In Proceedings of the 4th International Conference on Intelligent Robotics and Applications (ICIRIA), Aachen, Germany, December 2011

# 3D Body Pose Estimation using an Adaptive Person Model for Articulated ICP

David Droeschel and Sven Behnke

Autonomous Intelligent Systems Group, Computer Science Institute VI
University of Bonn, Bonn, Germany
droeschel@ais.uni-bonn.de, behnke@cs.uni-bonn.de

**Abstract.** The perception of persons is an important capability of today's robots that work closely together with humans. An operator may use, for example, gestures to refer to an object in the environment. In order to perceive such gestures, the robot has to estimate the body pose of the operator.
We focus on the marker-less motion capture of a human body by means of an *Iterative Closest Point* (ICP) algorithm for articulated structures. An articulated upper body model is aligned with the depth measurements of an RGB-D camera. Due to the variability of the human body, we propose an adaptive body model that is aligned within the sensor data and iteratively adjusted to the person's body dimensions. Additionally, we preserve consistency with respect to self-collisions. Besides that, we use an inverse data assignment, that is particularly utile for articulated models.
Experiments with measurements of a Microsoft Kinect camera show the advantage of the approach compared to the standard articulated ICP algorithm in terms of the *root mean squared* (RMS) error and the number of iterations the algorithm needs to converge. In addition, we show that our consistency checks enable to recover from situations where the standard algorithm fails.

**Keywords:** Human-Robot Interaction, Marker-less motion capture, Articulated ICP

## 1 Introduction

Today's robots need to operate in environments closely together with humans. For example, in household environments a *domestic service robot* has to interact with people, navigate around them or deliver objects to a user. Interacting with a user hereby could mean that the user refers to an object in the environment by pointing to it, rather than verbally describing it [8]. Pointing to an object is a way of communication where humans use their whole body as a medium. Therefore, the robot has to perceive the human's body pose , i. e., the individual joint angles and the location and orientation of the body parts, in order to detect a motion as a gesture, to determine the pointing direction, and map it to a target in the environment.

We focus on marker-less estimation of such body poses by means of an Iterative Closest Point (ICP) approach to fit a human body model in 3D point measurements from a depth camera. Hence, a precise and complete model of a human body is necessary.

The human body is an articulated object that is highly variable in its size and shape. It consists of a skeleton with many degrees of freedom covered with tissue and skin. In addition, the human body is usually covered in clothes that obfuscate its shape. Due to the variability in the human body shape, a static model is disadvantageous. Hence, a person-dependent model has to be adapted from a generic model.

In contrast to previous work, our approach leverages the advantage of an adaptive body model that is aligned within the measured 3D points and iteratively adjusted to the person's body dimensions. Besides that, we use an inverse data assignment that is particularly utile for articulated models. Our approach is based on depth images from a RGB-D camera [16] that provides depth and color information at high frame rates.

The remainder of this paper is organized as follows: After a brief review of related work, we describe the structure of the body model (Section 3.1) as well as the basic idea of articulated ICP (Section 3.2). Section 4 describes our extensions to the articulated ICP algorithm that enables an adaptive body model. Finally, we evaluate our extensions and compare the adaptive body model to the static body model.

## 2 Related Work

Perceiving humans has been studied in many research areas since decades. The vast majority of this work employs information from one or more color cameras to estimate the human body pose as surveyed by Moeslund et al. [17]. However, recently affordable depth cameras became available, which fosters research on depth-based approaches to human body pose estimation. Furthermore, these approaches do not suffer from varying lighting conditions.

Ganapathi et al. [10] investigate marker-less human pose tracking from monocular depth images. They combine an accurate generative model with a discriminative model that provides data-driven evidence about body part locations. The generative model applies a local model-based search that exploits the kinematic chain of a body model. The discriminative model utilizes a set of trained patch classifiers to detect body parts and is used for initialization and reinitialization if the local search loses track of the body, e. g., due to fast movements. The detection and localization algorithm has been published by Plagemann et al. [22]. They propose an interest point detector based on identifying geodesic extrema in point clusters that coincide with salient points of the body.

Several recent approaches focus on the extension of a well-established 3D registration method, the *Iterative Closest Point* (ICP) algorithm [4], to articulated models by fitting a static (non-adaptive) cylindrical human body model into 3D measurements. Demirdjian et al. [6], for example, estimate the pose of individual body parts using the ICP algorithm in 3D point clouds generated by dense stereo. Thereby, the poses for individual body parts are estimated independently and kinematic constraints are enforced after registration. These constraints are implemented by a support vector machine (SVM) classifier that is trained on data from a motion capture system.

Ogawara et al. [19] estimate the human body pose from an occupied volume from multiple video streams. They use a deformable skin model with joint structure consisting of Bézier surfaces, as proposed in [15]. The idea is inspired by Kehl et al. [11] who

use an extension of the ICP formulation to deformable objects [20] and an M-estimator, a generalized form of the least squares method, to minimize the ICP error function [23]. In contrast, Ziegler et al. [24] formulate the problem of tracking a body pose as state estimation problem, modeling the joint angles as a state vector in an unscented Kalman filter (UKF). Their approach is related to the ICP algorithm since they determine point correspondences by spatial neighborhood and iteratively refine their estimation of the joint angles.

The work by Mündermann et al. [18] and their more recent work [5] generalize the ICP algorithm to articulated models by jointly minimizing the distance from the registered data points to the model surface using a Levenberg-Marquardt minimization scheme. To overcome the variations in the human body, they propose to match a person against a database of articulated models [1]. A specific articulation model is chosen from the database that correlates in height and volume. Such a body model consists of a set of triangle meshes.

Pellegrini et al. [21] propose to divide the articulated body into parts that can be aligned rigidly using a closed-form solution. Therefore, the articulation structure is split into two branches and a single joint angle is adapted. Knoop et al. [14] also apply the ICP to each body part individually.

Azad et al. [3] use a particle filter to estimate the pose of a upper body model from a stereo camera. Their body model consists of fixed-sized cones connected by ball (shoulder) or hinge (elbow) joints. The head is directly connected to the abdomen, omitting a neck. By means of image-based cues, like edges or color they update the particle set. In [2], the authors extend their work by a 3D hand/head tracking as a separate cue for the particle filter. Kim et al. [12] combine the ICP algorithm with a particle filter. The human body is modeled by a set of cylinders and a sphere for the head. Each body part consists of a set of 3D points that model the surface of it. In [13], the same authors propose heuristics to speed up the assignment of correspondences.

All of the mentioned approaches employ a person model that is static during alignment. In contrast, we adapt a generic person model to a person-specific body model to account for the variability in the person's dimensions.

## 3 Basic Algorithm

### 3.1 Human Body Model

We model the human body as an directed acyclic graph. The rigid body parts $b_1, b_2, ...b_B$ are the vertices of the graph. Starting from the pelvis, which is the root node of the graph, body parts are connected to other body parts by edges $j_i$ representing the joints. Each body part $b_i$ is modeled by a cylinder with parameters $l_i$ (length) and $r_i$ (radius) and a transformation $\mathbf{T}_i$ that defines the orientation and translation to its parent. In case of the root node $b_r$, $\mathbf{T}_r$ describes the transformation of the complete body model to the coordinate system origin. In addition, every part has a set of points $\mathbf{m}_1^{b_i}, .., \mathbf{m}_L^{b_i}$ assigned to it that model the cylindrical surface, based on $l_i$ and $r_i$.

### 3.2 Articulated Iterative Closest Point Algorithm

The general formulation of the ICP algorithm [4] aims at finding a rigid transformation between a model point set $M$ and a scene point set $D$. For a set of $N$ corresponding point pairs a transformation $T$ that minimizes

$$E(\mathbf{T}) = \sum_{i=1}^{N} ||\mathbf{m}_i - \mathbf{T}d_i||^2 , \tag{1}$$

is determined by performing an iterative least squares minimization scheme. The solution can be determined by several closed-form algorithms. In the articulated case, Equation 1 is extended to

$$E(\mathbf{T}_1 ... \mathbf{T}_B) = \sum_{j=1}^{B} \sum_{i=1}^{L} \left|\left| \mathbf{T}_j m_i^{b_j} - d_i \right|\right|^2 . \tag{2}$$

Similar to [21], we split the complete body chain in two subsets at joint $j_s$ and align them with a rigid transformation with respect to $j_s$ which can be solved in closed-form. The splitting joint $j_s$ is chosen successively and varied in every iteration. In case of $j_s$ being the joint assigned to the root node, the entire model is aligned in the data.

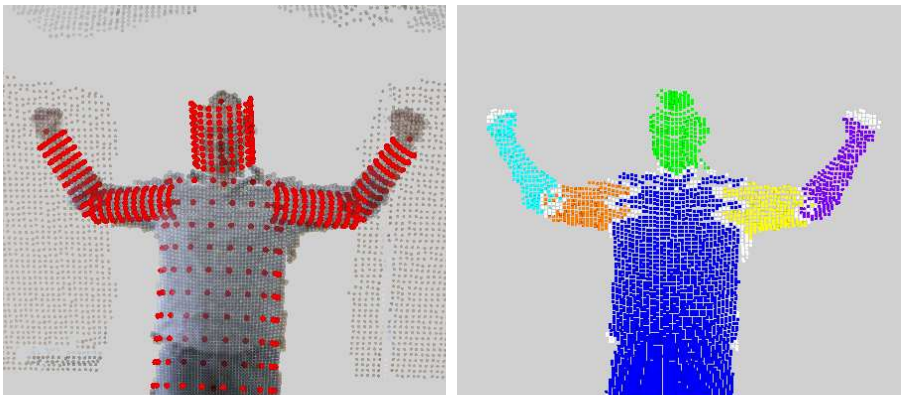
## 4 Proposed Extensions

### 4.1 Data Segmentation

We apply Euclidean clustering in the 3D point cloud to segment the input data and reduce the number of possible correspondences. Neighboring points are assigned to the same point cluster if the Euclidean distance between them does not exceed a threshold $\tau_d$. The distance threshold needs to be chosen appropriately to take the sensor's accuracy in distance measurements into account. For our setup we use a threshold $\tau_d = 5$ cm. We exclude clusters with less than $\tau_n = 500$ points from further processing. The person model is aligned to each remaining cluster with a standard ICP run, i. e., $j_s = j_r$, with $j_r$ being the root joint. After convergence the point cluster that minimizes Eq. 2 is assumed to be the point cluster that corresponds to the person and remaining clusters are removed from the point cloud.

### 4.2 Data Assignment

An important step in the ICP algorithm is the data assignment where point correspondences between the model point set and the scene point set are established. A common way is to determine the nearest neighbor $d_k$ for every model point $m_i$ in the scene point set. This can also be conveyed to articulated structures and correspondences are determined for every body part individually. However, this assignment is disadvantageous for scene points that are close to more than one body part, e. g., points close to a joint.



(a) (b)

**Fig. 1.** (a) Exemplary scene with aligned body model. (b) Resulting scene points after segmentation with color coded assignment to body parts. Points that are rejected due to distance ratio or absolute distance from any body part are colored white.

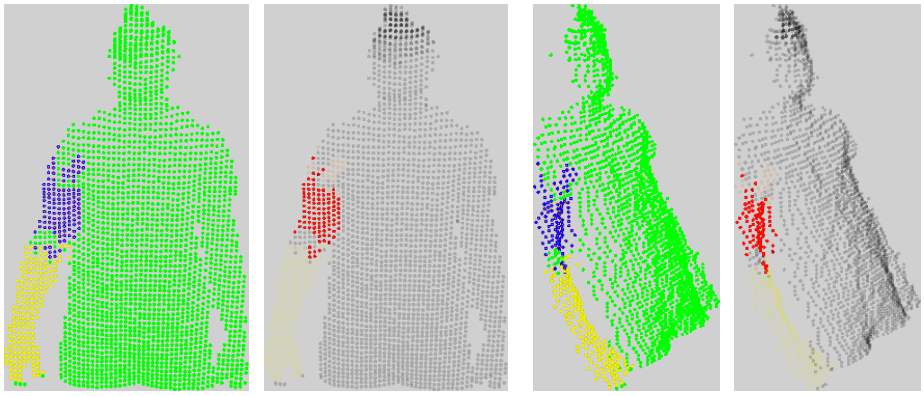
In contrast, we use an inverse data assignment where every scene point is assigned to its closest body part. We reject ambiguous correspondences that cannot be clearly assigned to one body part using the distance ratio between closest and the second-closest body part. For our setup, we reject all assignments in which the distance ratio exceeds $0.8$. Moreover, we reject points that are too far away from any body part. Fig. 1 shows the resulting assignment for an exemplary scene.

### 4.3 Model Adaption

After assigning correspondences and before estimating the joint angle, we adapt the model for each body part based on the surface of the assigned points. Assigned points hereby means the correspondences from the previous step. Similar to the splitting joint $j_s$, we choose the body part that is adapted successively and vary it in every iteration.

By means of a RANSAC [9] estimator, a cylindrical model is fitted into the data points corresponding to a body part. Randomly, data points are selected and the best cylindrical model in terms of the overall number of inliers is calculated, where inliers are points that are closer than $5$ cm to the cylinder model. Since the quality of the fitted cylindrical model increases with the number of inliers, a model estimation with less than 100 inliers is neglected. Fig. 2 shows the calculated inliers for an exemplary scene and body part (upper arm). The resulting cylinder is described by vector $\hat{d}$ and radius $\hat{r}$, where $\hat{d}$ corresponds to the direction of the cylinder in the coordinate origin. In order to get the length $l$ of a cylinder, the inliers are transformed by the inverse of $\hat{d}$ to align with the x-axis.

Since the measurements are subject to noise and the assignment of data points can be inaccurate, especially in the first iterations, the model parameters $p$ (i.e., radius $r$ and length $l$) are filtered over time. Parameter $p_k$ at time step $k$ is calculated by



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

**Fig. 2.** (a+c) Data assignment for the upper arm (violet) and forearm (yellow) from two different perspectives. (b+d) The resulting inliers (red) for a the estimated cylindrical upper arm model.

$$p_k = c\hat{p_k} + (c-1)p_{k-1}. \tag{3}$$

We trust a new estimation of the model parameter $\hat{p_k}$ at time step $k$ with $c = 0.1$. From the resulting model parameters a new set of model points $\mathbf{m}_1^{b_i}, .., \mathbf{m}_L^{b_i}$ for the body part $b_i$ is generated that replaces the previous model points.

In case of a wrong assignment or an insufficient cylindrical fit as well as for initialization, we use a prior for the radius and length of every body part. It prevents from growing or shrinking to abnormal dimensions. For instance, in our system a forearm has a minimum radius of 5 cm and a maximum radius of 15 cm.

### 4.4 Self-Collision Checking

After each ICP iteration, the current state of the model is checked for consistency with respect to self-collisions. A self-collision of a body part is detected by calculating the distance to every body part in the articulation chain except its direct parent and children (i. e. neighboring nodes in the graph). Thus, we allow neighboring body parts to collide with each other, e. g., the forearm can collide with the upper arm but not with the abdomen. In case of a self-collision, the transformation of the current iteration is inverted and applied to the selected joint $j_s$.

### 4.5 Model Initialization

For a good alignment of the model in the first data frames a proper initialization of the articulation structure is necessary. In order to do so, a body segmentation step as proposed in [7] can be used. In this approach, body features such as shoulder, elbow and hand are extracted from a point cluster, based on geodesic distances and geometric priors. The resulting body features can be used to initialize the joint states of the model.

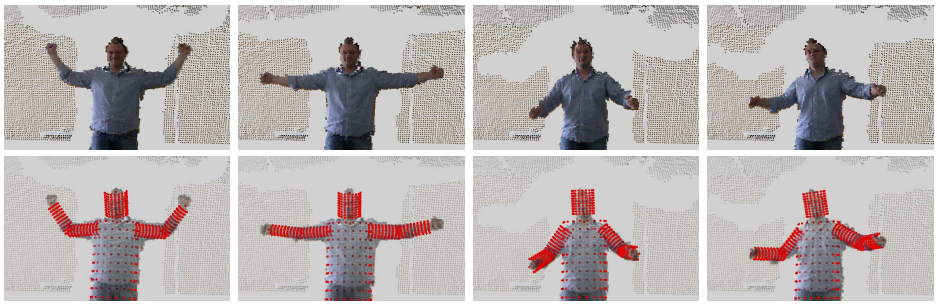


**Fig. 3.** Each column shows one frame from the test data set. The first row shows the raw sensor measurements. The second row shows the adapted body model (red points).

## 5 Experiments

We evaluate our approach with measurements of a Microsoft Kinect camera [16]. For the following experiments, we use a down-sampled depth image with QQVGA resolution (160 $\times$ 120). The average runtime of our current implementation on the down-sampled depth image is 112 ms on a 2.4 GHz single core of a Core2Duo laptop computer. The runtime depends on the number of iterations that are necessary to converge. In general, the number of iterations decreases after an initial model alignment and a frame rate of 8 Hz can be achieved. We focus on an upper body model, since the camera has a field-of-view of $58^\circ \times 45^\circ$. An adult person with typical European body proportions stands in front of the camera in two meter distance.

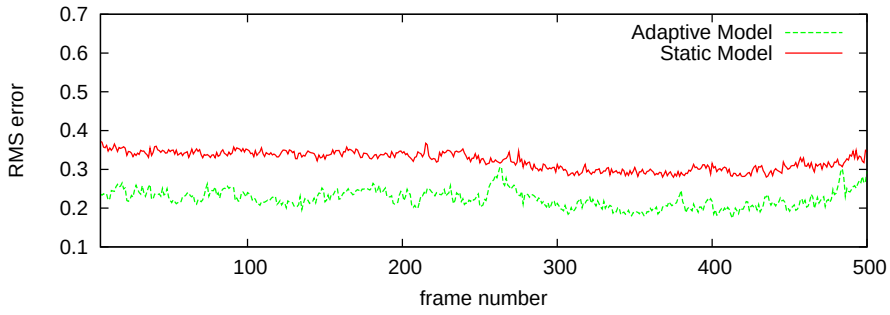



**Fig. 4.** RMS error for each data frame after convergence. The adaptive model (green dashed) is compared to the static model (red).

In a first experiment, we compare the *root mean squared* (RMS) error of the aligned body model with and without model adaption. The data set consists of 500 data frames of a person performing four different body poses. Fig. 3 shows the four body poses with the adapted body model. After convergence of the ICP algorithm, the RMS error

of a model configuration is calculated by Equation 2 for every data frame. Fig. 4 shows the RMS error of the entire data set for the adaptive and the fixed body model. It can be observed that the adaptive model is better aligned with the data. Besides that, using an adaptive model reduces the necessary number of iterations to converge, as shown in Fig. 5.

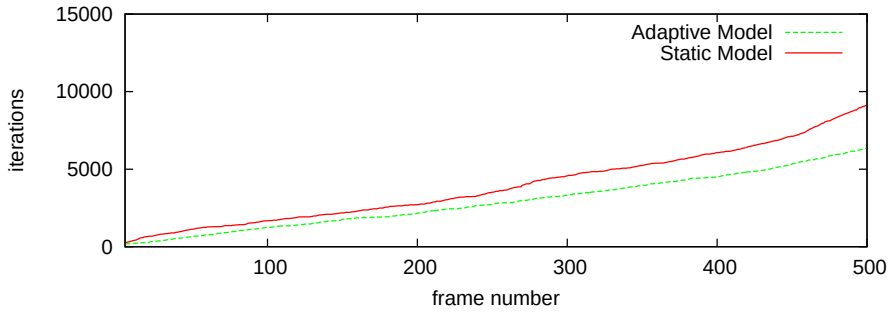



**Fig. 5.** Number of iterations that are necessary to converge, cumulated up to every frame, for the test data set.

In a second experiment, we demonstrate the effect of self-collision prevention. Here, the test person relaxes his arms and keeps them close to his body. Without self-collision checking, body parts of the articulated model may collide with each other, e. g., the forearm with the abdomen. This could result in wrong correspondences, e. g., points from the abdomen could be assigned as correspondences for the forearm. With self-collision checks, a minimal distance between body parts is maintained that prevents from these wrong assignments. Fig. 6 shows frames from the dataset with and without self-collision checks.

The third experiment, shown in Fig. 7, demonstrates how the algorithm aligns the body model when using an incorrect initialization. Here, the angle of the shoulder joint differs between test person and body model and the cylinder dimensions of the body parts are initialized too large (Fig. 7, top left). After 72 iterations (Fig. 7, bottom right), the body model is correctly aligned and the parameters of the model are properly adapted.

## 6 Conclusions

We propose an extension to the ICP algorithm for articulated models. Due to the variability in the human body shape, we use an adaptive body model that is aligned to 3D point measurements and iteratively adjusted to the person's body dimensions, in contrast to previous approaches, that rely on the correctness of a static model. Besides that, we use an inverse data assignment, that is particularly utile for articulated models. Our approach is based on depth measurements of a RGB-D camera. In experiments, we compare our approach to the standard articulated ICP algorithm with a static body

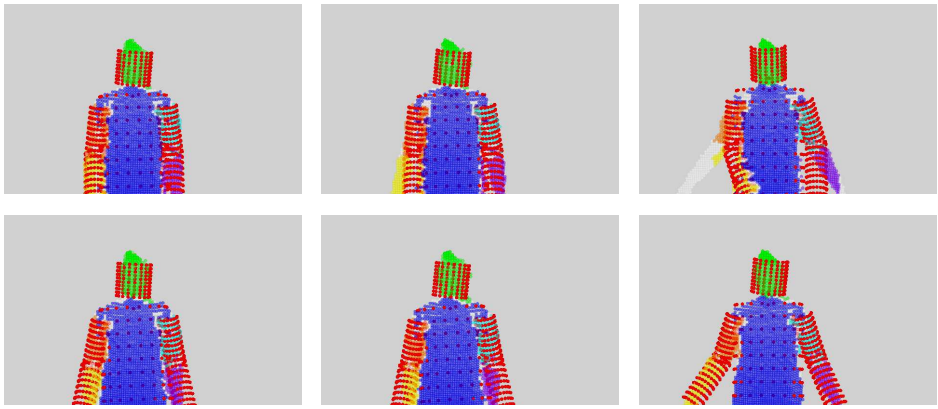


**Fig. 6.** Self-collision prevention. Without self-collision prevention (top row) false correspondences are assigned which results in a wrong alignment of the forearm. With enabled self-collision prevention (bottom row) the model can be aligned correctly.

model. The evaluation shows that an adaptive model aligns better with the data in terms of the RMS error.

We also implement a self-collision check and demonstrate its utility in an experiment. Furthermore, we show how an incorrect model initialization still results in a correct aligned body model.

Up to now, our system only relies on the depth measurements of the camera. However, in some situations, the color images might be beneficial. It is a matter of future work to integrate color information into the algorithm. Besides that, the extracted body pose can be used to interpret, e.g., pointing gestures and an intended pointing target. To do so, the system described in [7] will be adapted to the described body pose estimation. Another possibility for future work is a GPU-based implementation to benefit from the full resolution of the camera and achieve real-time performance.

## Acknowledgment

This work has been supported partially by grant BE 2556/2-3 of German Research Foundation (DFG).

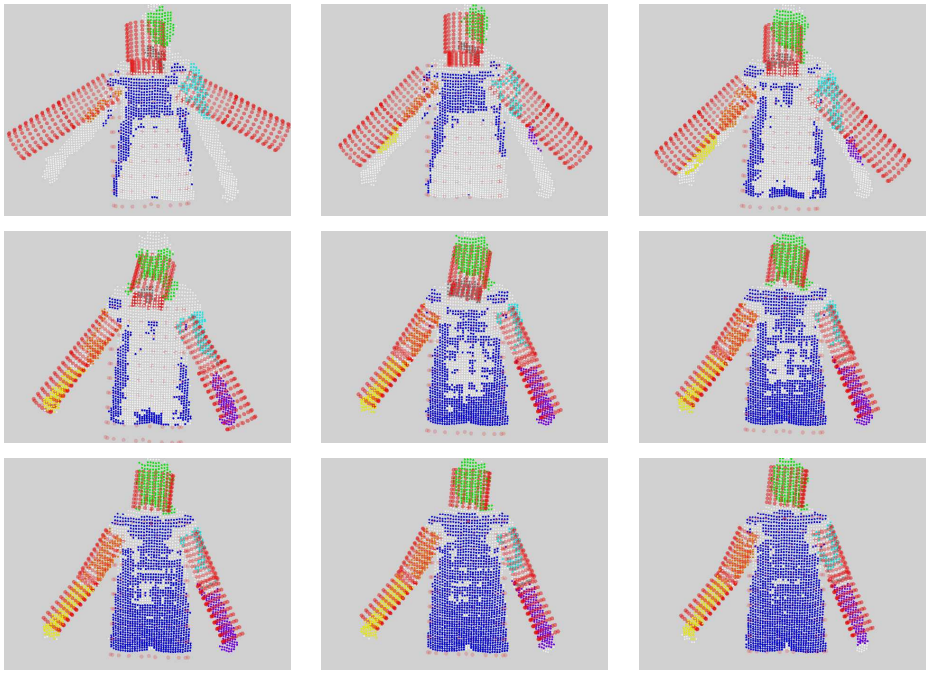


**Fig. 7.** The alignment of the model at different iterations for the same data frame. Even with an incorrect initialized model, the algorithm converges after 72 iterations.

## References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape completion and animation of people. In: Proc. of the 32nd International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). Los Angeles, California (2005)
2. Azad, P., Asfour, T., Dillmann, R.: Robust real-time stereo-based markerless human motion capture. In: IEEE/RAS International Conference on Humanoid Robots (Humanoids). pp. 700–707 (2008)
3. Azad, P., Ude, A., Asfour, T., Dillmann, R.: Stereo-based markerless human motion capture for humanoid robot systems. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 3951–3956 (2007)
4. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 14, 239–256 (1992)
5. Corazza, S., Mündermann, L., Gambaretto, E., Ferrigno, G., Andriacchi, T.P.: Markerless motion capture through visual hull, articulated ICP and subject specific model generation. International Journal of Computer Vision 87, 156–169 (2010)
6. Demirdjian, D., Ko, T., Darrell, T.: Constraining human body tracking. In: Proceedings of the IEEE International Conference on Computer Vision. p. 1071. IEEE Computer Society, Washington, DC, USA (2003)
7. Droeschel, D., Stückler, J., Behnke, S.: Learning to interpret pointing gestures with a time-of-flight camera. In: Proceedings of the 6th International Conference on Human-robot Interaction (HRI). pp. 481–488. ACM, New York, NY, USA (2011)

8. Droeschel, D., Stückler, J., Holz, D., Behnke, S.: Towards Joint Attention for a Domestic Service Robot – Person Awareness and Gesture Recognition using Time-of-Flight Cameras. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 1205–1210. Shanghai, China (2011)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–395 (1981)
10. Ganapathi, V., Plagemann, C., Thrun, S., Koller, D.: Real time motion capture using a single time-of-flight camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA (2010)
11. Kehl, R., Bray, M., Van Gool, L.: Full body tracking from multiple views using stochastic sampling. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pp. 129–136. Washington, DC, USA (2005)
12. Kim, D., Kim, D.: A novel fitting algorithm using the ICP and the particle filters for robust 3d human body motion tracking. In: Proceeding of the 1st ACM workshop on Vision networks for behavior analysis. pp. 69–76. VNBA, ACM, New York, NY, USA (2008)
13. Kim, D., Kim, D.: A fast ICP algorithm for 3-D human body motion tracking. Signal Processing Letters, IEEE 17(4), 402 –405 (2010)
14. Knoop, S., Vacek, S., Dillmann, R.: Modeling joint constraints for an articulated 3D human body model with artificial correspondences in ICP. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids). pp. 74 –79 (2005)
15. Komatsu, K.: Human skin model capable of natural shape variation. The Visual Computer 3, 265–271 (1988), springer
16. Microsoft: http://www.xbox.com/en-US/kinect (2010)
17. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104, 90–126 (2006)
18. Mündermann, L., Corazza, S., Andriacchi, T.: Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
19. Ogawara, K., Li, X., Ikeuchi, K.: Marker-less human motion estimation using articulated deformable model. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 46 –51 (2007)
20. Ogawara, K., Takamatsu, J., Hashimoto, K., Ikeuchi, K.: Grasp recognition using a 3D articulated model and infrared images. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1590 – 1595 (2003)
21. Pellegrini, S., Schindler, K., Nardi, D.: A generalization of the ICP algorithm for articulated bodies. In: British Machine Vision Conference (BMVC) (2008)
22. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Realtime identification and localization of body parts from depth images. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Anchorage, Alaska, USA (2010)
23. Wheeler, M.D., Ikeuchi, K.: Sensor modeling, probabilistic hypothesis generation, and robust localization for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 252–265 (1995)
24. Ziegler, J., Nickel, K., Stiefelhagen, R.: Tracking of the articulated upper body on multi-view stereo image sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 774–781. Washington, DC, USA (2006)