# $M^3$: Marker-free Model Reconstruction and Motion Tracking from 3D Voxel Data

Edilson de Aguiar, Christian Theobalt, Marcus Magnor, Holger Theisel, Hans-Peter Seidel
MPI Informatik, Saarbrücken, Germany
{edeaguia|theobalt|magnor|theisel|hpseidel}@mpi-sb.mpg.de

## Abstract

*In computer animation, human motion capture from video is a widely used technique to acquire motion parameters. The acquisition process typically requires an intrusion into the scene in the form of optical markers which are used to estimate the parameters of motion as well as the kinematic structure of the performer. Marker-free optical motion capture approaches exist, but due to their dependence on a specific type of a priori model they can hardly be used to track other subjects, e.g. animals. To bridge the gap between the generality of marker-based methods and the applicability of marker-free methods we present a flexible non-intrusive approach that estimates both, a kinematic model and its parameters of motion from a sequence of voxel-volumes. The volume sequences are reconstructed from multi-view video data by means of a shape-from-silhouette technique. The described method is well-suited for but not limited to motion capture of human subjects.*

## 1. Introduction

In computer animation, the generation of life-like human characters has always been a challenging problem. The most important aspects of human character animation are the generation of human physical appearance and the realistic recreation of his motion. A variety of techniques has been developed to assist the animator in the latter task. Standard techniques for motion generation are keyframing, in which the animator specifies a set of key body poses to be interpolated, physics-based animation, in which the character's motion is simulated considering forces and torques, and motion capture. In the latter method, motion parameters are recorded from a real human actor performing. A variety of motion capture techniques has been developed, spanning from mechanical devices over electromagnetic approaches to optical systems. The majority of optical motion capture systems relies on reflective markers on the body and multiple high-speed high-resolution video cameras to estimate motion parameters. In many application scenarios, for example when texture and motion information should be acquired at same time (e.g. surveillance and 3D video), visual interference with the recorded scene is not desirable. Thus, a large variety of marker-free optical motion capture algorithms has been developed that do without any visual intrusion into the scene.

The techniques mentioned so far have in common that they depend on an a priori human body model. The most widely used model type is a predefined skeleton of the body that represents the underlying kinematics via joints and interconnecting bones. For marker-based approaches, it has been demonstrated that it is possible to estimate joint locations and bone hierarchies from the 3D marker trajectories [23]. This way, different moving subjects, humans and animals, can be tracked by the same technique without requiring complete manual redesign of the body model. Some marker-free capturing methods for humans can also estimate parts of the body structure semi-automatically using a priori information (e.g. [12]). However, these approaches fall short of the general case of arbitrary moving subjects.

In order to extend the class of non-intrusive algorithms with some of the flexibility provided by marker-based motion capture systems, we present a novel approach that

- estimates the kinematic structure of the moving subject without requiring significant a priori knowledge;
- tracks the motion of the subject using volume data that is reconstructed without the use of optical markers;
- is flexible enough to be applied to a large class of moving subjects.

Input to our system are sequences of voxel volumes that are reconstructed from multi-view video streams by means of a shape-from-silhouette approach. At each time step the volumes are subdivided by fitting ellipsoidal shells to the voxel data, thereby approximating the shape of the moving subject. Exploiting the temporal dimension, we can identify correspondences between ellipsoids over time and thus identify coherent rigid body parts. Knowing the motion of

the rigid bodies over time, the joint locations of the kinematic chain are estimated, and the motion parameters of the recorded subject are calculated based on the derived skeleton. We demonstrate the performance of the approach using volume sequences of a moving person recorded in our acquisition environment, and explain how the approach can be applied to a more general class of moving subjects, e.g. animals.

## 2. Related Work

Commercial human motion capture systems can be classified as mechanical, electromagnetic, or optical systems [18]. Video-based systems used in the industry typically require the person to wear optical markers on the body to whose 3D locations a kinematic skeleton is fitted [23]. A method for acquisition of a deformable human model using a combination of silhouette information and marker-based tracking is shown in [22]. Since in many application scenarios no visual intrusion into the scene is desired, researchers in computer vision have investigated marker-free optical methods [10]. Some of these methods work in 2D and represent the body by a probabilistic region model [27] or a stick figure [15]. More advanced algorithms employ a kinematic skeleton assembled of simple shape primitives, such as cylinders [21], ellipsoids [6], or superquadrics [11]. Inverse kinematics approaches linearly approximate the non-linear mapping from image to parameter space [3, 28] to compute model parameters directly. Analysis-through-synthesis methods search for optimal body parameters that minimize the misalignment between image and projected model. To assess the goodness-of-fit, features, such as image discontinuities, are typically extracted from the video frames [11]. A force field exerted by multiple image silhouettes aligns a 3D body model in Ref. [8]. In Ref. [20] a combination of stereo and silhouette fitting is used to estimate human motion. A hardware-accelerated silhouette-based motion estimation is described in Ref. [4], and in Ref. [9] a particle filter is applied to estimate body pose parameters from silhoutte views.

Recently, sequences of shape-from-silhouette (visual hull) models have been considered as input data for human motion estimation. Ellipsoidal body models [6], kinematic skeletons [17], or skeleton models with attached volume samples [26] are fitted to the volume data. Other visual hull-based approaches fit a pre-defined kinematic model with triangular mesh surface representation [2] to the volumes, or employ a Kalman Filter and primitive shapes for tracking [19].

All previously mentioned marker-free techniques rely on some form of pre-designed body model or require a significant amount of a priori knowledge to generate the model from the data in a semi-automatic procedure. In contrast, we present an approach that estimates the moving subject's kinematics and its motion in tandem, thereby enabling motion capture without prior information about the body structure. We achieve this by combining a volume decomposition technique based on ellipsoidal shells with a motion tracking of these primitive shapes which enables automatic marker-free motion capture.

The idea of characterizing 3D point clouds by means of fitting primitive shapes is a common approach in 3D shape analysis (see [16] for a survey) where it is typically applied to static data. In Ref. [7], multiple superquadric shapes are used to decompose 3D point data into primitive sub-shapes. The same category of geometric primitives is used in computer vision for object recognition, range map segmentation [14] and analysis of medical data sets [1].

Most similar to our approach is the work by Cheung et al. [5], where a person's skeleton and motion are estimated from visual hulls, and the work by Kakadiaris et al. [12] where body models are estimated from multiple silhouette images. Our method differs from these approaches in that it does not require a dedicated initialization phase where prescribed motion sequences are to be performed with each limb separately. Thus, our method requires far less a priori information about the tracked subject.

## 3. The Big Picture

In Fig. 1 an overview of the main algorithmic workflow of our method is shown. The system expects a voxel volume $V(t)$ for each time step $t$ of video as input. In step 1, the Ellipsoid Fitting step, each $V(t)$ is filled with ellipsoidal shells using a split and merge approach (Sect. 5). The result is a set of fitted ellipsoids $E(t)$ and a list of associated voxel subsets $S(t)$ for each time instant. The correspondences between ellipsoids at different time instants are established by means of a dynamic programming method in step 2, the Ellipsoid Matching step (Sect. 6). The result of step 2 is a path for each primitive shape that describes its motion over time. Together, all ellipsoid paths form the path set $P$. Knowing their motion, the primitives are clustered into separate rigid bodies in step 3, the Body Part Identification step (Sect. 7). After step 3, the motion of each rigid body over time is known, and joint locations between neighboring bodies can be estimated in step 4, the Skeleton Reconstruction step (Sect. 8). This step also enables estimation of body motion parameters based on the constructed skeleton model. Optionally, steps 1-3 may be iterated on subsets of the volume data (Sect. 9).

## 4. Voxel Data Acquisition

The video sequences used as input to our system are recorded in our multi-view video studio [25]. IEEE1394

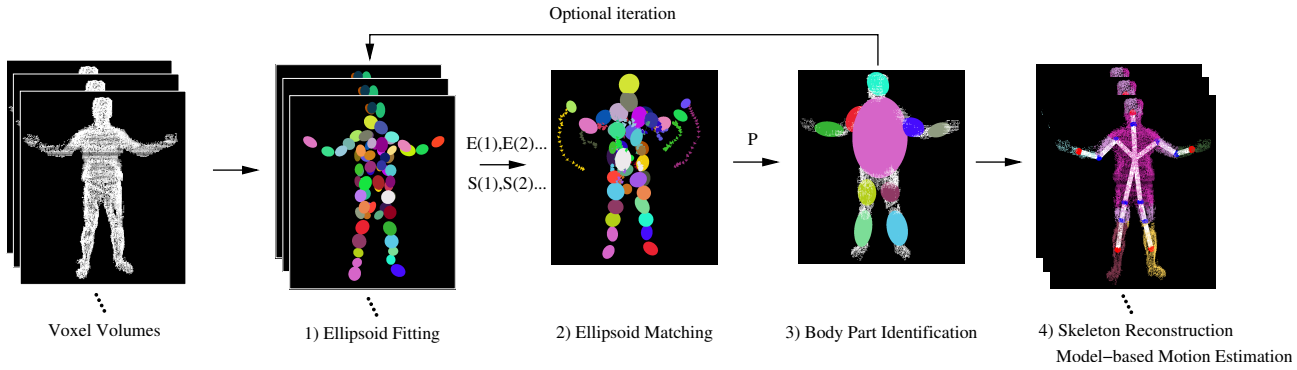


**Figure 1. Visualization of the individual processing steps. Steps 1-3 may optionally be iterated.**

cameras are placed in a convergent setup around the center of the scene. The video footage used in this paper is recorded from 8 static viewing positions arranged at approximately equal angles and distances around the center of the room. The cameras are synchronized via an external trigger, are recording at a resolution of 320x240 pixels and at a frame rate of 15 fps which is the technical limit for external synchronization. The cameras are metrically calibrated into a common coordinate system. In each video frame, the subject in the foreground is segmented via background subtraction, thereby creating silhouette images. From the silhouettes we reconstruct the voxel-based volume of the object in the foreground by means of a space-carving approach [13]. In addition to simple shape-from-silhouette reconstruction, this method employs a color-consistency criterion over multiple camera views to enhance the reconstruction quality. In our experiments, we carve surface voxel sets out of volume blocks of $256^3$ volume elements.

## 5. Ellipsoid Fitting

### 5.1. Fitting an Ellipsoid to Voxel Data

An ellipsoid is a closed surface defined as the solution of the implicit equation

$$F(x,y,z) = (\frac{x}{a_1})^2 + (\frac{y}{a_2})^2 + (\frac{z}{a_3})^2 \qquad (1)$$

where $a_1$, $a_2$ and $a_3$ are scaling factors along the three coordinate axes. Eq. 1 enables a simple test for deciding if a point $(x,y,z)$ lies inside $(F < 1)$, on the surface of $(F = 1)$, or outside $(F > 1)$ the primitive shape. An ellipsoid in a general position is described by three additional rotation parameters $(R_x, R_y, R_z)$ and three translation parameters $(T_x, T_y, T_z)$ with respect to the world origin. Thus, in order to fit an ellipsoid to a set of $N$ 3D points (in our case surface voxels) such that its surface comes as close as possible to all points nine shape parameters $[a_1, a_2, a_3, R_x, R_y, R_z, T_x, T_y, T_z]$ need to be determined. Using the following procedure we can robustly and quickly fit ellipsoids while avoiding a time-consuming numerical optimization. First, $T_x, T_y, T_z$ are found as the 3D location of the voxel set's center of gravity. The six remaining parameters are found via moment analysis [6], i.e. the directions of the main axes of variation in the 3D voxel set are found as the eigenvectors of the point set's covariance matrix. From these, the optimal radii $a_1, a_2, a_3$ and the optimal rotation parameters $R_x, R_y, R_z$ are derived.

This procedure computes an ellipsoidal fit very quickly, but it does not provide a direct measure of the fitting quality. Hence we calculate a fitting error (FE) $D$ that gives a numerical estimate of how well the ellipsoid approximates the point data:

$$D = \frac{\sqrt{a_1 a_2 a_3}}{N} \sum_{i=1}^{N} ||\overline{OP(i)}||_{rad} \cdot (F(x_i, y_i, z_i)^{\frac{1}{2}} - 1)^2 \qquad (2)$$

In Eq. 2 $||\overline{OP(i)}||_{rad}$ is the radial Euclidean distance between the $i$th point in the data set $P(i)$ and the intersection point of the line segment $OP(i)$ with the ellipsoid surface. Thus, an ellipsoid $\mathcal{E}$ is represented by 10 scalar values: $\mathcal{E} = [a_1, a_2, a_3, T_x, T_y, T_z, R_x, R_y, R_z, D]$.

### 5.2. Split and Merge

Using the method described in Sect. 5.1 for each time step, we fill the voxel volumes with ellipsoidal shells such that their total number and each individual ellipsoid's fitting error are as small as possible. We achieve this by applying a hierarchical *split and merge* approach [7]. The procedure starts with a split stage, approximating the whole voxel volume by one ellipsoid which is subdivided into two ellipsoids in case $D$ is greater than some threshold (Fig. 2). The

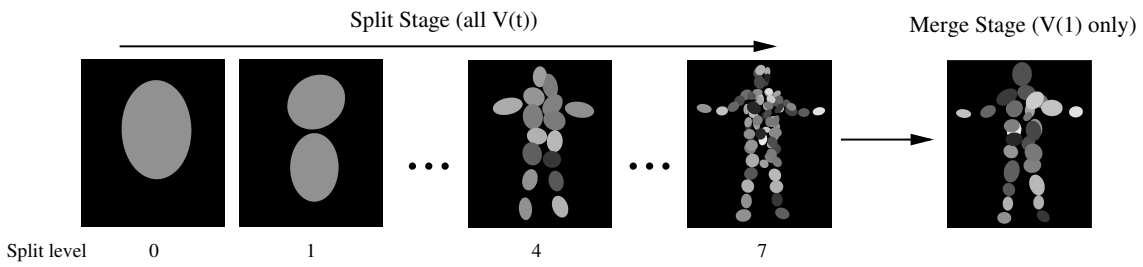


**Figure 2. Illustration of the split and merge procedure.**

split stage recursively processes each newly created ellipsoid in the same way, thereby producing a hierarchical decomposition of the voxel set. The split stage is performed for each voxel volume $V(t)$ individually.

The merge stage follows the split stage and improves the fitting result by merging pairs of neighboring ellipsoids into one. It is performed only for the voxel volume $V(1)$ of the first time step.

In the following the individual steps of the split stage and the merge stage are detailed.

**5.2.1. Split Stage** For each $V(t)$:

1 The whole set of 3D voxels $V(t)$ is approximated by one ellipsoid $\mathcal{E}$.

2 If the fitting error $D$ of $\mathcal{E}$ is less than some threshold $T_{SPLIT}$, the procedure stops. Otherwise, it proceeds to step 3.

3 The set of 3D voxels is split into two subsets $S_1$ and $S_2$ along the plane $\mathcal{P}$ orthogonal to the major inertial axis of the voxel set (Note that $\mathcal{P}$ contains the centroid of the set).

4 $S_1$ and $S_2$ are approximated individually by one ellipsoid each. For each subset, the procedure is repeated from step 2.

We obtain a set of ellipsoids $E_{split}(t)$ and a set of corresponding voxel subsets $S_{split}(t)$ that approximate the voxel model $V(t)$. After a sufficient number of subdivisions (in our case typically 7 and using a small $T_{SPLIT}$), there is a high likelihood that all points in one voxel subset belong to the same rigid body of the tracked subject's kinematic skeleton. Nonetheless, it is still possible that more than one ellipsoid is fitted to one rigid body (e.g. four ellipsoids to the upper arm), or that an ellipsoid was fitted to a position on the boundary between two adjacent rigid bodies (e.g. centered on the knee joint). In the latter case the voxel subset associated with the ellipsoid would belong to two different kinematic elements.

**5.2.2. Merge Stage** For $V(1)$ only:

1 For each subset of voxels $S_i \in S_{split}$, we determine the list $K_i = \{S_{n1}, .., S_{nk}\}$ of neighboring voxel subsets ($S_{n1}, .., S_{nk} \in S_{split}$).

2 For each possible pairing of the voxel set $S_i$ and one neighboring voxel set $S_j \in K_i$, a merged voxel set $M_j$ is created. A novel ellipsoid is fitted to each $M_j$ and a fitting error $D_j$ is computed. From all paired ellipsoids whose $D_j$ is below some threshold $T_{MERGE}$ the one with the lowest $D_j$ is chosen to replace the two ellipsoids it emerged from.

3 A new set of ellipsoids is obtained. The procedure is repeated from step 1. It terminates when all fitting errors are greater than $T_{MERGE}$.

We perform the merging step only on the first voxel volume $V(1)$. If we were considering voxel volumes from different time steps independently and merging ellipsoids only due to structural criteria, it would not be possible to prevent erroneous merges across rigid body boundaries. In addition, we keep the number of ellipsoids at each time step of the sequence constant. Merges in $V(1)$ that turn out to be implausible in later time steps of the motion sequence, are prevented by carefully tuning $T_{MERGE}$ at $t = 1$. The resulting set of ellipsoids is the starting point for the ellipsoid matching step (Sect 6) which exploits the temporal dimension to prevent merging across boundaries of separate bodies.

The result of the split and merge process is a set of ellipsoids $E(t)$ and a set of voxel subsets $S(t)$ for each $V(t)$.

## 6. Ellipsoid Matching

In this step a set of correspondences $C(t, t+1)$ between each pair of ellipsoid sets $E(t)$ and $E(t+1)$ from subsequent time steps is computed. The set of correspondences describes for each shape primitive in $E(t)$ to which member of $E(t+1)$ it is related. In other words, a correspondence for one ellipsoid tells us from which 3D position in $t$ to which location in $t+1$ it moves.

Assuming that we can keep the number of ellipsoids constant for all time instants, the correspondences enable the reconstruction of a complete motion path for each individual shape primitive over the duration of the whole input sequence. The ellipsoid matching procedure looks at each pair of ellipsoid sets $E(t)$ and $E(t+1)$ at subsequent time instants separately.

Since the number of shape primitives in the sets $E(t)$ and $E(t+1)$ may differ, we employ a two-stage procedure to establish the correspondences and to reorganize the ellipsoids such that their number at each time instant is constant. This way we establish a bijective correspondence mapping between ellipsoids at subsequent time steps.

In the first stage, a correspondence for each individual shape primitive is established to an ellipsoid at the subsequent time instant by means of a dynamic programming approach [24].

This optimization is based on an error function that is the weighted sum of two distance functions. The first distance value is the Euclidean distance between the ellipsoid centers. The second distance function is the absolute difference in the size of the voxel subsets that are associated with each of the two primitive shapes.

Dynamic programming establishes a first set of correspondences. Unfortunately, the first matching stage may lead to two cases of degenerate correspondences, that need to be corrected in the second stage (Fig. 3):

- Duplicated Ellipsoids: A duplicated ellipsoid occurs, for instance, if two or more ellipsoids from $E(t)$ are mapped to the same ellipsoid at $E(t+1)$. In this case, this ellipsoid at $E(t+1)$ is split and dynamic programming is applied again.
- Missing Correspondences: If the number of ellipsoids in $E(t+1)$ is initially greater than in $E(t)$, some ellipsoids in $E(t+1)$ are not assigned a partner in $E(t)$. In order to solve this problem, we merge ellipsoids in $E(t+1)$ without correspondences from $E(t)$ with the closest ellipsoid in $E(t+1)$ for which a correspondence has been established.

By this means, for each primitive in $E(t)$ exactly one partner from $E(t+1)$ is found. After processing all time steps in this way, each ellipsoid set contains the same number of shapes as the set $E(1)$. Note that in order to establish correct correspondences $C(t, t+1)$ the ellipsoid sets are altered as well. For each shape primitive in $E(1)$ a complete motion path over the whole sequence can be identified by linking subsequent correspondences. The so-created set of paths $P$ contains for each $\mathcal{E}_i \in E(1)$ a path $P_i$, $P_i$ being an ordered set of 3D coordinates $P_i = \{(x_i(t), y_i(t), z_i(t)) \mid t \text{ valid time step}\}$ of the ellipsoid center at time $t$.

Fig. 5(left) shows example paths that were reconstructed with this approach.

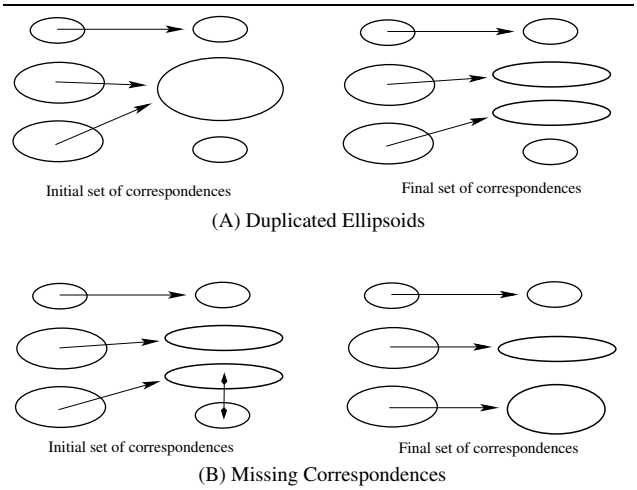


**Figure 3. Handling special cases during ellipsoid matching: Duplicate ellipsoids (top) and missing correspondences (bottom).**

## 7. Body Part Identification

The paths of $P$ provide all necessary information we need to identify separate rigid bodies in the kinematic skeleton of the tracked subject. In the case of tracking a human, this means that the paths enable us to identify, for example, the upper arm segment or the lower leg segment. Implicitly, we make the simplifying assumption that individual kinematic elements can be represented as rigid structures that do not undergo strong deformation.

In order to identify individual rigid bodies, we make use of the fact that the mutual Euclidean distance between any two points on the same body does not change while the skeleton is moving. Thus, if the mutual distance between the motion paths of two ellipsoids over time undergoes significant variation, it is most likely that the two primitives do not lie inside the same rigid body.

This criterion gives us a procedure at hand which enables clustering individual ellipsoids into separate kinematic elements. The procedure is based on the relative path of an ellipsoid $\mathcal{E}_k$, $P_r(\mathcal{E}_k)$. The relative path is obtained by subtracting the mean 3D position of the ellipsoid over time $(\overline{x_k}, \overline{y_k}, \overline{z_k})$ from each individual position along the path, $P_r(\mathcal{E}_k) = \{(x_k(t), y_k(t), z_k(t)) - (\overline{x_k}, \overline{y_k}, \overline{z_k}) \mid t \text{ valid time step}\}$. The relative path decouples information on motion variation from information about the location in space where the motion takes place. This way, comparison of ellipsoid motion is simplified.

In Fig. 4 (a) the paths of two ellipsoids belonging to the upper arm and forearm, respectively, are shown. The z-coordinates of their relative paths are also plotted (Fig. 4

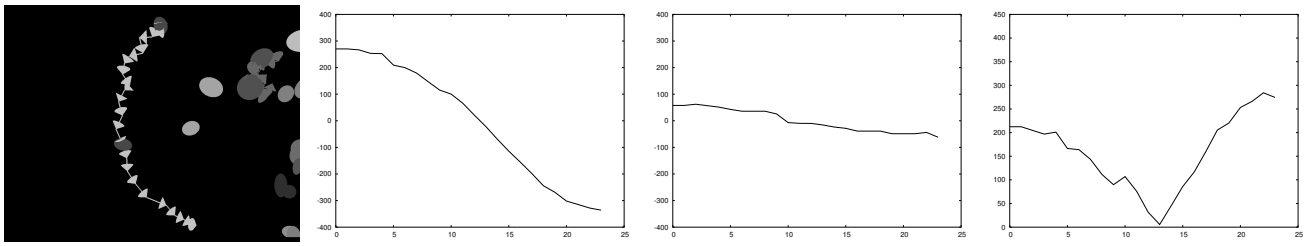

**Figure 4. Close-up of rendered motion paths of two ellipsoids (radii reduced for better visibility) in the forearm and upper arm respectively (a). Plots of z-coordinate values of the relative motion paths of each of the ellipsoids (b),(c). Graph of the distance curve for the ellipsoids' relative motion paths (d).**

(b),(c)). In order to decide whether the ellipsoids reside on the same rigid body we compute the distance curve between the relative paths (Fig. 4 (d)).

In detail, the steps which enable us to draw such a conclusion are as follows:

1 The relative path $P_r(\mathcal{E}_k)$ is calculated for each ellipsoid $\mathcal{E}_k \in E(1)$.

2 After selecting a seed ellipsoid $\mathcal{E}_{seed} \in E(1)$, a distance curve $DC_{seed,k}$ between $P_r(\mathcal{E}_{seed})$ and $P_r(\mathcal{E}_k)$ for each ellipsoid $\mathcal{E}_k \in E(1) \setminus \{\mathcal{E}_{seed}\}$ is computed. The value of $DC_{seed,k}$ at each time step is the Euclidean distance between the respective positions on the relative paths of $\mathcal{E}_{seed}$ and $\mathcal{E}_k$. One example is shown in Fig. 4 (d).

3 The integral of the difference curve $I(seed, k)$ is calculated to get an estimate of the area under the curve's plot. The value of $I(seed, k)$ indicates how similar the two paths are. Ellipsoid $\mathcal{E}_k$ is classified as belonging to the same body as $\mathcal{E}_{seed}$ if $I(seed, k) < T_{SB}$, $T_{SB}$ being a path similarity threshold.

4 The ellipsoids passing the test are assigned to the same rigid body as $\mathcal{E}_{seed}$.

5 The procedure iterates by restarting from step 2 and selecting a new seed from all ellipsoids that have not yet been assigned to a rigid body.

The seed $\mathcal{E}_{seed}$ in the first iteration is the ellipsoid nearest to the center of gravity (COG) of the voxel set $V(1)$. In the subsequent iterations, the selected seed is the ellipsoid nearest to the COG of the body part that was found in the preceding iteration. The threshold $T_{SB}$ used in the procedure is determined by performing statistical analysis on the set of integral values $I(seed, k)$. In case of a human subject this seed selection strategy leads to a subsequent identification of rigid bodies that belong to one limb (e.g. the arm). For each $V(t)$ it is now known which voxel subsets form a rigid body and how the rigid bodies move over time.

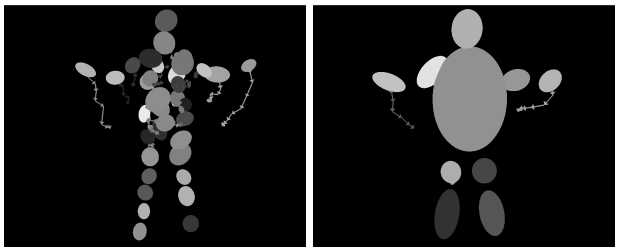

**Figure 5. Illustration of a few time steps of the computed motion paths for arm ellipsoids (left). Novel ellipsoids were fitted to identified rigid body segments (right).**

A novel ellipsoid is fitted to each such voxel subset (Fig. 5 (right)).

We achieved good results with this criteria for all the motion data we tested.

## 8. Skeleton Reconstruction

The final step of $M^3$ makes use of the detected rigid bodies and their motion to estimate the 3D locations of interconnecting joints. Joint finding is performed for each time step individually. Since we have no a priori information about which rigid bodies are connected by a joint in the skeleton hierarchy, we employ a heuristic approach to recover the connectivity. The temporal sequence in which the bodies were identified in the previous algorithmic step (Sect. 7) provides a clue on which adjacent rigid bodies are possibly connected. For each pair of potentially interconnected adjacent rigid bodies $B_a$ and $B_b$, the joint finding procedure is as follows:

1 A set of uniformly spaced point samples is created for $B_a$ and $B_b$. The point samples for each body are chosen such that they lie exactly on any of the three ma-

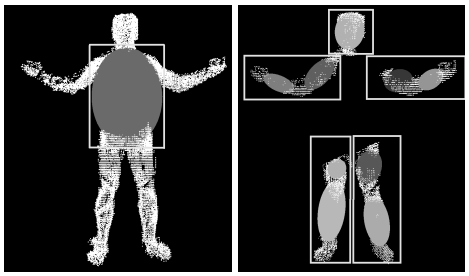

**Figure 6. After the first iteration the torso voxels (left) are identified and eliminated from each voxel set. Steps 1-3 of $M^3$ are then repeated on remaining isolated voxel subsets individually (right).**

jor axes of the ellipsoid which was fitted to $B_a$ and $B_b$ in the previous step of $M^3$.

2 For both ellipsoids: Using a growing strategy that starts in the ellipsoid center, the number of point samples on all three major axes is simultaneously increased (moving away from the center) until at least one of the samples lies inside the surface of the other ellipsoid respectively. This point (or in case of multiple points lying inside, their average 3D location) forms an estimate of the joint position between both bodies.

This is a simple but efficient approach which produced good results for our test data. Currently, we do not apply a temporal coherence criterion to stabilize the joint locations over time, hence the joint positions may jump. In a future version of the system, we plan to eliminate these artifacts by exploiting the temporal domain during joint localization.

Having a complete skeleton at each time instant, it becomes possible to describe the motion of the tracked subject in terms of the rotation parameters of the skeleton's joints and the translation of the root. Several example images of the reconstructed skeleton aligned with the moving body can be found in Fig. 7.

## 9. Results and Discussion

We evaluate the performance of our system using voxel data of a person performing simple gymnastics moves. The video footage was recorded in our multi-view video studio. Although the space carving approach eliminates most of the typical artifacts in shape-from-silhouette volumes that are due to insufficient visibility, some noise still appears in the form of bulky arms and legs. However, space carving is only one possibility to generate input data for for $M^3$. While simple visual hull reconstruction can run in real-time, space-carving takes, on average, 30 s on a Pentium IV 1.7 GHz to reconstruct the shape of the person per single time step from a $256^3$ voxel block. In total, the test data set contains 220 frames that were recorded at 15 fps.

A deterioration of tracking quality due to noise in the volume data can be prevented by applying an iterative variant of our method. In our experiments we applied an iterative implementation of $M^3$ in which the steps 1-3 are repeated (see Fig. 1). After each iteration, the largest rigid body is identified and, before the next iteration, all voxels belonging to this rigid body are eliminated from all volume data sets $V(t)$. Subsequently, steps 1-3 are applied in the same way to each newly found isolated voxel set. In the case of a human subject, this means that the first iteration identifies the torso segment, and in subsequent iterations, the algorithm proceeds with the arms, the legs and the head (Fig. 6).

The results in Fig. 7 show that our method is capable of reliably capturing the skeleton structure as well as the person's motion despite noise in the volume data. Due to the lack of ground truth data we assess the model estimation quality by visual inspection. The run-times of individual algorithmic components of our system are summarized in Tab. 1.

| | |
|---|---|
| split step | 6.5 s (single time step) |
| merge step | 13 s (first time step) |
| Body part identification | 51 s (whole sequence) |
| Skeleton reconstruction | 500 ms (single time step) |

**Table 1. Measured run-times of individual system components.**

An important advantage of our method over related approaches is that it estimates the body structure of the tracked subject with a minimum of a priori information. No special initialization motion is required to reconstruct the body model, any motion sequence is equally appropriate. One step in the algorithm where a knowledge-guided user interaction is feasible is the choice of the initial pose of the tracked subject. Although our experiments show that starting the system at different time instants of the test sequence produces similar results in terms of skeleton reconstruction, it is nonetheless desirable to have an initial body posture in which individual rigid bodies are not occluded.

Due to the significance of human motion data in computer animation we decided to demonstrate our approach using volume data of a moving person. We are convinced that our approach is equally suitable for a more general class of moving subjects that can be modeled by linked kinematic chains. Hence, we plan to present in the future results ob-

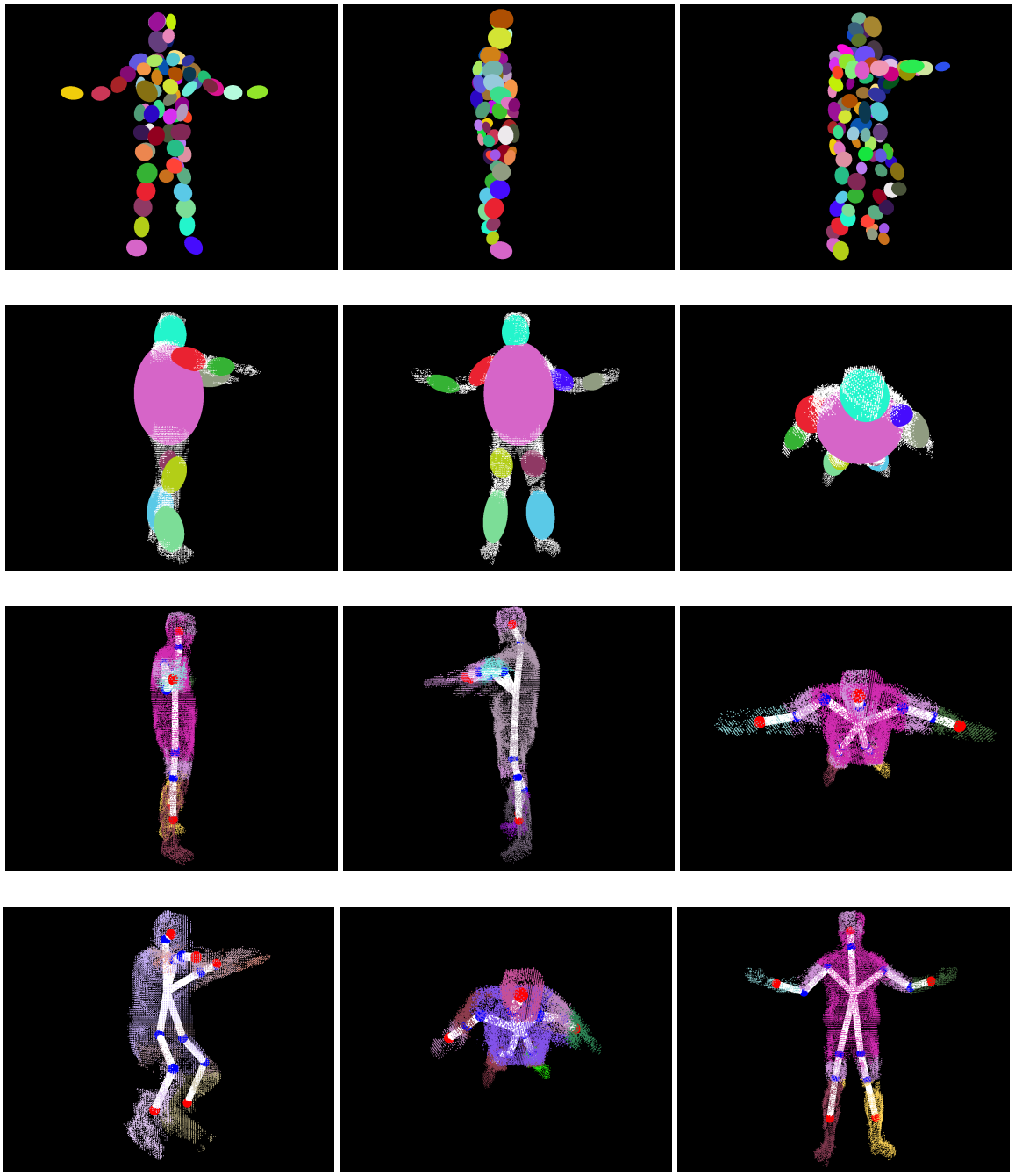

**Figure 7. Top row: Ellipsoids fitted to different body poses after the split stage. Second row: Discovered rigid bodies rendered as ellipsoidal shells inside the voxel volumes. Third and fourth row: Skeleton fitted to volume data at different time instants. Voxel sets belonging to different rigid bodies are drawn in different colors.**

tained from moving animals or moving mechanical structures.

In its current state, the system is subject to a couple of limitations. Even though we don't prescribe an initialization motion, two different adjacent rigid body segments can only be automatically identified if at least once in a sequence a relative motion between them can be observed. We consider this a principal problem of a non-informed tracking approach and not a limitation that is specific to our method. Furthermore, we expect that the system's performance will deteriorate if voxels of individual rigid bodies merge frequently with the rest of the volume (e.g. if the arms are often kept tight to the torso).

To conclude, we believe that, although $M^3$ does not operate on the same accuracy level as marker-based motion capture approaches, it is nonetheless a useful tool in situations where visual interference with the captured scene is inappropriate and no information about the structure of the tracked subject is available.

## 10. Conclusions and Future Work

We presented a novel approach for marker-free human motion capture that enables the simultaneous recovery of the kinematic chain and the motion parameters of a moving subject from volume data. It has been demonstrated that despite noise in the data, this novel approach robustly identifies the body structure of a moving person whose shape is reconstructed from silhouette images in multi-view video data. The algorithm is general enough to be applicable in similar form also to other moving subjects whose structure can be modeled as a linked kinematic chain, e.g. animals or mechanical devices.

In the future, we plan to enhance our split and merge procedure, as well as our joint localization approach, by further exploiting temporal coherence. In addition, we examine novel ways of improving our matching and body classification methods.

## References


[1] F. Banégas, M. Jaeger, D. Michelucci, and M. Roelens. The ellipsoidal skeleton in medical applications. In *Proceedings of the sixth ACM symposium on Solid modeling and applications*, pages 30–38. ACM Press, 2001.

[2] A. Bottino and A. Laurentini. A silhouette-based technique for the reconstruction of human movement. *CVIU*, 83:79–95, 2001.

[3] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. of CVPR 98*, pages 8–15, 1998.

[4] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In *Proceedings of SIGGRAPH2003*, pages 569–577, San Diego, USA, 2003. ACM.

[5] G. Cheung, B. S., and T. Kanada. Shape-from-silhouette of articulated objects and its use for human body kinematics estiamtion and motion capture. In *Proc. of CVPR*, 2003.

[6] K. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. In *Proc. of CVPR*, volume 2, pages 714–720, June 2000.

[7] L. Chevalier, F. Jaillet, and B. A. Segmentation and superquadric modeling of 3D objects. In *Proceedings of WSCG 2003*, 2003.

[8] Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with silhouettes. In *Proc. of ICCV 99*, pages 716–721, 1999.

[9] B. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proc. of CVPR'00*, 2000.

[10] D. Gavrila. The visual analysis of human movement. *CVIU*, 73(1):82–98, January 1999.

[11] D. Gavrila and L. Davis. 3D model-based tracking of humans in action: A multi-view approach. In *Proc. of CVPR 96*, pages 73–80, 1996.

[12] I. Kakadiaris and D. Metaxas. 3D human body model acquisition from multiple views. In *Proc. of ICCV'95*, pages 618–623, 1995.

[13] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *Int. J. Comput. Vision*, 38(3):199–218, 2000.

[14] A. Leonardis, A. Jaklic, and F. Solina. Superquadrics for segmenting and modeling range data. *IEEE PAMI*, 19(11):1289–1295, 1997.

[15] M. Leung and Y. Yang. First sight : A human body outline labeling system. *PAMI*, 17(4):359–379, 1995.

[16] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[17] J. Luck and D. Small. Real-time markerless motion tracking using linked kinematic chains. In *Proc. of CVPRIP02*, 2002.

[18] A. Menache. *Understanding Motion Capture for Computer Animation and Video Games*. Morgan Kaufmann, 1995.

[19] I. Mikić, M. Triverdi, E. Hunter, and P. Cosman. Articulated body posture estimation from multicamera voxel data. In *Proc. of CVPR*, 2001.

[20] R. Plaenkers and P. Fua. Tracking and modeling people in video sequences. *CVIU*, 81(3):285–302, March 2001.

[21] K. Rohr. Incremental recognition of pedestrians from image sequences. In *Proc. of CVPR 93*, pages 8–13, 1993.

[22] P. Sand, L. McMillan, and J. Popović. Continuous capture of skin deformation. *ACM Trans. Graph.*, 22(3):578–586, 2003.

[23] M.-C. Silaghi, R. Plaenkers, R. Boulic, P. Fua, and D. Thalmann. Local and global skeleton fitting techniques for optical motion capture. In *Modeling and Motion Capture Techniques for Virtual Environments*, number 1537 in LNAI, No1537, pages 26–40. Springer, 1998.

[24] M. Sniedovich. *Dynamic programming*. Marcel Dekker, Inc., 1992.

[25] C. Theobalt, M. Li, M. Magnor, and H.-P. Seidel. A flexible and versatile studio for synchronized multi-view video recording. In *Proceedings of Vision, Video and Graphics*, pages 9–16, 2003.

[26] C. Theobalt, M. Magnor, P. Schueler, and H.-P. Seidel. Combining 2D feature tracking and volume reconstruction for online video-based human motion capture. In *Proceedings of Pacific Graphics 2002*, pages 96–103, 2002.

[27] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, 1997.

[28] S. Yonemoto, D. Arita, and R. Taniguchi. Real-time human motion analysis and IK-based human figure control. In *Proceedings of IEEE Workshop on Human Motion*, pages 149–154, 2000. .