Automatic Learning of Articulated Skeletons from 3D Marker Trajectories

Edilson de Aguiar, Christian Theobalt, and Hans-Peter Seidel

MPI Informatik, Saarbrücken, Germany
{edeaguia, theobalt, hpseidel}@mpi-inf.mpg.de

Abstract. We present a novel fully-automatic approach for estimating an articulated skeleton of a moving subject and its motion from body marker trajectories that have been measured with an optical motion capture system. Our method does not require a priori information about the shape and proportions of the tracked subject, can be applied to arbitrary motion sequences, and renders dedicated initialization poses unnecessary. To serve this purpose, our algorithm first identifies individual rigid bodies by means of a variant of spectral clustering. Thereafter, it determines joint positions at each time step of motion through numerical optimization, reconstructs the skeleton topology, and finally enforces fixed bone length constraints. Through experiments, we demonstrate the robustness and efficiency of our algorithm and show that it outperforms related methods from the literature in terms of accuracy and speed.

1 Introduction

Marker-based optical motion capture (MOCAP) systems reconstruct the motion of moving subjects by measuring the 3D trajectories of optical beacons attached to the body [1,2,3]. In order to biomechanically analyze the motion of a person or in order to map real world performances onto virtual characters, the captured marker-trajectories have to be transformed into the motion parameters of a kinematic skeleton model. Although commercial tools exist that assist the motion capture professionals in performing this transformation, the estimation of kinematic skeletons and their motion parameters is still a labor-intensive, error prone and often inflexible process. Many commercial systems require the tracked subject to strike a dedicated initialization pose (T-pose) prior to actual motion recording or need specific initialization movements. Moreover, due to measurement noise in the marker-trajectories and non-rigid deformations of the body surface commercial software often fails to enforce fixed bone length constraints.

Despite the relevance of the skeleton reconstruction and joint parameter computation problem, astonishingly few papers have been published that aim at solving it in an automatic, robust, flexible and more efficient way than standard software packages. We present a new algorithm to estimate a skeleton model and its motion parameters that does not require a specific initialization pose, that relies on a minimum of a priori knowledge about the kinematic structure, and that reconstructs a model with fixed bone lengths from arbitrary motion sequences. Our main contributions are:

G. Bebis et al. (Eds.): ISVC 2006, LNCS 4291, pp. 485–494, 2006.

- A new method to identify individual rigid bodies from 3D marker trajectories.
- A method to determine joint positions at each time step of video, and to extract the topology of a skeleton with fixed bone lengths that optimally captures the true body pose at each time step.

The remainder of this paper is structured as follows: Sect. 2 reviews the most relevant related work. Sect. 3 details our clustering procedure that is used to identify individual rigid bodies from the markers' 3D motion. Sect. 4 details how correct joint positions are found for each time step of video and how the skeleton topology is automatically inferred. An optimal skeleton with fixed bone lengths is computed thereafter by the method described in Sect. 5. We have tested our method on a large number of publicly available motion capture sequences and compared it to most related methods from the literature. Furthermore, we have validated our method on synthetic sequences which provides us with accurate ground truth information about the model's kinematic structure, Sect. 6. The paper concludes in Sect. 7.

2 Related Work

Nowadays, marker-based motion capture systems have developed into a standard tool within the technical repertoire of professionals in computer animation and biomechanical analysis. Unfortunately, generating a moving kinematic skeleton model from raw marker trajectories with commercial tools is often still a semi-automatic procedure [1,2,3]. Commercial software frequently requires the use of body models with predefined topology making it hard to capture subjects which are not stored in the model database. Furthermore, many tools fall short of providing skeletons with constant bone lengths and the IK-based joint parameter estimation often does not produce satisfactory results.

Most algorithms from the literature aim at solving one particular sub-problem in the overall motion capture pipeline. Biomechanics researchers have developed several methods to accurately locate the joint of a subject from the motion of bones or markers [4,5,6]. Other approaches are able to solve the skeleton reconstruction problem by taking into account a priori information [7,8]. O'Brien et al. [9] present a technique for determining the joint parameters of an articulated skeleton hierarchy from magnetic tracking data. In their work, both position and orientation information of the markers are available, which simplifies the skeleton reconstruction procedure. In contrast, we present an automatic method for jointly estimating an articulated skeleton and its motion from marker trajectories. Our method does not impose any constraints on the type of motion or type of subject being captured.

Most similar to our approach are the methods by Silaghi et al. [10] and Kirk et al. [11]. Silaghi et al. describe a semi-automatic approach to locally find skeleton structures. An optimal skeleton is then assembled by matching a template to the different skeletons found over time. Although skeleton models can be reconstructed reliably, their method requires a substantial amount of user interaction.

Kirk et al. present an automatic approach for determining the kinematic structure of a subject when only a small number of markers have been attached to it. Also with this method, articulation structures can be reconstructed. However, the complexity of the involved optimization problem makes it hard to apply their algorithm to long motion capture sequences with many body markers.

Our method builds on and improves ideas from the literature. It enables us to automatically identify rigid bodies, to automatically compute joint positions and skeleton topology, and to automatically enforce fixed bone length constraints. It does not require any a priori information about the tracked subject, is computationally efficient, and the quality of the estimated skeletons matches the quality of body models that have been generated with commercial tools.

3 Rigid Body Clustering

The input to our system is raw optical MOCAP data, i.e. 3D marker trajectories that can be acquired with all commercial optical MOCAP systems available today. Although the positional marker tracking accuracy achievable today is very high, some noise in the measurements is unavoidable. It is also a common problem that, due to self-occlusions on the body, some of the markers are temporarily invisible or even completely lost. In a pre-processing step, we eliminate from the trajectory data all the markers that are not visible in all the frames. In principle, these markers could still be used for improving the quality of the skeleton reconstruction in a post-processing step (e.g. by the method proposed in Kirk et al.[11]). However, our experiments have shown that a robust rigid body identification is possible even if only a few complete marker trajectories are at our disposition.

The first step in our processing pipeline is to cluster markers into groups, each of them representing one rigid body part. To serve this purpose, we capitalize on the fact that the distance between any two markers on the same body part remains constant (within a measurement tolerance) over time, while it varies if they lie on different parts. To robustly decide which markers lie on the same body part, we employ a spectral clustering algorithm that examines the standard deviations of the mutual marker distances over time. We make use of a fast variant of spectral clustering that has proven its robustness on many point segmentation problems [12]. In our implementation we define the entries of the affinity matrix $\mathcal{A}$ as follows:

$$\mathcal{A}_{i,j} = \exp(-\rho_{i,j}/(2 * \sigma^2)), \tag{1}$$

where $\rho_{i,j}$ is the standard deviation of the mutual distance between markers i and j over all frames, and $\sigma = 1/N^2 * \Sigma_f(dist^f_{i,j})$ is a scaling term controlling the spectral clustering convergence. N is the number of frames and $dist^f_{i,j}$ is the distance between markers i and j in frame f. Intuitively, the affinity matrix encodes the likelihood of each pair of markers to belong to the same body segment. Instead of grouping the markers directly based on the individual values $\mathcal{A}_{i,j}$, spectral clustering uses the top eigenvectors of matrices derived from

$\mathcal{A}$ to cluster the markers. This leads to a more robust and kinematically more meaningful segmentation than, for instance, the application of simple K-means clustering [13]. As an additional benefit, the optimal number of clusters N can be automatically calculated based on the datasets eigen-gap. Fig. 2(left) shows that our approach robustly identifies individual segments in the human body.

4 Estimation of Joint Positions and Skeleton Hierarchy

Given a list of body segments and their associated markers, we now estimate the positions of interconnecting joints at each time step of a motion sequence, and thereafter reconstruct the topology of the interconnecting bone skeleton.

The method to achieve the first goal makes use of a relatively straightforward observation. If we assume that two rigid bodies are connected via a single three-degree-of-freedom (DOF) ball joint then the distance between each marker on either of the adjacent bodies and the common joint has to remain constant over time. Taking measurement noise and subtle non-rigid body deformations into account, a good estimate for the correct joint position sequence is the sequence of points that minimizes the variance in joint-to-marker distance for all markers of the adjacent parts at all frames.

Kirk et al. [11] put this criterion into practice by computing the joint positions between two interconnected segments at all time steps via solving a large optimization problem. However, their approach is only feasible for sequences where the number of frames N and the number of markers M are small, since an energy minimization in $N * M$ variables for each pair of segments is necessary. In contrast, we have developed a faster scheme which efficiently finds optimal skeletons even with sequences that are several thousand frames long and which feature several hundred markers.

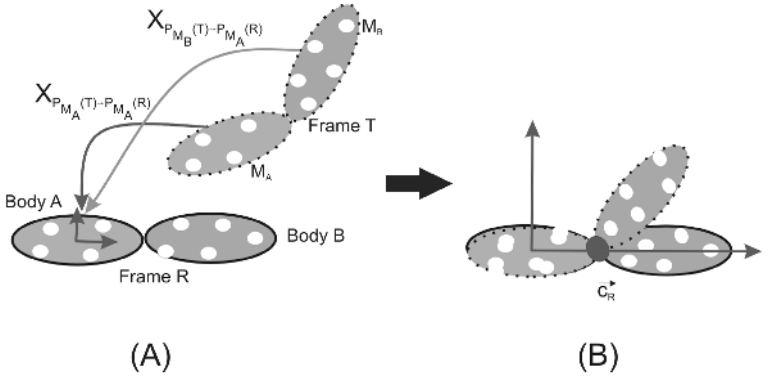

Fig. 1. Marker alignment: rigid body transformations are calculated (a) to align the position of markers for both segments in time step T with the markers of body segment A in the reference frame (b). After aligning the markers from all time steps the joint position $\boldsymbol{c_R}$ is found by minimizing (2).

Our scheme works as follows: Let A and B be two body segments, and let K be the set of the M body markers. Both body segments have associated sets of markers $M_A = \{a|a \in K\}$ and $M_B = \{b|b \in K\}$. At each time step f the markers in M_A and M_B have respective 3D positions $P_{M_A}(f) = \{\boldsymbol{p}(k,f)|k \in M_A\}$ and $P_{M_B}(f) = \{\boldsymbol{p}(k,f)|k \in M_B\}$. It is our goal to find the set $C = \{\boldsymbol{c_f}|f \in \{1,\ldots,N\}\}$, i.e. the set containing the 3D position $\boldsymbol{c_f}$ of the interconnecting joint at each time step. To this end, we define a reference frame number R which can be any of the frames but usually is the first frame of the sequence. First, for each time step $t \in \{0,\ldots,N\}$ we compute two rigid body transforms $X_{P_{M_A}(t)\rightarrow P_{M_A}(R)}$ and $X_{P_{M_B}(t)\rightarrow P_{M_A}(R)}$ that align the positions of the markers in both marker sets with the positions of the markers M_A at the reference time step [14], as shown in Fig. 1.

The positions of all markers at all time steps are now aligned with the marker positions at the reference time step. We are now able to solve for the joint location at the reference frame $\boldsymbol{c_R}$ by minimizing the following energy functional:

$$CF(\boldsymbol{c_R}) = 1/2 * \sum_{a \in M_A} (\sigma_a(\boldsymbol{c_R}) + \alpha * \overline{d}_a(\boldsymbol{c_R})) + 1/2 * \sum_{b \in M_B} (\sigma_b(\boldsymbol{c_R}) + \alpha * \overline{d}_b(\boldsymbol{c_R})) \quad (2)$$

where

$$\sigma_a(\boldsymbol{c_R}) = 1/N * \sum_{i=2}^{N} (\|\boldsymbol{c_R} - X_{P_{M_A}(i)\rightarrow P_{M_A}(R)} * \boldsymbol{p}(a,i)\| - \overline{d}_a(\boldsymbol{c_R}))^2 \quad (3)$$

and

$$\overline{d}_a(\boldsymbol{c_R}) = 1/N * \sum_{i=2}^{N} \|\boldsymbol{c_R} - X_{P_{M_A}(i)\rightarrow P_{M_A}(R)} * \boldsymbol{p}(a,i)\| \,. \quad (4)$$

The definitions of $\sigma_b(\boldsymbol{c_R})$ and $\overline{d}_b(\boldsymbol{c_R})$ correspond to (3) and (4). In (2), α is the coefficient that controls the influence of a distance penalty term. We employ the distance penalty term to prevent the algorithm from erroneously positioning the joint far away from either segment (e.g. infinitely away), where the variance $\sigma_a(\boldsymbol{c_R})$ and $\sigma_b(\boldsymbol{c_R})$ are minimal. Through experimental evaluation we have found that a value of $\alpha = 1/5$ leads to the best results. After finding $\boldsymbol{c_R}$, the joint position at all other frames can be computed by $\boldsymbol{c_f} = X^{-1}_{P_{M_A}(f)\rightarrow P_{M_A}(R)} * \boldsymbol{c_R}$.

Since we do not use a priori information about the topology of the subject, we perform the above procedure for each possible pair of body segments. Fortunately, the final values of the error term (2) enable us to automatically infer the skeleton topology and to discard invalid pairings of segments. To do so, we employ a graph-based method similar to the one presented in [9]. A skeleton graph is constructed in which each body part represents a node, and joints form the edges between them. Each edge is assigned a weight that corresponds to the value of (2) that we obtained for the pair of nodes (segments) that it connects. The topology of the skeleton can be determined by constructing the minimal spanning tree [15] of the skeleton graph.

Our method efficiently and robustly computes joint positions even for very long sequences with complex motion, as seen in Figs. 2 and 3.

5 Enforcing Constant Bone Lengths

Up to now, the lengths of the bones in the skeleton can vary from time step to time step. However, eventually one wants to express the motion parameters based on a single skeleton with constant dimensions. To serve this purpose, we have developed a simple and efficient way to enforce fixed bone length constraints in a separate processing step. We employ a least-squares fitting technique to appropriately adjust the joint positions that we have found by means of the approach described in Sect. 4.

Our algorithm follows the hierarchy of the estimated skeleton from the root down to the leaves (i.e hand/feet) and solves a least-squares problem for each pair of subsequent joints in the kinematic chain. By this means it is more efficient than related methods [10] that enforce fixed bone length constraints by solving a least-squares problem for the whole model at once.

Let us assume that $\boldsymbol{c_f}^i$ is the position of a joint i at frame f, and $\boldsymbol{c_f}^{i-1}$ is the 3D location of its parent joint at frame f. The optimal fixed length of the bone connecting joints $i-1$ and i, $l_{i-1,i}$, as well as the new joint positions of i, $\boldsymbol{oc_f}^i$, for all f can be found by minimizing the following cost function:

$$V(\boldsymbol{oc_1}^i, \ldots, \boldsymbol{oc_N}^i, l_{i-1,i}) = \sum_{f=0}^{N} \|\boldsymbol{c_f}^i - \boldsymbol{oc_f}^i\|^2 + (\|\boldsymbol{oc_f}^i - \boldsymbol{c_f}^{i-1}\| - l_{i-1,i})^2 \quad (5)$$

In (5) the first term is used to keep the new optimal joint positions as close as possible to the old positions, while the second term constrains the bone length to be the same in all frames. The dimension of the parameter space in (5) can be further reduced by expressing the new position of joint i in terms of the normalized direction vector $\boldsymbol{e}_{i-1,i}$ between $i-1$ and i. Replacing $\boldsymbol{oc_f}^i$ by $\boldsymbol{c_f}^{i-1} + \boldsymbol{e}_{i-1,i} * l_{i-1,i}$ in (5):

$$V(l_{i-1,i}) = \sum_{f=0}^{N} \|\boldsymbol{c_f}^i - (\boldsymbol{c_f}^{i-1} + \boldsymbol{e}_{i-1,i} * l_{i-1,i})\| \quad (6)$$

Eq. (6) is independently solved for each pair of subsequent joints in the hierarchy. The final result of our processing pipeline is a skeleton model of correct topology that, at each time step of motion, stands in a correct pose.

6 Results

We have tested our algorithm on a large number of optical motion capture sequences from the CMU motion capture database [16]. They were recorded with Vicon MX40 cameras. The motion sequences we used for testing comprise of 180-4000 frames and show, for example, simple gymnastic exercises, athletic performances and dancing sequences. After pre-processing of the raw data, on average around 110 non-interrupted marker trajectories were available for body model estimation. Fig. 2(left) shows the automatic segmentation result for a

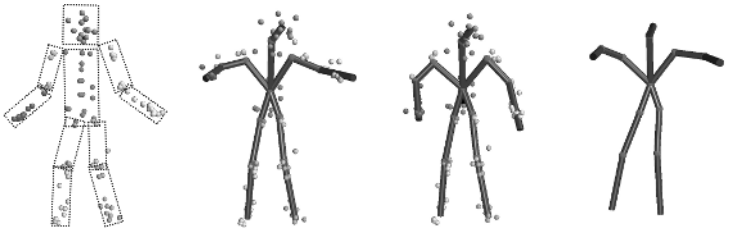

Fig. 2. Gymnastics sequence: Segmentation into body parts - boxes drawn for illustration purpose (left), two poses of the optimal skeleton shown together with the 3D marker positions (middle), and the fixed bone length skeleton at a different time step (right)

gymnastics sequence that was generated by our spectral clustering method described in Sect. 3. Individual segments of the human body have been correctly identified. Unfortunately, in the sequences that were at our disposition no significant foot or hand motion relative to the legs and arms respectively can be observed. In consequence, our algorithm cannot identify hand and feet as separate body segments. However, this is by no means a limitation of our method, but a general problem that is hard to solve for any learning-based approach.

Spectral clustering leads to much better segmentation results than, e.g., simple k-means clustering, since the clustering is far less deteriorated by noise in the data. While a purely distance based segmentation produces many kinematically meaningless rigid bodies, our variance-based scheme in conjunction with spectral clustering produces plausible body segmentations.

Fig. 2 shows different poses of the optimal kinematic skeleton reconstructed for some frames of a gymnastics sequence. One can see that both the topology of the bone skeleton and the positions of the joints have been faithfully estimated. The body models exhibit a high level of detail that is comparable to the complexity of skeletons usually used in animation and biomedical analysis. Fig. 3 shows further reconstruction results that we obtained by applying our algorithm to a dancing sequence.

Our method for estimating the joint locations and skeleton topology performs better than the method proposed by Kirk et al. [11] which is the most closely

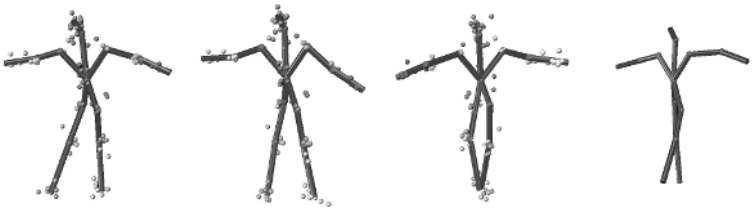

Fig. 3. Dancing sequence: The first three images show the markers and the estimated skeleton in three different poses. The image on the right shows the skeleton with constant bone length at another time step. Joint positions and skeleton topology have been faithfully reconstructed.

related approach from the literature (see also Sect. 4). As shown in Tab. 1, the runtimes of our method on complete motion capture sequences are orders of magnitude faster. We measured them on a Pentium IV with 3.0 GHz using the L-BFGS-B method to solve the minimization problems [17]. The comparison suggests that our approach is well-suited for processing long motion capture sequences as they are commonly recorded for most computer game and movie productions.

We have evaluated the accuracy of our system both visually and qualitatively. Unfortunately, we do not have ground truth measurements of the skeleton structures at our disposition. We thus compared our estimation results against the best possible reference data, which are the body models estimated with commercial software and that are provided by CMU together with their data. Fig. 4 shows visual comparisons for two different time steps. Our algorithm has reliably captured pose and dimension of the body. Please note that although our method reconstructed the topology of the root/spine area in a different way, the overall mobility is the same as in the reference model.

In order to get a qualitative error estimate, we have tested our method on synthetic data. Both test sequences (a walking robot and a jumping snowman) have been generated in 3D Studio Max by animating triangle meshes with handcrafted kinematic skeletons. In both test cases we use randomly selected vertices of the triangle meshes as markers for the reconstruction. Fig. 5(a) shows the robot in one pose and the respective skeleton reconstructed by our method. Kinematic structure and pose have been correctly identified. Fig. 5(b) shows that our approach correctly reconstructs model and pose in the case of the jumping snowman, too. In either test cases, the joint positions estimated by our method deviate on average by around 3% (relative to the model's height) from the true joint positions. This error is much lower than the position inaccuracy that we obtain with the method by Kirk et al. [11] which is in the range of 7%.

Our approach is subject to a few limitations. If the accuracy of marker trajectories is strongly deteriorated by noise or significant non-rigid body deformations (e.g. of the skin) are observed, joint positions may be improperly estimated. However, this is a general problem that commercial systems often fail to handle as well as the reference data provided by CMU suggest. Furthermore, it is impossible to distinguish two rigid body segments if at no time during a motion sequence a relative motion between them is observed. This is not a limitation specific to our approach but a general conceptual limitation of learning-based methods.

Table 1. Comparison of the runtime of our method to the runtime of the method proposed by Kirk et al. [11] on 4 different MOCAP sequences

Sequence	Number of Frames	Kirk et al. [11]	Our method (Sect. 4)
1	189	1649s	103s
2	307	2320s	175s
3	591	4515s	307s
4	1134	11247s	590s

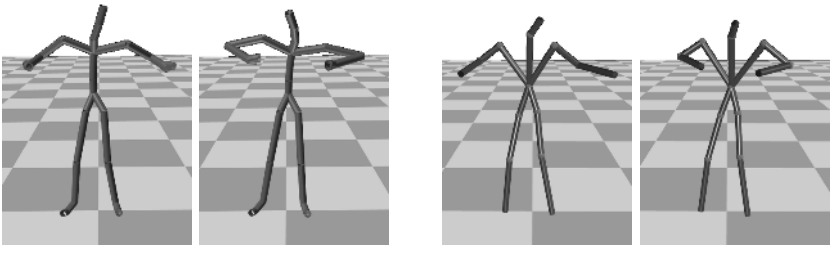

Fig. 4. Visual comparison of the skeletons that are provided with the MOCAP data (two images on left) and our learned skeleton in the same poses (two images on right)

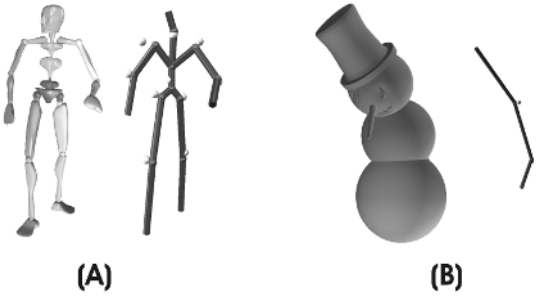

Fig. 5. Evaluation using synthetic data: (a) animated robot mesh (left) and reconstructed kinematic skeleton (right). Ground truth joints are shown as white spheres. (b) Animated snowman (left) and reconstructed skeleton with estimated joints (gray spheres) and ground truth joints (white spheres). The method is able to estimate the kinematic skeleton of general subjects accurately.

Despite these restrictions, our algorithm is an efficient and robust tool that can greatly simplify the motion capture pipeline. As shown, our method can be applied in the same way to motion data of arbitrary subjects including animals, generating accurate skeleton reconstructions.

7 Conclusion

We have presented a fully-automatic system for learning an articulated skeleton model with constant bone lengths and its poses from 3D marker trajectories. Our approach does with no a priori information about the kinematics of the captured individual and can be applied to arbitrary subjects including humans and animals. Through experimental evaluation we have shown that it performs better in terms of speed and accuracy than the most closely related methods from the literature. The learned models are comparable to the ones obtained with commercial software in terms of accuracy and detail. As future work, we plan to integrate our method with an automatic non-intrusive surface reconstruction approach in order to automatically learn complete virtual characters.

Acknowledgments. This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

References

1. Builder, V.B.: (http://www.vicon.com/products/bodybuilder.html)
2. Builder, M.A.S.: (http://www.motionanalysis.com/)
3. Motion, S.: (http://www.simi.com/)
4. Schwartz, M.H., Rozumalski, A.: A new method for estimating joint parameters from motion data. Journal of Biomechanics **38, Issue 1** (2005) 107–116
5. Gamage, S.S.H.U., Lasenby, J.: New least squares solutions for estimating the average centre of rotation and the axis of rotation. Journal of Biomechanics **35, Issue 1** (2002) 87–93
6. Spiegelman, J.J., Woo, S.L.: A rigid-body method for finding centers of rotation and angular displacements of planar joint motion. Journal of Biomechanics **20, Issue 7** (1987) 715–721
7. Cameron, J., Lasenby, J.: A real-time sequential algorithm for human joint localization. In Proc. SIGGRAPH'05, Posters(111) (2005)
8. Ringer, M., Lasenby, J.: A procedure for automatically estimating model parameters in optical motion capture. In: BMVC. (2002)
9. O'Brien, J.F., Bodenheimer, R., Brostow, G., Hodgins, J.K.: Automatic joint parameter estimation from magnetic motion capture data. In: Graphics Interface. (2000) 53 – 60
10. Silaghi, M.C., Plänkers, R., Boulic, R., Fua, P., Thalmann, D.: Local and global skeleton fitting techniques for optical motion capture. In: CAPTECH '98: Proceedings of the International Workshop on Modelling and Motion Capture Techniques for Virtual Environments, London, UK, Springer-Verlag (1998) 26–40
11. Kirk, A.G., O'Brien, J.F., Forsyth, D.A.: Skeletal parameter estimation from optical motion capture data. In: CVPR 2005. (2005) 782–788
12. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm (2001)
13. Duda, R.O., Hart, P.E.: Pattern Classification (2nd Edition). Wiley, New York - London - Sydney (2001)
14. Horn, B.: Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America **4(4)** (1987) 629–642
15. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society **7** (1956) 48–50
16. Database, C.G.L.M.C.: (http://mocap.cs.cmu.edu/)
17. Byrd, R., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comp. **16** (1995) 1190–1208