# Articulated Soft Objects for Multi-View Shape and Motion Capture

Ralf Plänkers and Pascal Fua*
Computer Vision Lab
Swiss Federal Institute of Technology (EPFL)
1015 Lausanne, Switzerland
Email: Pascal.Fua@epfl.ch

IEEE PAMI, 25(10), 2003

**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 synchronized 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.

Such systems are promising. However, they typically use oversimplified models, such as cylinders or ellipsoids attached to articulated skeletons. These models are too crude for precise recovery of both shape and motion. We propose 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:

*This work was supported in part by the Swiss Federal Office for Education and Science under contract 99.0230-2 and in part by the Swiss National Science Foundation.

- **Effective use of stereo and silhouette data:** Defining surfaces implicitly allows us to define a distance function from data points to models that is both differentiable and computable without search.
- **Accurate shape description from 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:** The model can be projected into images to predict the expected location of image features and occluded areas, thereby making silhouette extraction more robust.

The *Articulated Soft Object Model* depicted by Figure 1(a-d) is at the root of our approach. It is equally applicable to other vertebrates, such as the horse and cow of Figure 1(e), and, more generally, to non-polyhedral articulated objects.

The main contribution of this paper is to show that the articulated implicit surface formalism is powerful enough to yield qualitatively good results under very difficult conditions, even in the absence of the sophisticated probabilistic frameworks that many other approaches rely on. Furthermore, it is amenable to a mathematically elegant and simple implementation by extending well known robotics results [7].

We integrate our formalism into a complete framework for tracking and modeling and demonstrate its robustness using trinocular 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. 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 robust information.

In the remainder of this paper, we first describe related approaches to articulated shape and motion recovery. We then describe our model in more detail and introduce our optimization framework. Finally, we present reconstruction results on complex human motions.

# 2 Related Work

Modeling humans from images involves recovering both body-shape and motion. Most existing approaches can be classified as addressing one or the other of these two problems, which is what we do in this section. However, in the following sections, we will argue that these two issues are intimately connected and one of the original features of our approach is that it simultaneously captures both shape and motion.

## 2.1 Human Shape Reconstruction

Digital photogrammetry and structured light have long been used to quickly acquire the stereo information needed to model live subjects. By using large numbers of cameras and space carving techniques [28, 8], it is possible to go even further and to capture the shape at frame-rate. However, these approaches mainly focus on re-rendering captured scenes from new viewpoints using disparity maps and multiple textures to interpolate between observed viewpoints. This is in part because the data produced in this way cannot be directly animated to create new motions: Animating the body shape requires partitioning it into body parts and attaching those parts to an articulated skeleton [2, 13, 30].

At the other end of the cost and complexity range are methods that accept simple photographs as input [18, 21]: They use silhouettes to deform generic 3-D models and create 3–D facsimiles of individual people that are suitable to represent articulated movement in a virtual world. Such systems are low-cost but do not produce a realistic body shape for a specific individual. Instead, they rely on texture mapping to hide the deficiencies of the approximated shape.

## 2.2 Human Motion Capture

In recent years there has been much interest in capturing complex motions solely by analyzing video sequences. Single camera solutions such as [4, 32, 29] would be ideal to process standard image sequences. However, they are not always robust, in part because image data is inherently noisy and in part because it is inherently ambiguous [24].

In contrast, using multiple cameras leads to a considerable reduction in the size of the search space and a considerable improvement in robustness, at the cost of having to deal with a large amount of data, most of which is redundant. Systems such as those proposed by [6, 10, 14], among many others, handle this problem effectively. Many of these approaches rely on sophisticated statistical models [12, 9, 5] to further reduce the size of the search space involved in modeling the whole human body. These approaches, however, 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 and it is this shortcoming that our approach addresses. The physics-based spring system model proposed in [19, 20] is one the rare exceptions to this modeling approach: Limbs are represented by volumetric primitives that are attached to one another by springs. Relatively loose connections between the body parts replace the usual underlying rigid structure. This allows for more flexibility in the model. However, with such a model it is difficult to constrain joint angles in order to enforce anatomically correct postures and the method has been mostly demonstrated on very clean silhouette data.

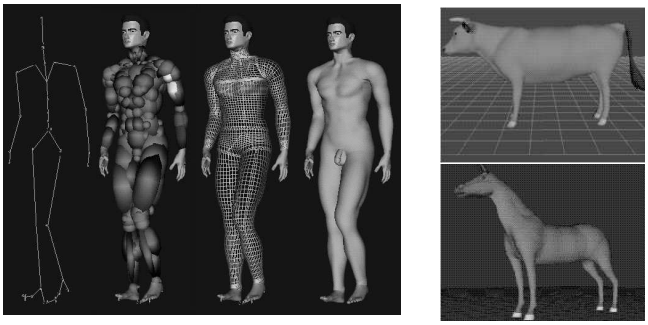
For a more detailed description of existing approaches, we refer the interested reader to recent surveys of visual motion capture research [1, 17, 23].

# 3 Articulated Soft Objects

The human body model depicted by Figure 1(a-d) is at the root of our approach. It was originally developed solely for animation purposes [31]. Smooth implicit surfaces, also known as *metaballs* or *soft objects* [3, 11], are attached to an articulated skeleton and are arranged in an anatomically-based approximation.

This particular human body model includes 230 metaballs. In order to prevent body parts from blending into each other, the body is segmented into ten distinct parts: Arms, legs and torso are split into upper and lower parts. When computing the implicit surfaces, only metaballs belonging to the same segments are taken into account. We constrain the left and right limbs to be symmetric. The head, hands and feet are explicit surfaces that are attached to the body. For display purposes, a polygonal skin surface is constructed via B-spline patches over control points computed by ray casting [31].

Our goal is to use video-sequences to estimate the model's shape and derive its position in each frame. To this end, in the following sections, we reformulate the animation model as an *articulated soft object*. We outline this formalism below and refer the interested reader to earlier publications [26, 25] for additional details.



(a) (b) (c) (d) (e)

Figure 1: Articulated soft objects: (a) Skeleton. (b) Volumetric primitives used to simulate muscles and fat tissue. (c) Polygonal surface representation of the skin. (d) Shaded rendering. (e) A cow and a horse modeled using the same technique.

## 3.1 State Vector

Body shape and position are controlled by a *state vector* $\Theta$, which is a set of parameters defining joint locations and limb sizes. We assign to each body part variable length and width coefficients. 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 motion is given by rotational degrees of freedom of the articulated skeleton's joints in all frames and by six parameters of global position and orientation.

## 3.2 Metaballs

The skin metaball surface $\mathcal{S}$ is a generalized algebraic surface that is defined as a level set of the summation over $n$ 3-dimensional Gaussian density distributions, each called a *primitive* [3]. $\mathcal{S}$ is the implicit surface defined by the level set $F(x, y, z) = L$:

$$\mathcal{S} = \left\{ [x, y, z] \in \mathbf{R}^3 \mid F(x, y, z) = L \right\} \quad (1)$$

$$F(x, y, z) = \sum_{i=1}^{n} f_i(x, y, z) \quad (2)$$

$$f_i(x, y, z) = \exp(-2d_i(x, y, z)) \ . \quad (3)$$

Here $d_i$ represents the algebraic ellipsoidal distance introduced in Eq. 4 and $L$ is taken to be 0.5. For simplicity, we omit the index $i$ for specific primitives wherever the context is unambiguous.

## 3.3 3–D Quadratic Distance Function

We use ellipsoidal metaballs because they allow accurate modeling of human limbs with relatively few primitives. To express the transformations of these implicit surfaces that is caused by their attachment to an articulated skeleton, we write the ellipsoidal distance functions $d$ of Eq. 3 as follows:

$$d(\mathbf{x}, \Theta) = \mathbf{x}^T \cdot \mathbf{S}_\Theta^T \cdot \mathbf{Q}_\Theta^T \cdot \mathbf{Q}_\Theta \cdot \mathbf{S}_\Theta \cdot \mathbf{x} \ , \quad (4)$$

where $\mathbf{Q}_\Theta$ and $\mathbf{S}_\Theta$ are $4 \times 4$ matrices that represent the metaball's shape and its skeleton-induced transformation respectively, and $\mathbf{x} = [x, y, z, 1]^T$ is a 3–D point.

More specifically, $\mathbf{Q}_\Theta$ defines the scaling and translation along the metaball's principal axes. The size and position of a metaball is relative to the segment it is attached to. A length parameter not only specifies the length of a skeleton segment but also the shape of all attached metaballs in the direction of the segment. Width parameters influence the metaballs' shape in the other directions. $\mathbf{S}_\Theta$ is a $4 \times 4$ rotation-translation matrix from the world frame to the frame to which the metaball is attached. It is defined as a series of matrix multiplications where each matrix corresponds to the transformation introduced by one joint.

We can now compute the global field function $F$ of Eq. 2 by plugging Eq. 4 into the individual field functions 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 [26].

# 4 Estimation Framework

The expected output of our system is the instantiated state vector $\Theta$ of Section 3.1 that describes the model's shape and motion. This is a highly non-linear problem because the model consists of an articulated set of implicit surfaces. As a result it contains rotations in Euclidean space as well as quadratic and exponential distance functions. Simplifying the volumetric models, replacing the perspective transform by an orthographic one, and using a different representation for rotational joints can be used to linearize parts of the problem [4]. Doing so, however, tends to lose in generality. Instead, we chose to solve full non-linear problem using the Levenberg-Marquart least squares estimator to minimize the distance between observations and model.

## 4.1 Least Squares Estimator

In practice, we use the image data to write $n_{\text{obs}}$ observation equations of the form

$$F(\mathbf{x}_i, \Theta) = L + \epsilon_i \ , \ 1 \leq i \leq n_{\text{obs}} \ , \tag{5}$$

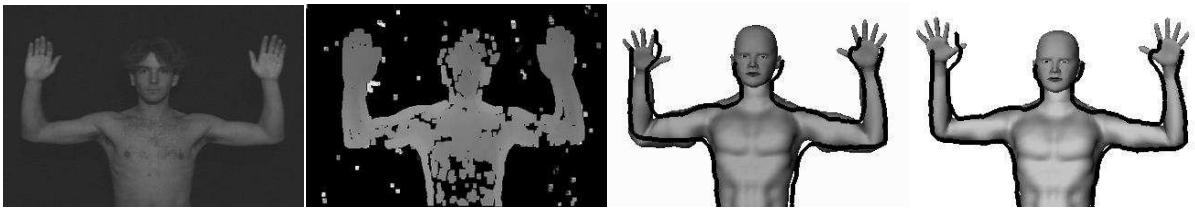
where $F$ is the global field function of Eq. 2, $L$ is the level value of Eq. 1, $\mathbf{x}_i$ is a data point, and $\epsilon_i$ is treated as an independently distributed gaussian error term. We then minimize $v^T P v$, where $v = [\epsilon_1, \ldots, \epsilon_{n_{\text{obs}}}]$ is the vector of residuals and $P$ is a diagonal weight matrix associated with the observations. Our system must be able to deal with observations coming from different sources, here stereo or silhouettes, that may not be commensurate with each other. We therefore use the following heuristic that has proved to be very effective. To ensure that the minimization proceeds smoothly, we multiply the weight $p_i^{\text{type}}$ of the $n_{\text{type}}$ individual observations of a given type by a global coefficient $c_{\text{type}}$ computed as follows:

$$\begin{aligned} G_{\text{type}} &= \frac{\sqrt{\sum_{1 \leq i \leq n_{\text{obs}}, j=\text{type}} \left\| \nabla_\Theta F(\mathbf{x}_i^{\text{type}}, \Theta) \right\|^2}}{n_{\text{type}}} \\ c_{\text{type}} &= \frac{\lambda_{\text{type}}}{G_{\text{type}}} \end{aligned} \tag{6}$$

where $\lambda_{\text{type}}$ is a user-supplied coefficient between 0 and 1 that indicates the relative importance of the various kinds of observations. This guarantees that, initially at least, the magnitudes of the gradient terms for the various types have appropriate relative values.

## 4.2 Data Constraints

In this work, we concentrate on combining stereo and silhouette data. Figure 2 illustrates their complementarity: In this example, we used a single stereo pair. In Figure 2(c) only stereo-data,



(a) (b) (c) (d)

Figure 2: The importance of silhouette information for shape modeling. (a) First image of a stereo pair. (b) Corresponding disparity map. (c,d) Two different fitting results. The black curves are the actual body outlines in (a). In (c) no silhouette constraints were used and the fitting puts the model too far away from the cloud. The system compensates by incorrectly enlarging the primitives. (d) depicts the result of the fitting using the silhouette constraints provided by the black outlines.

in the form of a cloud of 3–D points derived from the disparity map, was used. The stereo data is too noisy and shallow to sufficiently constrain the model. As a result, the fitting algorithm tends to move it too far away from the 3–D data and to compensate by inflating the arms to keep contact with the point cloud. This behavior is very similar to that observed when fitting ellipses to noisy 2–D points, especially in the case of non uniformly distributed points such as those obtained from stereo reconstruction [15]. Using the silhouettes in addition to the stereo data, however, sufficiently constrains the fitting problem to obtain the much improved result of Figure 2(d).

Because our field-function $F$ is both well-defined and differentiable, the observations and their derivatives can be computed both simply and without search using the matrix formalism of Section 3.3. We sketch these computations below and, again, refer the interested reader to earlier publications [26, 25] for additional details.

#### 4.2.1 3–D Point Observations

Disparity maps such as those of Figure 2 are used to compute clouds of noisy 3–D points. Each one is used to produce one observation of the kind described by Eq. 5. Minimizing the corresponding residuals tends to force the fitted surface to be as close as possible to these points. Because of the long range effect of the exponential field function in the error function $F$ of Eq. 2, the fitting succeeds even when the model is not very close to the data. Also, during least-squares optimization, an error measure that approaches zero instead of becoming ever greater with growing distance has the effect of filtering outliers.

To compute the Jacobian of the error function of Eq. 2, we must differentiate the individual field functions of Eq. 3. Derivatives with respect to parameter $\theta \in \Theta$ can be computed as:

$$\frac{\partial}{\partial\theta} f(d(\mathbf{x},\Theta)) = \frac{\partial f}{\partial d}\frac{\partial d}{\partial\theta} = -2\frac{\frac{\partial}{\partial\theta}d(\mathbf{x},\Theta)}{e^{2d(\mathbf{x},\Theta)}} \quad , \tag{7}$$

$$\frac{\partial}{\partial\theta} d(\mathbf{x},\Theta) = 2 * \mathbf{x}^T \cdot \mathbf{S}_\Theta^T \cdot \mathbf{Q}_\Theta^T \cdot \left[\frac{\partial}{\partial\theta}\mathbf{Q}_\Theta \cdot \mathbf{S}_\Theta\right] \cdot \mathbf{x} \quad , \tag{8}$$

where the $\left[\frac{\partial}{\partial\theta}\mathbf{Q}_\Theta\mathbf{S}_\Theta\right]$ term can be computed simply with a few matrix multiplications [26]. In short, these derivatives consist of modules that either are inexpensive to compute or need only be computed once because they are constant over all observations and constraints.

### 4.2.2 Silhouettes Observations

A silhouette point in the image defines a line of sight tangential to the surface. For any $\theta \in \Theta$, we define the implicit surface $\mathcal{S}(\theta) = \{[x, y, z] \in \mathbf{R}^3, F(x, y, z, \theta) = L\}$. Let $[x(\theta), y(\theta), z(\theta)]$ be the point on the line of sight where it is tangential to $\mathcal{S}(\theta)$. By definition, $[x(\theta), y(\theta), z(\theta)]$ satisfies two constraints:

1. The point is on the surface, therefore $F(x(\theta), y(\theta), z(\theta), \theta) = L$.
2. The normal to $S(\theta)$ is perpendicular to the line of sight at $[x(\theta), y(\theta), z(\theta)]$.

We integrate silhouette observations into our framework by performing an initial search along the line of sight to find the point $\mathbf{x}$ that is closest to the model in its current configuration. This point is used to add one of the observations described by Eq. 5. By construction, the point on the ray with the lowest field value satisfies the second constraint.

A change in model position or size induces a motion of $\mathbf{x}$ along the line of sight so that $\mathbf{x}$ remains the point closest to the model. This involves computing first and second order derivatives for the Jacobian entries of the form

$$\begin{aligned}
\frac{\partial^2 d}{\partial x_i \partial x_j} &= 2 * \frac{\partial \mathbf{x}}{\partial x_i}^T \cdot \mathbf{S}_\Theta^T \cdot \mathbf{Q}_\Theta^T \cdot \mathbf{Q}_\Theta \cdot \mathbf{S}_\Theta \cdot \frac{\partial \mathbf{x}}{\partial x_j} , \\
\frac{\partial^2 d}{\partial x_i \partial \theta} &= 2 * \mathbf{x}^T \cdot \left( \left[ \frac{\partial}{\partial \theta} \mathbf{S}_\Theta^T \cdot \mathbf{Q}_\Theta^T \right] \cdot \mathbf{Q}_\Theta \cdot \mathbf{S}_\Theta + \mathbf{S}_\Theta^T \cdot \mathbf{Q}_\Theta^T \cdot \left[ \frac{\partial}{\partial \theta} \mathbf{Q}_\Theta \cdot \mathbf{S}_\Theta \right] \right) \cdot \frac{\partial \mathbf{x}}{\partial x_i} ,
\end{aligned}$$

where $x_i, x_j$ represent translational degrees of freedom, while $\theta$ stands for one of the rotational ones [26]. Again this involves evaluating the same $\left[\frac{\partial}{\partial\theta}\mathbf{Q}_\Theta\mathbf{S}_\Theta\right]$ terms as before and, therefore, very little extra computation.

# 5 Implementation and Results

We initialize the model in one frame of the sequence by clicking on the approximate locations of several key joints in two images. This gives us a rough scaling of the skeleton and an approximate model pose. The system then goes through the following two steps.

1. **Frame-to-frame tracking:** For a given time step, the system extracts clouds of 3–D points from our synchronized input video sequences using a correlation-based approach [16]. These points are used to create observations of the kind described in Section 4.2.1. The *tracking* process adjusts the model's joint angles by minimizing the objective function of Section 4.1 with respect to the joint angle values relative to that frame. This modified posture serves as the 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 silhouette observations such as those introduced in Section 4.2.2.

2. **Global fitting:** The results from the *tracking* in all frames serve as initialization for global *fitting* to refine the postures in all frames and to adjust the skeleton and/or metaball parameters to improve the model's shape. To this end, it minimizes again the objective function of Section 4.1, over all frames and with respect to the full state vector.

Global fitting is required to correctly model the proportions of the skeleton and derive the exact position of the articulations inside the skin surface by simultaneously using as much information as possible. 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. In the examples shown below, optimization is performed on the 27 degrees of freedom that define the shape and motion of the upper body.

## 5.1 Using Stereo Alone

The images of Figure 3 are from 10-second sequences acquired by three progressive-scan cameras arranged in an inverted "L" configuration and capturing non-interlaced images at 30 Herz with an effective resolution of $640 \times 400$. These sequences feature highly complex motions of the subject's arms and upper body. Note the frequent self-occlusions as well as merging of body parts, such as grasping both hands in frame 110.[1]

Using our correlation-based approach, we created approximately 4000 3–D point observations per frame of the type described in Section 4.2.1. In the middle row, we show a shaded version of the recovered 3–D body model reprojected into the images. In the bottom row, we overlay the outline of this model on the original images. Note that in frame 100 the system places the left elbow too far from the body but recovers before frame 110. In effect the arm is "sliding" along the stereo data. In Section 5.2, we will show that using silhouettes during the reconstruction fixes this problem. In this example, neither hand nor head motion were modeled. Both are interpreted as rigid objects rigidly attached to neighboring body parts. This is why, in frame 40, the left hand is not at the right place, and, in frame 110, the model seems to look in the wrong direction when the subject moves his shoulder with respect to the head.

[1]Note to reviewers: Higher resolution versions of Figs. 3, 4, 5, and 6 will be made available as mpeg movies.

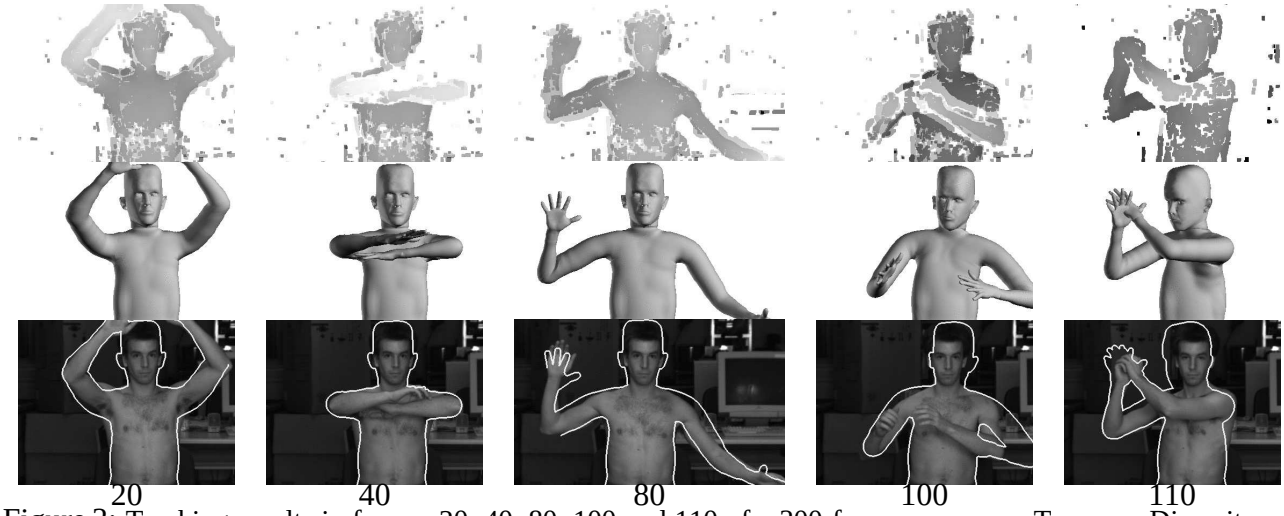



Figure 3: Tracking results in frames 20, 40, 80, 100 and 110 of a 300-frame sequence. Top row: Disparity maps. Middle row: Tracking and fitting results. Bottom row: Projections of the recovered model outline overlaid on the original images. The system correctly tracks until frame 80, misplaces the elbow in frame 100, and recovers by frame 110. As shown in Fig 6, using silhouette information will eliminate the error in the frame 100.

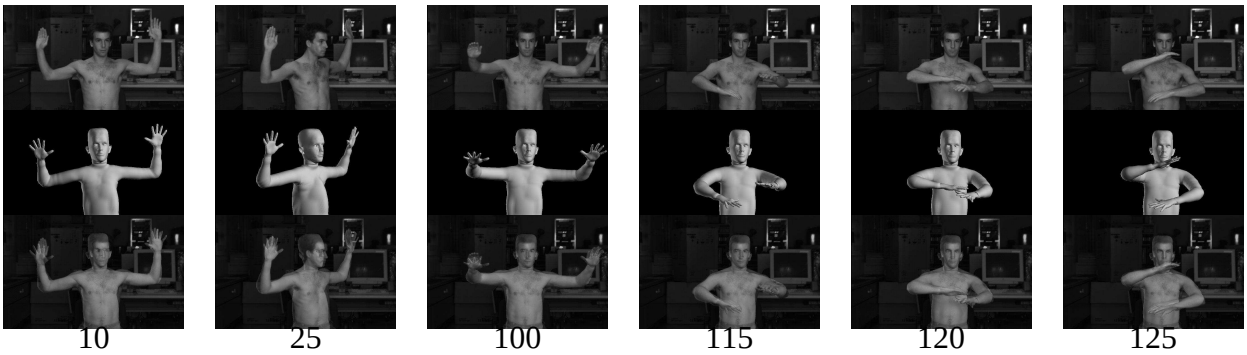



Figure 4: Selected frames of a trinocular sequence in which the subject performs complex upper body motions coupled with abrupt arm waving. Top row: One of the three original images at a given time. Center row: Reconstructed body model. Bottom row: Superposition of the original image and of the shaded model.

Figure 4 depicts the tracking of another very complex motion: In addition to the unpredictable arm movements, the torso bends and twists away from the cameras.

In Figure 5 we consider the case of somebody wearing a baggy sweater instead of being bare-chested. Edge-based methods would fail on such a sequence because the various deformations of the cloth hide the person's contour. The stereo data we obtain from the images is discriminating enough for our estimator to find the correct posture. In this case, we optimized only with respect to the motion parameters, the actual body shape being unobservable. Careful examination of those results, however, reveals some of the limitations of the shoulder model we use: A single ball-and-socket joint is convenient for motion estimation but it cannot capture realistic shoulder poses and deformations. A more anatomically correct model [22] is needed.

Figure 5: A few frames from a sequence in which the subject wears a baggy sweater, thus complicating the task of the tracker. Top row: One of the three original images at a given time. Bottom row: Tracking results.

## 5.2 Using Silhouette Information

At video rates, we can take the initial guess for the silhouette location to be the projected body outline in the previous frame. We then optimize this position using an active contour. However, in the presence of a cluttered and unknown background, even a good initial body-outline estimate does not guarantee convergence of an active contour to the correct solution. We therefore exploit our disparity maps to reject those gradients that are too far from areas whose distance from the cameras make them likely to be part of the subject and to correspond to actual silhouettes [27].

We then rerun our fitting algorithm on the 3–D stereo data augmented by this hypothesized silhouette outline. Even though the active contour may occasionally miss parts of the true body outline, thanks to the implicit surface treatment of outliers and to the strong stereo information, the system is usually able to find the correct model pose.

In a final refinement step, the model, now in its correct pose, is again projected into the camera frame and reoptimized. The full 3–D model is then fitted again using this improved silhouette and the stereo information. This approach yields the results of Figure 6. These results show that this new algorithm is able to overcome the errors that occurred when using stereo alone as depicted by Figure 3.

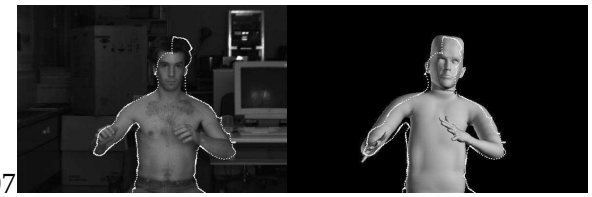
97  100 

Figure 6: Applying the model-based silhouette extraction method of Section 5.2 to the sequence of Figure 3. The tracking errors around frame 100 have been corrected by combining stereo and silhouette information. The snake-optimized contours are overlaid on the original images and on the reconstructed body model.

# 6 Conclusion

We have presented 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 noisy image data. This has allowed us to validate our approach using trinocular video-sequences featuring complex fully 3–Dimensional motions without engineering the environment or adding markers. Even though we use a straightforward approach to optimization instead of a sophisticated probabilistic one, our system can handle complicated motions that involve self-occlusions, temporary merging of body parts and noise introduced by clothing.

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. To this end, we have developed 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 [12, 9, 5]. 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 among 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 [28, 8]. 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


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

[2] E. Bittar, N. Tsingos, and M.-P. Gascuel. Automatic Reconstruction of Unstructured 3D data: Combining Medial Axis and Implicit Surfaces. In H.-P. Seidel and P. J. Willis, editors, *Computer Graphics forum*, volume 14 of *EUROGRAPHICS*, pages 457–468. Eurographics Ass., September 1995.

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

[4] Ch. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. In *Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

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

[6] M. Covell, A. Rahimi, M. harville, and T. Darrell. Articulated-Pose Estimation using Brightness and Depth-Constancy Constraints. In *CVPR*, Hilton Head Island, SC, 2000.

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

[8] 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.

[9] 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.

[10] Q. Delamarre and O. Faugeras. 3D Articulated Models and Multiview Tracking with Physical Forces. *Computer Vision and Image Understanding*, 81:328–357, March 2001.

[11] M. Desbrun and M.P. Gascuel. Animating Soft Substances with Implicit Surfaces. *Computer Graphics, SIGGRAPH Proceedings*, pages 287–290, 1995.

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

[13] I. Douros, L. Dekker, and B. Buxton. An Improved Algorithm for Reconstruction of the Surface of the Human Body from 3D Scanner Data Using Local B-Spline Patches. In *ICCV Workshop on Modeling People*, Corfu, Greece, September 1999.

[14] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[15] M. Fitzgibbon, A. W.and Pilu and R. B. Fisher. Direct least-squares fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):476–480, May 1999.

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

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

[18] A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun. Virtual People: Capturing Human Models to Populate Virtual Worlds. In *Computer Animation*, Geneva, Switzerland, May 1999.

[19] I. Kakadiaris and D. Metaxas. 3D Human Body Model Acquisition from Multiple Views. In *International Conference on Computer Vision*, 1995.

[20] I.A. Kakadiaris and D. Metaxas. Model based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. In *Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 1996.

[21] W.S. Lee, J. Gu, and N. Magnenat Thalmann. Generating Animatable 3D Virtual Humans from Photographs. In *Computer Graphics Forum, Eurographics*, pages C1–C10, Interlaken, Switzerland, August 2000.

[22] W. Maurel and D.Thalmann. Human Shoulder Modeling Including Scapulo-Thoracic Constraint and Joint Sinus Cones. *Computers and Graphics*, 24:203–218, 2001.

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

[24] D. Morris and J. Rehg. Singularity Analysis for Articulated Object Tracking. In *Conference on Computer Vision and Pattern Recognition*, pages 289–296, 1998.

[25] R. Plänkers. *Human Body Modeling from Video Sequences*. PhD thesis, EPFL, Lausanne, Switzerland, November 2001.

[26] 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.

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

[28] 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.

[29] C. Sminchisescu and B. Triggs. Covariance Scaled Sampling for Monocular 3D Body Tracking. In *Conference on Computer Vision and Pattern Recognition*, Hawaii, 2001.

[30] W. Sun, A. Hilton, R. Smith, and J. Illingworth. Building layered animation models of captured data. In *Eurographics Workshop on Computer Animation and Simulation*, Milano, Italy, 1999.

[31] 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.

[32] S. Wachter and H.-H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 73(3):174–192, June 1999.