problem of model acquisition in detail and provided the results of experiments on a single subject in different poses. No prior measurements of the subject were used. The only prior data used was a simple graph-based model of the human subject and an approximate relation between the stature of an average human subject and the length of the subject's long bones, as well as approximate locations of the shoulder, neck and pelvic joints with respect to the trunk. We performed experiments on four different subjects to check the success of the proposed algorithm on different human subjects. We also propose to compare the completely estimated model with the 3D-scan data of the same subject to verify accuracy of the computation. Once the model is available we can also accurately estimate the pose using the steps postulated in this section, without modifying the model parameters.

Our tracking algorithm uses multiple cues to perform robust and accurate tracking of pose using a complex human body model. We have made contributions at several levels, notably in the use of multiple cues, where we have combined the spatial cues in an intuitive manner. We use a mask for each body segment to compute the pixel displacement. We use a non-linear parametric model to compute the pixel displacement for the mask. We also allow for large values of displacement by using a combination of search and optimise algorithm. It is possible to reduce the degrees of freedom for certain joints such as the elbow joint. If we restrict the degrees of freedom of a joint, ideally the estimation is more robust and accurate. However, the estimation is very sensitive to the correct initialisation of the orientation axes. If there is a slight error in the definition of the axes, then the estimator may not be able to estimate the pose correctly. We, therefore, find it is better to not limit the degrees of freedom of special joints at this stage.

6: Acknowledgements

We thank Dr. Thomas Andriacchi, Dr. Lars Mündermann, Dr. Ajit Chaudhari and Dr. Stefano Corazza of the the Biomotion Laboratory at Stanford University for helpful comments and discussions. We are also thankful to James Sherman for his help in the data capture and calibration as well as Fritz McCall and Wesley Madoo of the UMIACS Computing Facilities for their help in the data capture at the Keck Laboratory. We are also thankful to James Sherman, Ashok Veeraraghavan and Anirudh Kembhavi for their contribution to the data collection process.

References

[1] J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[2] D. Anguelov, D. Koller, H. Pang, P. Srinivasan, and S. Thrun. Recovering articulated object models from 3-D range data. In *Uncertainty in Artificial Intelligence Conference*, 2004.

[3] D. Anguelov, L. Mündermann, and S. Corazza. An iterative closest point algorithm for tracking articulated models in 3-D range scans. In *Summer Bioengineering Conference, Vail*, 2005.

[4] N. I. Badler, C. B. Phillips, and B. L. Webber. *Simulating Humans*. Oxford University Press, Oxford, UK, 1993.

[5] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[6] Matthew Brand. Charting a manifold. In *Neural Information Processing Systems*, 2002.

[7] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, pages 8–15, 1998.

[8] Claudette Cedras and Mubarak Shah. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, March 1995.

[9] Tat-Jen Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *Computer Vision and Pattern Recognition*, volume 2, June 1999.

[10] Chi-Wei Chu, Odest Chadwicke Jenkins, and Maja J. Mataric. Markerless kinematic model and motion capture from volume sequences. In *CVPR (2)*, pages 475–482, 2003.

[11] Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with silhouettes. In *ICCV*, pages 716–721, 1999.

[12] David Demirdjian, T. Ko, and Trevor Darrell. Constraining human body tracking. In *ICCV*, pages 1071–1078, 2003.

[13] Jonathan Deutscher, Andrew Blake, and Ian Reid. Articulated body motion capture by annealed particle filtering. In *CVPR*, pages 2126–2133, 2000.

[14] Asi Elad and Ron Kimmel. On bending invariant signatures for surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1285–1295, 2003.

[15] Lars Mündermann et al. Conditions that influence the accuracy of anthropometric parameter estimation for human body segments using shape-from-silhouette. In *SPIE-IS and T Electronic Imaging*, volume 15, pages 268–277, 2005.

[16] Lars Mündermann et al. Estimation of the accuracy and precision of 3d human body kinematics using markerless motion capture and articulated icp. In *Summer Bioengineering Conference, Vail*, 2005.

[17] Lars Mündermann et al. Validation of a markerless motion capture system for the calculation of lower extremity kinematics. In *International Society of Biomechanics and American Society of Biomechanics, Cleveland*, 2005.

[18] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[19] D.M. Gavrila and L.S. Davis. 3-D model-based tracking of humans in action: A multi-view approach. In *Computer Vision and Pattern Recognition*, pages 73–80, 1996.

[20] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 38–44, 1996.

[21] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion. *IEEE PAMI*, 22(12):1453–1459, December 2000.

[22] I. A. Kakadiaris and D. Metaxas. 3D human body model acquisition from multiple views. In *Fifth International Conference on Computer Vision*, page 618. IEEE Computer Society, 1995.

[23] S. Baker K.M. Cheung and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 77–84, June 2003.

[24] N. Krahnstoever and R. Sharma. Articulated models from video. In *Computer Vision and Pattern Recognition*, pages 894–901, 2004.

[25] T. Tan L. Wang, W. Hu. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, March 2003.

[26] Xiangyang Lan and Daniel P. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *CVPR (1)*, pages 722–729, 2004.

[27] Ivana Mikic, Mohan Trivedi, Edward Hunter, and Pamela Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53(3):199–223, 2003.

[28] T.B. Moeslund and E. Granum. Multiple cues used in model-based human motion capture. In *International Conference on Face and Gesture Recognition*, 2000.

[29] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, pages 231–268, 2001.

[30] R. Plänkers and Pascal Fua. Articulated soft objects for video-based body modeling. In *ICCV*, pages 394–401, 2001.

[31] Deva Ramanan and David A. Forsyth. Finding and tracking people from the bottom up. In *CVPR (2)*, pages 467–474, 2003.

[32] J. M. Rehg and D.D. Morris. Singularity analysis for articulated object tracking. In *Computer Vision and Pattern Recognition*, pages 289–296, June 1998.

[33] Jim Rehg, Daniel D. Morris, and Takeo Kanade. Ambiguities in visual tracking of articulated objects using two- and three-dimensional models. *International Journal of Robotics Research*, 22(6):393 – 418, June 2003.

[34] K. Rohr. *Human Movement Analysis Based on Explicit Motion Models*. Kluwer Academic, 1997.

[35] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[36] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[37] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV*, pages 702–718, 2000.

[38] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard. Tracking loose-limbed people. In *CVPR*, pages 421–428, 2004.

[39] Leonid Sigal, Michael Isard, Benjamin H. Sigelman, and Michael J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *NIPS*, 2003.

[40] C. Sminchisescu and Bill Triggs. Kinematic jump processes for monocular 3D human tracking. In *International Conference on Computer Vision & Pattern Recognition*, pages I 69–76, June 2003.

[41] Cristian Sminchisescu and Bill Triggs. Covariance scaled sampling for monocular 3D body tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA*, volume 1, pages 447–454, Dec 2001.

[42] Aravind Sundaresan and Rama Chellappa. Markerless acquisition of articulated human body models using multiple cameras. In *European Conference on Computer Vision, Graz*, 2006. (submitted).

[43] Aravind Sundaresan and Rama Chellappa. Multi-camera tracking of articulated human motion using motion and shape. In *Asian Conference on Computer Vision, Hyderabad*, 2006. (accepted).

[44] Aravind Sundaresan, Amit RoyChowdhury, and Rama Chellappa. Multiple view tracking of human motion modelled by kinematic chains. In *International Conference on Image Processing, Singapore*, 2004.

[45] Tomáš Svoboda, Daniel Martinec, and Tomáš Pajdla. A convenient multi-camera self-calibration for virtual environments. *PRESENCE: Teleoperators and Virtual Environments*, 14(4), 2005. To appear.

[46] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[47] Christian Theobalt, Joel Carranza, Marcus A. Magnor, and Hans-Peter Seidel. Combining 3D flow fields with silhouette-based human motion capture for immersive video. *Graph. Models*, 66(6):333–351, 2004.

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

[49] M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *CVPR*, pages 664–665, 1991.

[50] M. Yamamoto, A. Sato, S. Kawada, T. Kondo, and Y. Osaki. Incremental tracking of human actions from multiple views. In *CVPR*, pages 2–7, 1998.