Tracking human motion with multiple cameras using an articulated model

Davide Moschini and Andrea Fusiello

Dipartimento di Informatica, Università di Verona,
Strada Le Grazie 15, 37134 Verona, Italy
davide.moschini@gmail.com
andrea.fusiello@univr.it

Abstract. This paper presents a markerless motion capture pipeline based on volumetric reconstruction, skeletonization and articulated ICP with hard constraints. The skeletonization produces a set of 3D points roughly distributed around the limbs' medial axes. Then, the ICP-based algorithm fits an articulated skeletal model (stick figure) of the human body. The algorithm fits each stick to a limb in a hierarchical fashion, traversing the body's kinematic chain, while preserving the connection of the sticks at the joints. Experimental results with real data demonstrate the performances of the algorithm.

1 Introduction

Tracking or capturing the motion of a human subject is a problem that has a long history in Computer Vision (see [1] for a survey) and several real-world applications, such as human-computer interfaces, motion transfer, animation of virtual characters, activity/gesture/gait recognition, biomechanical studies. Marker-based commercial systems are available that work at very high frame rates and very high precision. While it is out of doubt that such speed/accuracy combination is necessary in biomechanics, it is questionable whether it is needed when animating a virtual character in a videogame or building a user-interface. Therefore, there is a niche for less expensive markerless systems that work at a reduced speed. In this paper we present some preliminary results of an ongoing project aimed at building a system with those characteristics.

The literature on markerless body tracking in three dimensions can be broadly split into two groups: those using a stick model for the human body [2, 3], roughly corresponding to its skeleton, and those using a full 3D model of the body's shape, in the form of a polygonal mesh or a volumetric model [4–6]. Since we aim at a real-time system, we are forced to work with a stick model. Indeed, a stick (or skeletal) model has fewer dependencies on anthropometric parameters than a shape model and can be tracked much faster because of its simplicity.

Our system bases on volumetric reconstruction from multiple cameras (*shape from silhouette* [7]) followed by skeletonization and model fitting. Proper skeletonization algorithms, like [8, 9], are too computationally demanding to process

more than a few images per second, hence we are proposing here a novel strategy that produces a very coarse – but fast – approximation of the centerline of the human body.

The model fitting is based on the well-known Iterative Closest Point (ICP) algorithm [10]: the model is an articulated stick figure representing the body and it's kinematics, the data are 3D points roughly distributed around the centerline of the limbs. The data are registered to the model using a hierarchical approach that proceeds by traversing the kinematic chain.

Previous work on using ICP on articulated bodies include [11, 6, 12]. In [11] each segment is aligned independently to the data and articulated constraints are enforced *a-posteriori* by projection on the constraints surface. Likewise, [6] uses ICP to find a solution to a problem with relaxed joint constraints, and then forces hard constraints on that solution, thereby interfering with the result of ICP, which is optimal in the least-squares sense. Differently from these works we enforce joints constraint *during* the registration process. The only work with this feature is [12], that have been independently proposed. For the sake of clarity, the discussion on the differences is postponed to Section 5.

Other related approaches include those that optimize the same objective function as the articulated ICP (namely: the sum of squared distance between data and model with respect to the pose parameters of all the segments of the structure) with a different strategy, e.g. Expectation-Maximization [3] or Levenberg-Marquardt. The first is too computationally demanding for a tracking application, whereas the latter have been reported [12] to suffer from convergence to local minima more than articulated ICP.

2 Human Body Model

In this section we describe the articulated model representing the human body pose we used in the paper. It consists of a kinematic chain of ten sticks and nine joints, as depicted in Figure 1. The torso is at the root of tree, children represents limbs, each limb being described by a fixed-length stick and the corresponding rotation from its parent. Hence, the motion of one body segment can be described as the motion of the previous segment in the kinematic chain and an angular motion around a body joint. Only the torso contains a translation that accounts for the translation of the whole body. Rotations are represented with 3×3 matrices. For the sake of simplicity, all the joints are spherical (three d.o.f.) with no angle limits.

3 Shape from Silhouette

Shape from silhouette consist in recovering a volumetric approximate description of the human body (the *visual hull* [13]) from its silhouettes projected onto a number of cameras (three, in our case). Its main advantage over other reconstruction techniques is that it seamlessly integrates the information from multi-

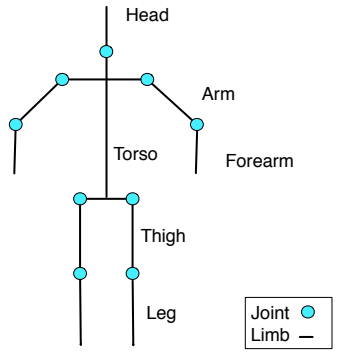

Fig. 1. The stick figure body model.

ple cameras. Moreover, implementations have been demonstrated that achieves real-time performances [14] by exploiting the graphical hardware.

Silhouettes are obtained by background subtraction with the software distributed with the HumanEva dataset [15]. The reconstruction is accomplished using the technique described in [16, 14], with a plane parallel to the floor sweeping the working volume (see Fig. 2). At each step the silhouettes are projected onto the current plane, using the projective texture mapping feature of OpenGL and the GPU acceleration, as described in [17]. The slice of the volume corresponding to the plane is reconstructed by doing the intersection of the projected silhouettes.

4 Skeletonization

The medial axis (or skeleton) of a 3D object is the locus of the centers of maximal spheres contained in the object. In principle it is a surface, even if it can degenerate to a curve or a point. A close relative is the *centerline* (or *curve-skeleton*) that is a curve in 3D space that captures the main object's symmetry axes and roughly runs along the middle of an object. This definition matches with the stick-figures model, hence the data onto which the model is to be registered will be points on the body's centerline.

There are many techniques in literature to find skeletons or centerlines of a 3D object (see [18] for a survey). However, they are too computationally demanding to fit our design, hence we introduce a new method based on slicing the volume along three axis-parallel directions (see Fig. 3). In each slice – which is a binary image – we compute the centroid of every connected component and add it to the set of centerline points. The slicing along the Z-axis comes for free from the previous volumetric reconstruction stage, whereas slicing along X and Y must be done expressly, but uses the same procedure with GPU acceleration.

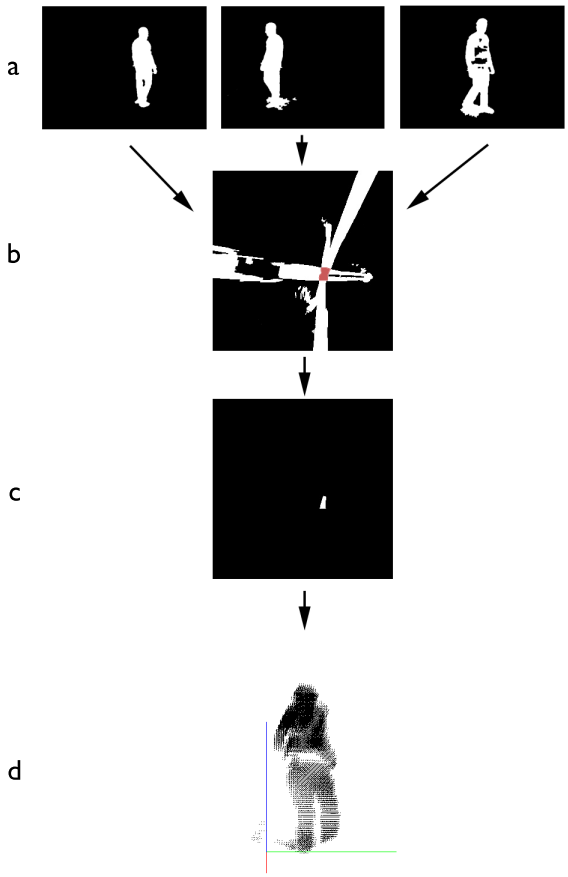

Fig. 2. a) Silhouettes; b) projection onto the sweeping plane; c) intersection (slice) d) final volumetric reconstruction

Our method is similar to [16] which computes the centerline of a body by finding the centroids of the blobs produced by intersecting the body with planes orthogonal to Z-axes. Using a single sweep direction has some problems with some configurations of the body. Consider for example the "T" pose: using only the scan along the Z-axes we completely loose the arms because by cutting the body at the arms height produces one single elongated blob containing a slice of the torso and the two arms, whose centroid is located on the vertebral column.

Our method solves this problem using three sweeps, thus it can be considered as a refinement of [16]. On the other hand, it can also be regarded as a coarse approximation of [19], where first 2D skeletons are extracted for each axis-parallel 2D slice of the 3D volume and then they are intersected to obtain the 3D centerline of the object. When the centroid belongs to the centerline our method

returns a subsampling of the centerline, and this is approximately the case for most configurations of the human body. Yet, when the 2D shape is strongly non convex and the centroid falls outside the shape itself the method yields spurious points. However, the subsequent fitting procedure, that will be described in the following section, has been designed to be robust with respect to outliers.

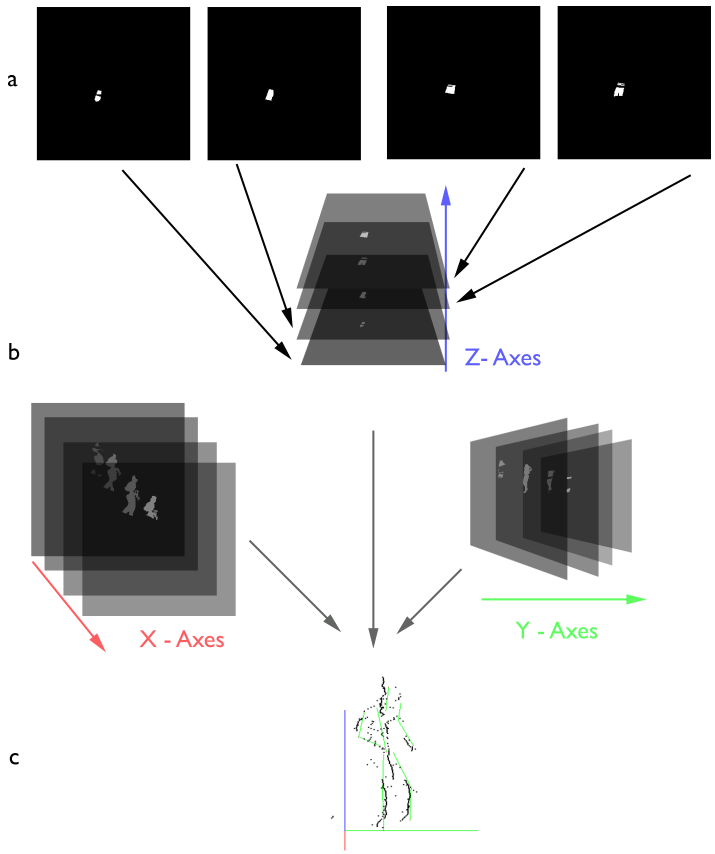

Fig. 3. a) slices along Z; b) slices along X and Y; c) centerline points with the stick figure overlaid.

5 Hierarchical Articulated ICP

This section describes the Hierarchical Articulated ICP algorithm for registering an articulate stick model to a cloud of points. It is based on the well-known Iterative Closest Point (ICP) [20, 10] that estimates the rigid motion between a given set of $3D$ data points and a set of $3D$ model points.

We assume that the data are 3D points distributed roughly around the centerline of the body's segments. The data are registered to the model using a hierarchical approach that starts from the torso and traverse the kinematic chain down to the extremities. At each step ICP computes the best rigid transformation of the current limb that fits the data while preserving the articulated structure.

The closest point search works from the data to the model, by computing for each data point its closest point on the body segments. Only the matches with the current segment are considered, all the other should be – in principle – discarded.

However, the rotation in 3D space of a line segment cannot be computed unambiguously, for the rotation around the axis is undetermined. In order to cope with this problem we formulate a *Weighted* Extended Orthogonal Procrustes Problem and give a small but non-zero weight also to points that match the descendants of the current segment in the kinematic chain. In this way they contribute to constrain the rotation around the segment axis. Think, for example, of the torso: by weighting the points that match the limbs as well, even if they cannot be aligned with single rigid transformation, the coronal (aka frontal) plane can be correctly recovered.

In order to improve ICP robustness against false matches and spurious points, following [21], we discard closest pairs whose distance is larger than a threshold computed using the X84 rejection rule. Let e_i be the closest-point distances, a robust scale estimator is the Median Absolute Deviation (MAD):

$$\sigma^* = 1.4826 \operatorname{med}_i |e_i - \operatorname{med}_j e_j|. \tag{1}$$

The X84 rejection rule prescribes to discard those pairs such that $|e_i - \operatorname{med}_j e_j| > 3.5\sigma^*$.

The Hierarchical Articulate ICP is is described step by step in Algorithm 1.

Point 6, where a transformation is computed given some putative correspondences, deserves to be expanded, in order to make the paper self-contained. The problem to be solved is an instance of the *Extended Orthogonal Procrustes Problem* (EOPP) [22], which can be stated as follows: transform a given matrix A into a given matrix B by a similarity transformation (rotation, translation and scale) in such a way to minimize the sum of squares of the residual matrix. More precisely, since we introduced weights on the points, we shall consider instead the *Weighted Extended Orthogonal Procrustes Problem* (WEOPP) problem. In formulae:

$$\arg\min_R \left\| (cRA + \mathbf{t}\mathbf{u}^\top - B)W \right\|_F^2 \quad \text{subject to } R^T R = I \tag{2}$$

where matrices A and B are $(3 \times p)$ matrices containing p corresponding point in 3-D space, R is (3×3) orthogonal rotation matrix, $\mathbf{t}$ is a (3×1) translation vector, c is scale factor, $\mathbf{u}$ is a $p \times 1$ vector of ones, W is a $(p \times p)$ diagonal matrix weighting the p points, and $\|\cdot\|_F$ denotes the Frobenius norm.

Algorithm 1 Hierarchical Articulate ICP

Input: The model $\mathcal{S}$ composed by segments and the data set $\mathcal{A}$ of 3D points
Output: a set of rigid motions (referred to the kinematic chain) that brings the model onto the data

1. Traverse the body model tree structure using a level-order or a preorder traversal method.
2. Let $s_j \in \mathcal{S}$ be the current body segment.
3. Compute the closest points:
   (a) For each data point $\mathbf{a}_i \in \mathcal{A}$ and for each segment $s_\ell \in \mathcal{S}$ compute its projection $\mathbf{p}_{i\ell}$ onto the line containing s_ℓ ;
   (b) if $\mathbf{p}_{i\ell} \in s_\ell$ then add $\mathbf{p}_{i\ell}$ to $\mathcal{M}$ (the set of the closest-point candidates), otherwise add the endpoint of s_ℓ to $\mathcal{M}$.
   (c) Find $\mathbf{b}_i$, the closet point to $\mathbf{a}_i$ in $\mathcal{M}$.
4. Weight the points: If $\mathbf{b}_i$ belongs to s_j than its weight is 1, otherwise it is ε (chosen heuristically) for all the descendant and 0 for all the others.
5. If the distance of $\mathbf{b}_i$ to $\mathbf{a}_i$ is above the X84 threshold then the weight is set to 0.
6. Solve for the transformation of s_j.
7. Apply the transformation to s_j and its descendants.
8. Repeat from step 3 until the weighted average distance between closest points points is less than a given threshold.

The solution to the the problem (derived in [23]) is based on the Singular Value Decomposition (SVD). Let

$$UDV^\top = A_w \left(I_p - \frac{\mathbf{u}_w \mathbf{u}_w^\top}{\mathbf{u}_w^\top \mathbf{u}_w} \right) B_w^\top \tag{3}$$

be the SVD decomposition of the matrix on the right-hand side[1], where $A_w = AW$, $B_w = BW$, and $\mathbf{u}_w = W\mathbf{u}$. The sought transformation is given by (we omit the scale c that is not needed in our case):

$$R = V \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(VU^T) \end{bmatrix} U^\top \tag{4}$$

$$\mathbf{t} = (B_w - RA_w) \frac{\mathbf{u}_w}{\mathbf{u}_w^\top \mathbf{u}_w} \tag{5}$$

The diagonal matrix in (4) is needed to ensure that the resulting matrix is a rotation [24]

The *Weighted Orthogonal Procrustes Problem* (WOPP) problem is a special case of WEOPP and the solution can be derived straightforwardly by setting $\mathbf{u} = \mathbf{0}$. In our case we use WEOPP for the torso and WOPP (only rotation) for the limbs.

[1] Please note that $A\frac{\mathbf{u}_w \mathbf{u}_w^\top}{\mathbf{u}_w^\top \mathbf{u}_w}$ is a matrix of the same size as A with identical columns, each of them equal to the centroid of the points contained in A.

The hierarchical articulate ICP is deterministic, every limb is considered only once and brought into alignment with ICP. The transformation that aligns a limb s_j is determined mostly by the points the matches s_j and secondarily by the points that matches its descendants. The transformation is applied to s_j and its descendants, considered as a rigid structure. The output of the algorithm represents the pose of the body. In a tracking framework, the pose obtained at the previous time-step is used as the initial pose for the current frame.

A similar algorithm has been independently proposed in [12]. The main difference is in the way the basic ICP is applied to the articulated structure, which leads to different schema. In [12] at each step of the algorithm the subtree of the selected joint is rigidly aligned using ICP with no weights, i.e., all the descendants of the joint plays the same role in the minimization. As a result, the same joint needs to be considered more than once to converge to the final solution. In this regard our approach is less computationally demanding. On the other hand one error in the alignment of a limb propagates downward without recovery, whereas in [12] a subsequent sweep may be able to correct the error, hence [12] seems to be more tolerant to a looser initialization.

6 Experimental Results

The body tracker has been tested on sequences taken from the HumanEva-I dataset [15]. All the sequences in HumanEva-I have been calibrated using the Vicon's proprietary software and the motion data saved in the common `c3d` file format. The dataset contains multiple subjects performing a variety of actions like walking, running, boxing, etc. In particular we used the sequences called *"S2 Jog"*, *"S2 Throwcatch"*. Figures 4 and 5 show some sample frames from these sequences together with the output of the silhouette extraction.

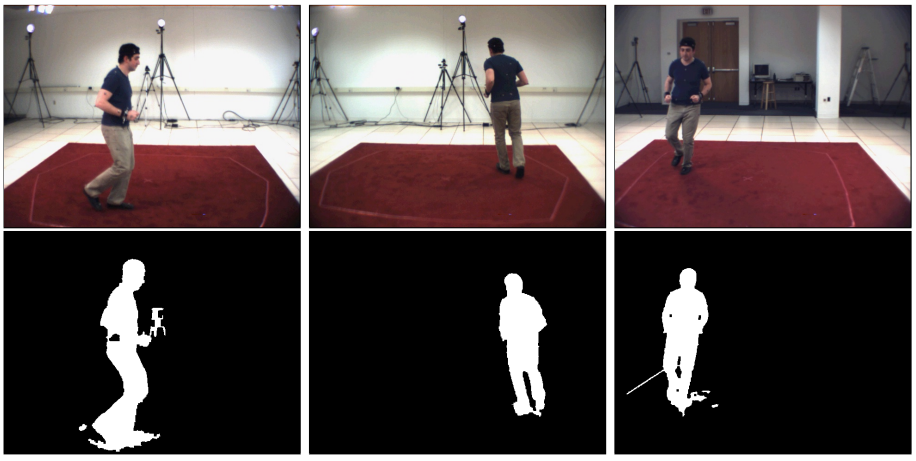

Fig. 4. Sample frames of *"S2 Jog"* and silhouettes.

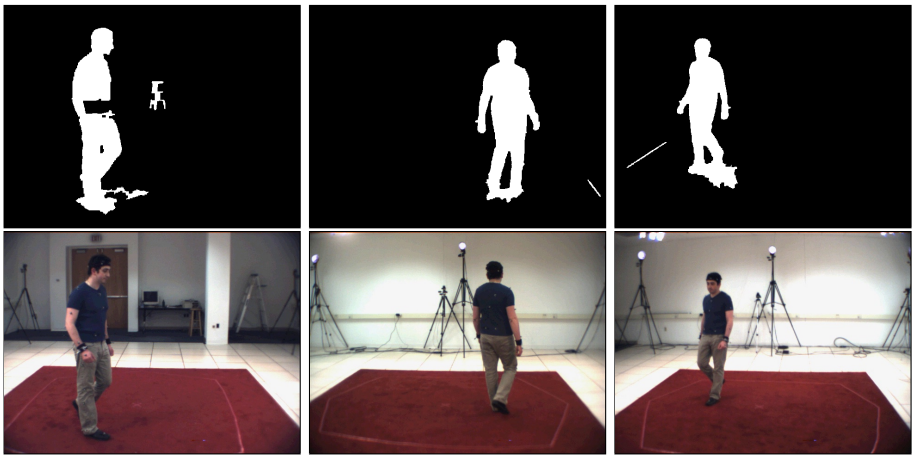

Fig. 5. Sample frames of *"S2 Throwcatch"* and silhouettes.

Validation of the algorithm is done by comparing the angles of the ground-truth with the angles of the computed model. Figure 6 reports the ground truth and estimated joint angles of the torso, right shoulder and right elbow in the two sequences. It can be seen that the estimated angles follows fairly closely the ground truth. There some spikes where the error grows but the tracker is able to recover in the subsequent frames. We expect that a Kalman filter will be able to smooth out significantly those spikes.

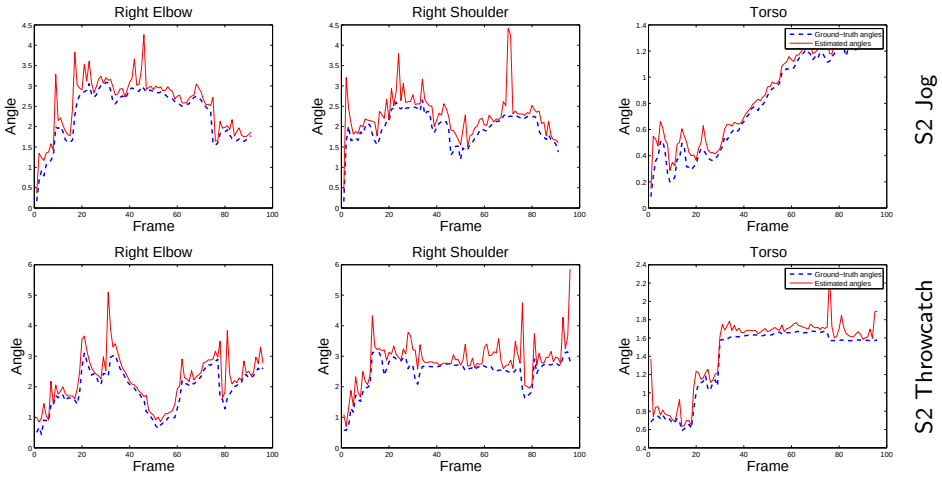

Fig. 6. Plots comparing ground truth and estimated joint angles of the torso, right shoulder and right elbow in the two sequences used for the experiments (the sequence name is on the right).

This results are remarkable if one considers the coarseness of the volumetric reconstruction, due to the small number of cameras (three) and the poor quality of image silhouettes.

For a quantitative comparison we computed the following angular error for each joint, in each frame of the sequence:

$$e(R_1, R_2) = \angle(R_1 R_2^\top) \tag{6}$$

where $\angle(\cdot)$ denotes the angle of the axis-angle representation of the rotation, and can be computed with $\angle(R) = \arccos((\mathrm{tr}(R) - 1)/2)$.

Mean and standard deviation of the error are shown in Table 1. The magnitude of the error is still higher than the target standard, which is about three degrees, as reported in [25]. We expect, however, that a Kalman filter will be able to smooth out significantly the aforementioned spikes and thus reduce significantly the error.

		S2 Jog	*S2 Throw-catch*
Torso	*Mean*	0.29	0.21
	Std. dev.	0.40	0.57
Neck	*Mean*	0.59	0.82
	Std. dev.	0.74	1.09
Left shoulder	*Mean*	0.73	0.44
	Std. dev.	0.81	0.61
Right shoulder	*Mean*	0.63	0.53
	Std. dev.	0.79	0.53
Left hip	*Mean*	0.96	0.56
	Std. dev.	0.12	0.87
Right hip	*Mean*	0.76	1.66
	Std. dev.	1.08	1.67
Left elbow	*Mean*	0.68	0.55
	Std. dev.	0.81	0.80
Right elbow	*Mean*	0.48	0.56
	Std. dev.	0.65	0.81
Left knee	*Mean*	0.51	0.25
	Std. dev.	0.70	0.25
Right knee	*Mean*	0.33	0.70
	Std. dev.	0.38	0.77

Table 1. Mean and standard deviation of the errors (in radians) for each joint of the body for the sequences used in the experiments.

7 Conclusions and Future Work

This paper has proposed a new ICP-based algorithm for tracking articulated skeletal model of a human body. The proposed algorithm takes as input multiple calibrated views of the subject, computes a volumetric reconstruction and the centerlines of the body and fits the skeletal body model in each frame using a hierarchic tree traversal version of the ICP algorithm that preserves the connection of the segments at the joints. The proposed approach uses only the kinematic constraints and no other assumptions are made on the position of the body. This implies that we can recognize potentially all the body configuration.

The results presented here demonstrate the feasibility of the approach, which is is intended to be used in complete system for vision-based markerless human body tracking.

The current Matlab implementation takes about 4 seconds to process a frame on a laptop with an Intel Core Duo Processor T2250. However, being the algorithm still in a prototypal stage, we are confident that a careful implementation in C/C++ could achieve nearly real-time performances. Indeed all the design choices focused on computational efficiency: the use of a simple stick model, the volumetric reconstruction on the GPU, the fast approximated skeletonization, the hierarchical ICP.

Future work will be aimed at optimizing the implementation and tackling the issue of pose initialization.

Acknowledgments

This paper was partially supported by PRIN 2006 project 3-SHIRT. Thanks to A. Giachetti for inspiring discussions.

References

1. Moeslund, T., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding **103**(2-3) (November 2006) 90–126
2. Brostow, G.J., Essa, I., Steedly, D., Kwatra, V.: Novel skeletal representation for articulated creatures. In: ECCV04. (2004) Vol III: 66–78
3. Ménier, C., Boyer, E., Raffin, B.: 3d skeleton-based body pose recovery. In: Proceedings of the 3rd International Symposium on 3D Data Processing, Visualization and Transmission, Chapel Hill (USA) (June 2006)
4. Anguelov, D., Koller, D., Pang, H.C., Srinivasan, P., Thrun, S.: Recovering articulated object models from 3D range data. In: Proc. of the 20th conference on Uncertainty in Artificial Intelligence, Arlington, Virginia, United States (2004) 18–26
5. Knoop, S., Vacek, S., Dillmann, R.: Modeling joint constraints for an articulated 3D human body model with artificial correspondences in ICP. In: Proc. 5th IEEE-RAS International Conference on Humanoid Robots. (2005) 74–79

6. Mundermann, L., Corazza, S., Andriacchi, T.: Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. (June 2007) 1–6
7. Martin, W.N., Aggarwal, J.K.: Volumetric descriptions of objects from multiple views. IEEE Transactions on Pattern Analysis and Machine Intelligence **5**(2) (March 1983) 150–158
8. Sharf, A., Lewiner, T., Shamir, A., Kobbelt, L.: On-the-fly curve-skeleton computation for 3d shapes. Comput. Graph. Forum **26**(3) (2007) 323–328
9. Dey, T.K., Sun, J.: Defining and computing curve-skeletons with medial geodesic function. In: SGP '06: Proceedings of the fourth Eurographics symposium on Geometry processing, Aire-la-Ville, Switzerland, Switzerland (2006) 143–152
10. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Transaction on Pattern Analysis and Machine Intelligence **14**(2) (1992) 239–256
11. Demirdjian, D., Ko, T., Darrell, T.: Constraining human body tracking. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA (2003) 1071
12. S. Pellegrini, K.S., Nardi, D.: A generalisation of the icp algorithm for articulated bodies. In: Proceedings of the British Machine Vision Conference. (2008)
13. Laurentini, A.: The visual hull concept for silhouette-based image understanding. IEEE Trans. Pattern Anal. Mach. Intell. **16**(2) (1994) 150–162
14. Li, M., Magnor, M., Seidel, H.P.: Hardware-accelerated visual hull reconstruction and rendering. In: In Graphics Interface 2003. (2003) 65–71
15. Sigal, L., Black, M.: Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, Department of Computer Science (2006)
16. Michoud, B., Guillou, E., Bouakaz, S.: Human model and pose Reconstruction from Multi-views. In: International Conference on Machine Intelligence. (November 2005)
17. Everitt, C., Rege, A., Cebenoyan, C.: Hardware shadow mapping. In: In ACM SIGGRAPH 2002 Tutorial Course no.31: Interactive Geometric Computations. (2002) 38–51
18. Cornea, N.D., Silver, D., Min, P.: Curve-skeleton applications. In: IEEE Visualization Conference. (October 2005) 95–102
19. Telea, A., van Wijk, J.J.: An augmented fast marching method for computing skeletons and centerlines. In: VISSYM '02: Proceedings of the symposium on Data Visualisation 2002, Aire-la-Ville, Switzerland, Switzerland (2002) 251–ff
20. Chen, Y., Medioni, G.: Object modelling by registration of multiple range images. Image and Vision Computing **10**(3) (1992) 145–155
21. Fusiello, A., Castellani, U., Ronchetti, L., Murino, V.: Model acquisition by registration of multiple acoustic range views. In: Proceedings of the European Conference on Computer Vision. (2002) 805–819
22. Schnemann, P., Carroll, R.: Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika **35**(2) (June 1970) 245–255
23. Akca, D.: Generalized procrustes analysis and its applications in photogrammetry. Technical Report, ETH, Swiss Federal Institute of Technology Zurich, Institute of Geodesy and Photogrammetry (2003)
24. Kanatani, K.: Geometric Computation for Machine Vision. Oxford University Press (1993)
25. Rosenhahn, B., Brox, T., Kersting, U., Smith, A., Gurney, J., Klette, R.: A system for marker-less motion capture. Knstliche Intelligenz **20**(1) (January 2006) 45–51