MARKERLESS FULL BODY SHAPE AND MOTION CAPTURE FROM VIDEO SEQUENCES

P. Fua[a], A. Gruen[b], N. D'Apuzzo[b], R. Plänkers[a]*

[a] VrLab, EPFL, 1015 Lausanne, [b] IGP, ETH-Hönggerberg, 8093 Zürich

KEY WORDS: Body Modeling, Motion Capture, Stereo, Silhouettes, Least-Squares Matching.

ABSTRACT

We develop a framework for 3–D shape and motion recovery of articulated deformable objects. We propose a formalism that incorporates the use of implicit surfaces into earlier robotics approaches that were designed to handle articulated structures. We demonstrate its effectiveness for human body modeling from video sequences. Our method is both robust and generic. It could easily be applied to other shape and motion recovery problems.

1 INTRODUCTION

Recently, many approaches to tracking and modeling articulated 3–D objects have been proposed. They have been used to capture people's motion in video sequences with potential applications to animation, surveillance, medicine, and man-machine interaction. See [Aggarwal and Cai, 1999, Gavrila, 1999, Moeslund and Granum, 2001] for recent reviews.

Such systems are promising. However, they typically use oversimplified models, such as cylinders or ellipsoids attached to articulated skeletons. Such models are too crude for precise recovery of both shape and motion. In our work, we have proposed a framework that retains the articulated skeleton but replaces the simple geometric primitives by soft objects. Each primitive defines a field function and the skin is taken to be a level set of the sum of these fields. This implicit surface formulation has the following advantages:

- **Effective use of stereo and silhouette data:** Defining surfaces implicitly allows us to define a distance function of data points to models that is both differentiable and computable without search.
- **Accurate shape description by a small number of parameters:** Varying a few dimensions yields models that can match different body shapes and allow both shape and motion recovery.
- **Explicit modeling of 3–D geometry:** Geometry can be taken into account to predict the expected location of image features and occluded areas, thereby making the extraction algorithm more robust.

Our approach, like many others, relies on optimization to deform the generic model so that it conforms to the data. This involves computing first and second order derivatives of the distance function of the model to the data points. This turns out to be prohibitively complex and slow if done in a brute-force fashion. The main contribution of this approach is a mathematical formalism that greatly simplifies these computations and allows a fast and robust implementation of articulated soft objects. It extends the traditional robotics approach that was designed to handle articulated bodies [Craig, 1989] and allows the use of implicit surfaces. For additional details, we refer the interested reader to our earlier publications [Plänkers and Fua, 2001, Plänkers and Fua, 2002].

We have integrated our formalism into a complete framework for tracking and modeling and demonstrate its robustness using video sequences of complex 3–D motions. To validate it, we focus on using stereo and silhouette data because they are complementary sources of information, as illustrated by Figure 1. Stereo works well on both textured clothes and bare skin for surfaces facing the camera but fails where the view direction and the surface normal is close to being orthogonal, which is exactly where silhouettes provide information. To increase the performance of our system, we have also developed an improved approach to extracting stereo-data using least-squares tracking methods.

In the remainder of this paper we first introduce our models. We then discuss our approach to extracting 3–D information from the video sequences and, finally, to fitting the 3–D body models to it.

2 ARTICULATED MODEL AND SURFACES

The human body model we use in this work [Thalmann *et al.*, 1996] is depicted by Figures 1(a,b). It incorporates a highly effective multi-layered approach

*This work was supported in part by the Swiss National Science Foundation.

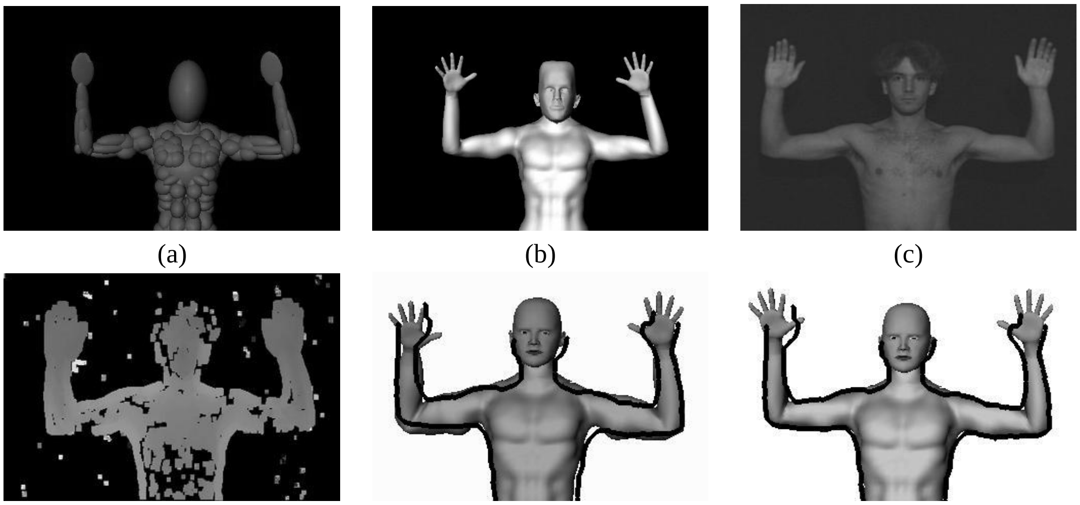

Figure 1: Models and silhouettes. (a) Metaballs attached to an articulated skeleton. (b) Skin surface computed by ray casting. (c) One image of a stereo pair used to estimate the parameters of the model in (b). (d) Corresponding disparity map. (e) The real body outlines overlaid on the skin surface. In this case the model was fitted using stereo only. As a result, it ends up too far from the actual data points and the system compensates by incorrectly enlarging the primitives. (f) Using the silhouettes during the fitting process provides stricter constraints that yield a better result.

for constructing and animating realistic human bodies. The first layer is a skeleton that is a connected set of segments, corresponding to limbs and joints. A joint is the intersection of two segments, which means it is a skeleton point around which the limb linked to that point may move.

Smooth implicit surfaces, also known as *metaballs* or *soft objects*, form the second layer [Blinn, 1982]. They are used to simulate the gross behavior of bone, muscle, and fat tissue. The metaballs are attached to the skeleton and arranged in an anatomically-based approximation. The head, hands and feet are explicit surfaces that are attached to the body. For display purposes a third layer, a polygonal skin surface, is constructed by ray casting [Thalmann *et al.*, 1996].

The body shape and position are controlled by a *state vector* Θ, which is a set of parameters controlling joint locations and limb sizes. In this section, we first describe this state vector in more detail and, then, our implicit surface formulation.

2.1 State Vector

Our goal is to use video-sequences to estimate our model's shape and derive its position in each frame. Let us therefore assume that we are given N consecutive video frames and introduce position parameters for each frame.

Let B be the number of body parts in our model. We assign to each body part a variable length and width coefficient. These dimensions change from person to person but we take them to be constant within a particular sequence. This constraint could be relaxed, for example to model muscular contraction.

The model's *shape* and *position* are then described by the combined state vector

$$\Theta = \{\Theta^w, \Theta^l,\ \Theta^r, \Theta^g\}\ \ , \tag{1}$$

where Θ is broken into sub-vectors that control the following model components:

- Shape
  - $\Theta^w = \{\theta^w_b \mid b = 1..B\}$, the width of body parts.
  - $\Theta^l = \{\theta^l_b \mid b = 1..B\}$, the length of body parts.
- Motion
  - $\Theta^r = \{\theta^r_{j,f} \mid j = 1..J, f = 1..N\}$, the rotational degree of freedom of joint j of the articulated skeleton for all frames f
  - $\Theta^g = \{\theta^g_f \mid f = 1..N\}$, the six parameters of global position and orientation of the model in the world frame for all frames f

The size and position of the metaballs is relative to the segment they are attached to. A length parameter not only specifies the length of a skeleton segment but also the shape of the attached metaballs in the direction of the segment. Width parameters only influence the metaballs' shape in the other directions.

2.2 Metaballs

Metaballs [Blinn, 1982] are generalized algebraic surfaces that are defined by a summation over n 3-dimensional Gaussian density distributions, each called a *primitive*. The final surface $\mathcal{S}$ is found where the density function F equals a threshold T, taken to be 0.5 in this work:

$$\begin{aligned} \mathcal{S} &= \left\{ [x,y,z]^T \in \mathbf{R}^3 \mid F(x,y,z) = T \right\} , & (2) \\ F(x,y,z) &= \sum_{i=1}^{n} f_i(x,y,z) , & (3) \\ f_i(x,y,z) &= exp(-2d_i(x,y,z)) , & (4) \end{aligned}$$

where d_i represents the algebraic ellipsoidal distance described below. For simplicity's sake, in the remainder of the paper, we will omit the i index for specific metaball sources wherever the context is unambiguous.

2.3 3–D Quadratic Distance Function

We use ellipsoidal primitives because they are simple and, at the same time, allow accurate modeling of human limbs with relatively few primitives because metaballs result in a smooth surface, thus keeping the number of parameters low. To express simply the transformations of these implicit surfaces that is caused by their attachment to an articulated skeleton, we write the ellipsoidal distance function d of Eq. 4 in matrix notation as follows. For a specific metaball and a state vector Θ, we define the 4×4 matrix

$$\mathbf{Q}_\Theta = \mathbf{L}_{\Theta^{w,l}} \cdot \mathbf{C}_{\Theta^{w,l}} \quad . \qquad (5)$$

where $\mathbf{L}$ and $\mathbf{C}$ are radii and position of the primitive respectively. The skeleton induced transformation $\mathbf{S}_\Theta$ is introduced as the rotation-translation matrix from the world frame to the frame to which the metaball is attached. These matrices will be formally defined in the appendix.

Given the $\mathbf{Q}_\Theta$ and $\mathbf{S}_\Theta$ matrices, we combine the quadric and the articulated skeleton transformations by writing the distance function of Eq. 3 as:

$$d(\mathbf{x}, \Theta) = \mathbf{x}^T \cdot \mathbf{S}_\Theta^T \cdot \mathbf{Q}_\Theta^T \cdot \mathbf{Q}_\Theta \cdot \mathbf{S}_\Theta \cdot \mathbf{x} \ . \qquad (6)$$

This formulation will prove key to effectively computing the Jacobians required to implement the optimization scheme of Section 4.

We can now compute the global field function F of Eq. 3 by plugging Eq. 6 into the individual field functions of Eq. 4 and adding up these fields for all primitives. In other words, the field function from which the model surface is derived can be expressed in terms of the $\mathbf{Q}_\Theta$ and $\mathbf{S}_\Theta$ matrices, and so can its derivatives as will be shown in the appendix. These matrices will therefore constitute the basic building blocks of our optimization scheme's implementation.

3 DATA ACQUISITION

From the video sequences, two kinds of 3–D information are extracted: A three dimensional surface measurement of the visible parts of the human body for each time step and 3–D trajectories of points on the body. The process consists of the following three steps:

- **Data Acquisition and Calibration:** The used image acquisition system consists of three synchronized progressive scan CCD cameras arranged in a triangle form in front of the subject. The cameras are connected to a frame grabber which digitizes the images at the resolution of 640x480 pixels with 8 bits quantization.

  The system is precalibrated using a 3–D reference frame with signalized points and then finely calibrated using thorough bundle adjustment techniques. The results of the calibration process are the exterior orientation of the cameras, the parameters of the interior orientation, the parameters for the symmetric radial and decentering distortion of the lenses and two additional parameters modeling differential scaling and shearing effects. A thorough determination of these parameters is required to achieve high accuracy in the measurement.

- **Matching Process and 3–D Point Cloud:** Our approach for the matching process [D'Apuzzo, 2002] is based on the adaptive least squares method with the additional geometrical constraint of the matched point to lie on the epipolar line. Starting from few seed points, the matcher produces a dense set of corresponding points relatively fast, e.g. on a Pentium III 600 MHz machine, about 20,000 points are matched in approximately 10 minutes.

  The seed points are generated automatically applying the Foerstner interest operator on the template image to determine points where the matching process may perform robustly; the corresponding points in the other two images are then established automatically by searching for the best matching results along the epipolar line.

The template image is then divided into polygonal regions according to which of the seed points is closest (Voronoi tessellation). Starting from the seed points, the set of corresponding points grows automatically by sequential horizontal and vertical shifts, until the entire polygonal region is covered. If the quality of the match is not satisfactory, the algorithm works adaptively by changing parameters (e.g. smaller shift, bigger size of the patch). The process is repeated for each polygonal region until the whole image is covered. The result is a dense set of corresponding points in the three images. The 3–D coordinates of the matched points are computed by forward ray intersection using the orientation and calibration data of the cameras. The mean achieved accuracy of the 3-D points is about 2 mm.

- **Surface Tracking:** The tracking process [D'Apuzzo *et al.*, 2000] is also based on least squares matching techniques. Its basic idea is to track triplets of corresponding points in the three images through the sequence and compute their 3–D trajectories. The spatial correspondences between the three images at the same time and the temporal correspondences between subsequent frames are determined with a least squares matching algorithm. The results of the tracking process are the coordinates of a point in the three images through the sequence, thus the 3–D trajectory is determined by computing the 3-D coordinates of the point at each time step by forward ray intersection. Velocities and accelerations are also computed.

  The tracking process is applied to all the points matched in the region of interest, resulting in a vector field of trajectories (position, velocity and acceleration), that can be checked for consistency and local uniformity of the movement. Key points can be defined and tracked in the vector field, producing 3–D information that can be used to establish the approximative posture of the body, e.g. position of joints.

Figure 2 depicts the output of this tracking process on one of the image triplets we work with.

4 MODEL FITTING

We use the body model of Section 2 both to track the human figure and to recover shape parameters. Our system is intended to run in batch mode, which means that we expect the two or more video sequences we use have been acquired before running our system. It goes through the following steps:

- **Initialization:** We initialize the model interactively in one frame of the sequence. The user has to enter the approximate position of some key joints, such as shoulders, elbows, hands, hips, knees and feet. Here, it was done by clicking on these features in two images and triangulating the corresponding points. This initialization gives us a rough shape, this is a scaling of the skeleton, and an approximate model pose. Techniques such as those proposed by [Barron and Kakadiaris, 2000, Taylor, 2000] could eliminate most of the currently necessary interaction.

- **Data Acquisition:** We use either clouds of 3–D points derived from the input stereo-pairs or triplets using either a simple correlation-based algorithm [Fua, 1993] or the higher quality data derived using the techniques introduced in Section 3. In the first case, the 3–D points form a noisy and irregular sampling of the underlying body surface. To reduce the size of the cloud and begin eliminating outliers, we robustly fit local surface patches to the raw 3–D points [Fua, 1997] and use the center of those patches as input to our system.

- **Frame-to-frame tracking:** At a given time step the *tracking* process adjusts the model's joint angles by minimizing, with respect to the joint angle values that relate to that frame, the distance of the model to the 3–D point clouds. This modified posture is saved for the current frame and serves as initialization for the next one. Optionally The system may use the model's projection into the images to derive initial silhouette estimates, optimize these using image gradients and derive from the results silhouette observations that it uses to constrain the minimization [Plänkers and Fua, 2002].

- **Global fitting:** The results from the *tracking* in all frames serve as initialization for global *fitting*. Its goal is to refine the postures in all frames and to adjust the skeleton and/or metaball parameters to make the model correspond more closely to the person. To this end, it optimizes over all frames simultaneously, again by minimizing the sane distance as before but, this time, with respect to the full state vector including the parameters that control the length and width of body parts.

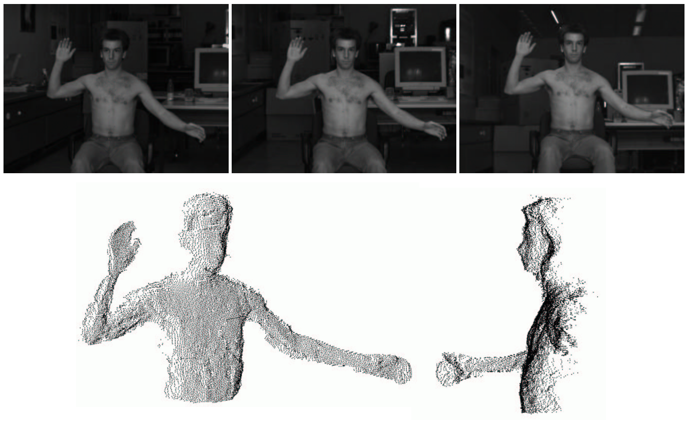

Figure 2: Top row: An image triplet. Bottom row: Measured 3–D point cloud.

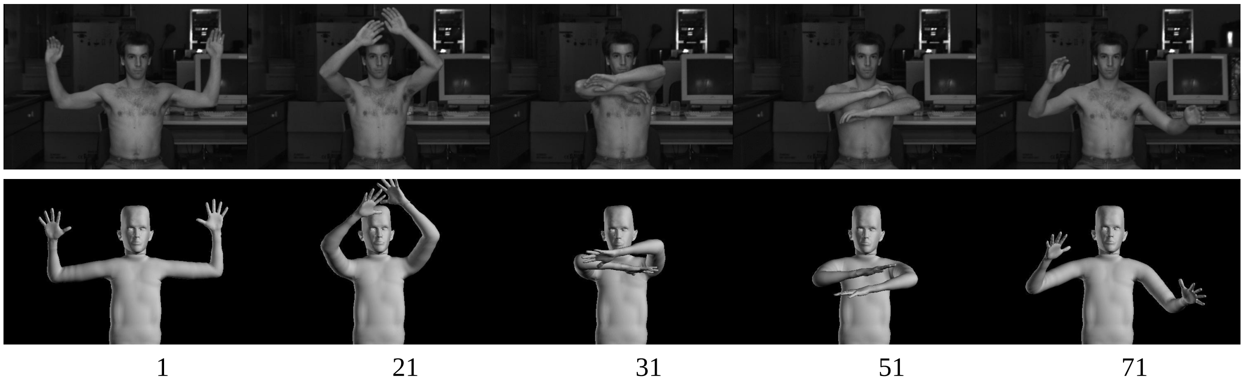

Figure 3: Tracking results in frames 1, 21, 31, 51, and 71 of a 300-frame sequence exhibiting a complex fully 3–Dimensional motion. Top row: Frames from one of three synchronized video sequences. Bottom row: Shaded represenation of the recovered model.

The final fitting step is required to correctly model the proportions of the skeleton and derive the exact position of the articulations inside the skin surface. This must be done over many frames and allows us find a configuration that conforms to every posture. To stabilize the optimization, we add to our objective function additional observations that favor constant angular speeds. Their weight is taken to be small so that they do not degrade the quality of the fit but, nevertheless, help avoid local minima in isolated frames and yield smoother and more realistic motions. Figure 3 depicts the results on a difficult fully 3–Dimensional motion.

5 CONCLUSION

In this work, we use a flexible framework for video-based modeling using articulated 3–D soft objects. The volumetric models we use are sophisticated enough to recover shape and simple enough to track motion using potentially noisy image data. This has allowed us to validate our approach using complex video-sequences featuring fully 3–Dimensional motions without engineering the environment or adding markers.

The implicit surface approach to modeling we advocate extends earlier robotics approaches designed to handle articulated bodies. It has a number of advantages for our purposes. First, it allows us to define a distance function from data points to models that is both differentiable and computable without search. Second, it lets us describe accurately both shape and motion using a fairly small number of parameters. Last, the explicit modeling of 3–D geometry lets us predict the expected location of image features such as silhouettes and occluded areas, thereby increasing the reliability of image-based algorithms.

Our approach relies on optimization to deform the generic model so that it conforms to the image data. This involves computing first and second derivatives of the distance function from model to data points. The main contribution of this paper is a mathematical formalism that greatly simplifies these computations and allows a fast and robust implementation. This is in many ways orthogonal to recent approaches to human body tracking as we address the question of how to best represent the human body for tracking and fitting purposes. The specific optimization scheme we use could easily be replaced by a more sophisticated one that incorporates statistics and can handle multiple hypotheses [Deutscher *et al.*, 2000, Davison *et al.*, 2001, Choo and Fleet, 2001]. Another natural extension of this work would be to develop better body and motion models: The current model constrains the shape and imposes joint angle limits. This is not quite enough under difficult circumstances: A complete model ought to also include more bio-mechanical constraints that dictate how body parts can move with respect to each other, for example in terms of dependencies between joint angles.

In our current work, we rely on cheap and easily installed video cameras to provide data. This, we hope, will lead to practical applications in the fields of medicine, athletics and entertainment. It would also be interesting to test our approach using high quality data coming from a new breed of image or laser-based dynamic 3–D scanners [Saito and Kanade, 1999, Davis *et al.*, 1999]. Our technique will provide the relative position of the skeleton inside the data and a standard joint angle based description of the subject's motion. Having high-resolution front and back data coverage of the subject should allow us to recover very high-quality animatable body models.

REFERENCES

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

[Barron and Kakadiaris, 2000]C. Barron and I. Kakadiaris. Estimating anthropometry and pose from a single image. In *Conference on Computer Vision and Pattern Recognition*, volume 1, Hilton Head Island, South Carolina, June 2000.

[Blinn, 1982]J. F. Blinn. A Generalization of Algebraic Surface Drawing. *ACM Transactions on Graphics*, 1(3):235–256, 1982.

[Choo and Fleet, 2001]K. Choo and D.J. Fleet. People tracking using hybrid monte carlo filtering. In *International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[Craig, 1989]J.J. Craig. *Introduction to robotics: mechanics and control*, chapter 5. Electrical and Computer Engineering. Addison-Wesley, 2nd edition, 1989.

[D'Apuzzo *et al.*, 2000]N. D'Apuzzo, R. Plänkers, and P. Fua. Least squares matching tracking algorithm for human body modeling. *International Archives of Photogrammetry and Remote Sensing*, 33(B5/1):164–171, 2000.

[D'Apuzzo, 2002]N. D'Apuzzo. Modeling human faces with multi-image photogrammetry. In *Three-Dimensional Image Capture and Applications V, Proc. of SPIE*, volume 4661, San Jose, USA, 2002.

[Davis *et al.*, 1999]L. Davis, E. Borovikov, R. Cutler, D. Harwood, and T. Horprasert. Multi-perspective analysis of human action. In *Third International Workshop on Cooperative Distributed Vision*, November 1999.

[Davison *et al.*, 2001]A. J. Davison, J. Deutscher, and I. D. Reid. Markerless motion capture of complex full-body movement for character animation. In *Eurographics Workshop on Computer Animation and Simulation*. Springer-Verlag LNCS, 2001.

[Deutscher *et al.*, 2000]J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *CVPR*, Hilton Head Island, SC, 2000.

[Fua, 1993]P. Fua. A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features. *Machine Vision and Applications*, 6(1):35–49, Winter 1993.

[Fua, 1997]P. Fua. From Multiple Stereo Views to Multiple 3–D Surfaces. *International Journal of Computer Vision*, 24(1):19–35, August 1997.

[Gavrila, 1999]D.M. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, 73(1), January 1999.

[Moeslund and Granum, 2001]T.B. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3), March 2001.

[Plänkers and Fua, 2001]R. Plänkers and P. Fua. Articulated Soft Objects for Video-based Body Modeling. In *International Conference on Computer Vision*, pages 394–401, Vancouver, Canada, July 2001.

[Plänkers and Fua, 2002]R. Plänkers and P. Fua. Model-Based Silhouette Extraction for Accurate People Tracking. In *European Conference on Computer Vision*, Copenhagen, Denmark, May 2002.

[Saito and Kanade, 1999]H. Saito and T. Kanade. Shape Reconstruction in Projective Grid Space from Large Number of Images. In *Conference on Computer Vision and Pattern Recognition*, Ft. Collins, CO, June 1999.

[Taylor, 2000]C.J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80(3):349–363, December 2000.

[Thalmann *et al.*, 1996]D. Thalmann, J. Shen, and E. Chauvineau. Fast Realistic Human Body Deformations for Animation and VR Applications. In *Computer Graphics International*, Pohang, Korea, June 1996.