# Robust Model-based 3D Torso Pose Estimation in RGB-D sequences

Markos Sigalas, Maria Pateraki, Iason Oikonomidis, Panos Trahanias
Institute of Computer Science
Foundation for Research and Technology Hellas (FORTH)
{msigalas, pateraki, oikonom, trahania} @ ics.forth.gr

## Abstract

*Free-form Human Robot Interaction (HRI) in naturalistic environments remains a challenging computer vision task. In this context, the extraction of human-body pose information is of utmost importance. Although the emergence of real-time depth cameras greatly facilitated this task, issues which limit the performance of existing methods in relevant HRI applications still exist. Applicability of current state-of-the art approaches is constrained by their inherent requirement of an initialization phase prior to deriving body pose information, which in complex, realistic scenarios, is often hard, if not impossible.*

*In this work we present a data-driven model-based method for 3D torso pose estimation from RGB-D image sequences, eliminating the requirement of an initialization phase. The detected face of the user steers the initiation of shoulder areas hypotheses, based on illumination, scale and pose invariant features on the RGB silhouette. Depth point cloud information is subsequently utilized to approximate the shoulder joints and model the human torso based on a set of 3D geometric primitives and the estimation of the 3D torso pose is derived via a global optimization scheme. Experimental results in various environments, as well as using ground truth data and comparing to OpenNI* User generator *middleware results, validate the effectiveness of the proposed method.*

## 1. Introduction

Free-form partial or full human body pose recovery is a challenging task for a variety of Human Robot Interaction (HRI) applications, as well as for other application domains such as security, telepresence, gaming and surveillance. Recently, the introduction of real-time depth cameras (henceforth referred as RGB-D cameras), and most importantly the emergence of the Microsoft Kinect™ [5], has greatly facilitated the pose recovery task.

Contemporary methods that employ RGB-D sensory input have managed to push the state-of-the-art in the recovery of human-body pose [10, 24]. This is particularly true in the case of controlled or semi-controlled environments. However, when users are allowed to act freely in natural environments, the performance of human pose recovery may be significantly degraded. An example scenario may be that of a service robot (e.g. a robotic salesman or a robotic bar-tender), with multiple users that move, act and interact independently in the scene, seeking attention and possibly service by the robotic agent. In such cases, torso [20] and/or face pose estimation [8, 18, 21] are identified as important attentive cues and are further utilized by fusion modules to initiate interaction with specific user(s). A specific drawback that many pose recovery approaches present [19, 20, 4], is the explicit (e.g. specific body pose) or implicit (e.g. short time of movements) requirement of an initialization phase in order to register the user and assume recovery of body pose parameters. This impedes their applicability in real life scenarios, where initialization is often hard, if not impossible.

The current work focuses on free-form HRI, with multiple humans arbitrarily entering and leaving the scene, and independently (inter)acting. In this context, we are interested in extracting information about the user's body orientation in 3D, as an indicator for attention seeking, under the assumption that an initialization phase isn't possible and shouldn't be required to commence interaction.

To cope with the above, we rely on robust face identification which triggers detection and segmentation of the human body, assuming that the user roughly looks to the camera and, therefore, his/her face is visible. Based on illumination, scale and pose invariant features on the RGB body-silhouette, the shoulder areas are subsequently delineated. Following that, depth point cloud information is further utilized to approximate the shoulder joints which, in turn, are used to model the torso part of the human body. Shoulder joints are modeled as spheres, while the torso is modeled as an ellipsoid. While previous works have also used similar modeling, e.g. Gavrila *et al*. [9], in the current work, however, the modeling is steered by features invariant to scale, pose and body type of users, thus less prone

to produce outliers. A set of anthropometric quantitative measures is utilized to evaluate shoulder joint hypothesis, and initialize and constrain the regression step for the torso fitting. The latter is performed via a non-linear regression method that robustly fits an appropriate body model (ellipsoid) to the corresponding data. The method primarily focuses on overcoming the requirement of large training data and initialization constraints. Moreover, the regression step is performed using a custom iterative algorithm that effectively facilitates real-time operation.

To evaluate our approach we performed extensive experiments using realistic interaction scenarios in different environments, with varying number of active users. We further compared our method with the OpenNI [19] skeleton extraction tool, using ground truth derived from marker-based sequences. Both qualitative and quantitative results are presented that illustrate the method's performance.

The proposed approach is also capable to estimate and track the pose of the arms. Although not thoroughly developed in the current paper, initial promising results attest on its general applicability for upper body pose tracking.

### Related Work

Recent approaches in human pose recovery are reviewed in [6, 17, 23], while the emergence of real-time depth sensors has stimulated new research. Both discriminative and generative approaches have been used in contemporary works, such as Random Forests [10, 12, 24], also combined with Graph Cuts optimization [12] and Probabilistic Graphical Models [4] as well as hybrid approaches, such as Connected Poselets [13].

Grest *et al*. [11] use Iterative Closest Point to extract and track the skeleton while Zhu and Fujimura [28] build heuristic detectors to locate upper body parts (head, torso, arms). Moreover, in [16], vertices are classified and segmented into different body parts, while, Plagemann *et al*. [22] build 3D mesh in order to form geodesic maps for the detection of the head, hand and foot.

Ilic and Fua [14] use *implicit surfaces*, extracted from arbitrary triangulated meshes, to model the upper human body. Stoll *et al*. [26] model both human body and image domain as sums of Gaussians to perform fast articulated motion tracking, while Sigal *et al*. [25] employ a loose-limbed body model to track humans using a non-parametric belief propagation optimization schema. John *et al*. [15] formulate the tracking problem as a multi-dimensional non-linear optimization, solved using particle swarm optimization, while Gall *et al*. [7] propose a multi-layer generative system combining global optimization (simmulated annealing), filtering to smooth out jitter and local optimization, to refine the estimated pose.

As already stated, many of the pose recovery approaches need an initialization step in order to perform effectively. For example, in [11] the model (human) size is known and the starting position predefined, whereas in [28] the system requires a T-pose initialization to size the model. Additionally, the latest version of the OpenNI [19] skeletonization module, needs to register and track the user for a number of frames or seconds as an initialization phase. Nevertheless, and despite the fact that precise details of the end-to-end tracking algorithm are not publicly available, OpenNI is a widely used library in the computer vision and HRI fields.

On the other hand, recent classification approaches ( [1, 24, 27]), although quite robust in (semi)controlled set-ups, exhibit certain weaknesses which limit performance in real life scenarios. Occlusions, loss of sensor data, interactions among multiple users and/or objects may lead to false positive detections, either by misclassifying body parts or by classifying non-human objects as human parts.

In the current work, we aim to overcome such shortcomings and come up with a methodology able to robustly and accurately infer the body pose of multiple users, which act and move freely in naturalistic environments and set-ups.

## 2. Methodology Overview

An overall schematic representation of the steps employed in our method for human-torso 3D pose recovery is given in Fig 1. Initial face identification triggers a segmentation step that delineates the human-body area. Based on that, approximation and 3D modeling of shoulder joints is then performed, which subsequently steers modeling -via a non-linear regression schema- of the torso area as a 3D ellipsoid. The latter is used to extract the torso 3D pose parameters. More specifically, the major steps of the proposed approach are:

- **Agent Segmentation**. Based on face detection and tracking, the human body silhouette is extracted for the detected users in the scene.
- **Shoulder joint approximation**. Given the location of the face, we select sets of points on the RGB silhouette, delineating possible shoulder or armpit areas. Selection is based on pose and scale invariant features satisfying certain geometric constraints. The selected silhouette points are used to define the 2D area of the shoulder joint and thus, using the depth information, the 3D shoulder joint point cloud. Shoulder joints are approximated by least squares fitting of 3D spheres on the selected areas on the point cloud and a set of anthropometric criteria is used to evaluate the estimation and eliminate possible outliers.
- **Body pose estimation**. Driven by the detection of the shoulder joints, a set of 3D points, approximately along the user's upper-body, is selected and used for torso approximation via an ellipsoid. A custom iterative algorithm, resembling gradient descent optimiza-

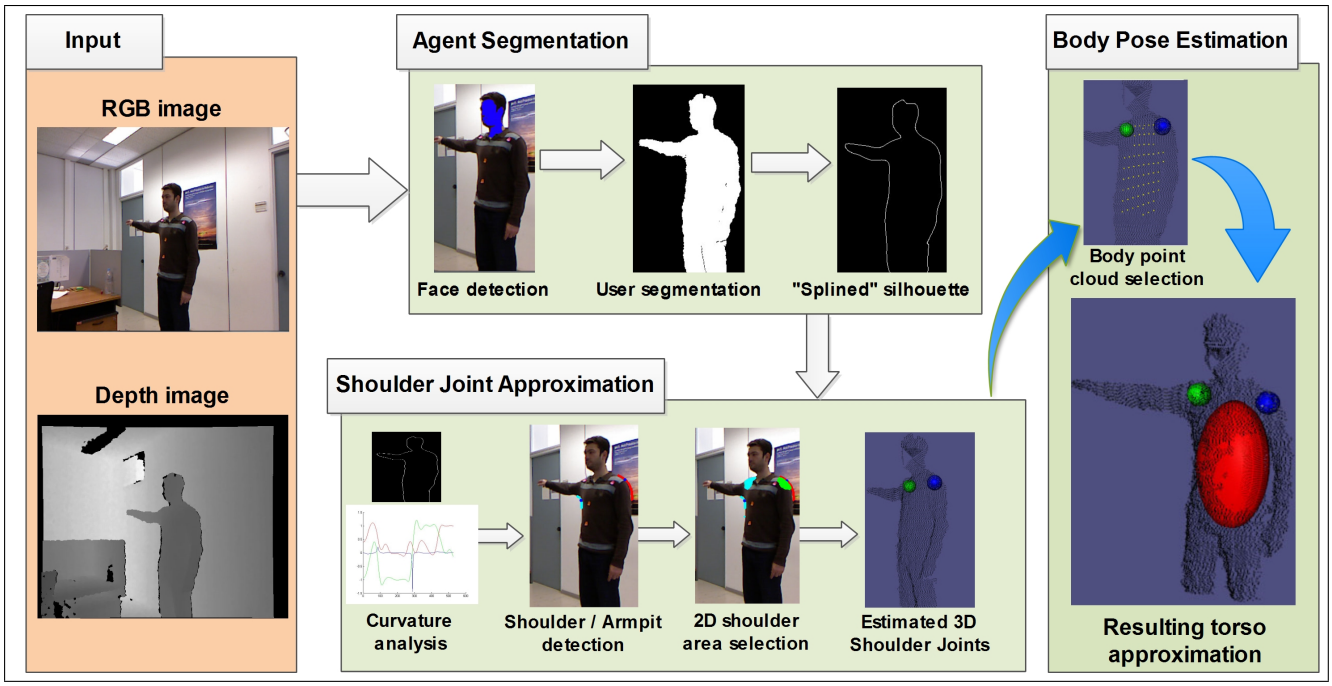



Figure 1. Methodology overview. Assuming the location of the face, the user is segmented from the rest of the scene and his silhouette is extracted. Points along the silhouette are then selected and used to estimate the location of the shoulder joints. Finally, the resulting shoulder joints define the area of the user's torso, on which the ellipsoidal torso model is fitted, in order to infer the upper body pose.

tion, is used to fit the ellipsoid model on the selected point cloud.

In the following we elaborate on the above described steps. It is noted that specific quantitative parameters regarding the human-body are required as input in our method. Since an initialisation step is intentionally excluded in our method, we rely on established anthropometric proportions to set these parameters relative to the human-body height; the latter is readily available as a by-product of the position of the detected face, assuming that the user is entering the scene in upright position.

### 2.1. Agent Segmentation

The first step in agent segmentation regards detection of the user's face, which is based on previous work on skin color detection and tracking [2, 3]. An initial background segmentation (using a standard segmentation technique [29]), is applied and foreground pixels are characterized according to their probability to depict human skin and grouped together into solid skin color blobs using hysteresis thresholding and connected components labeling. The location and the speed of each blob is modeled as a discrete time, linear dynamical system which is tracked using the Kalman filter equations, according to the propagated pixel hypotheses algorithm, as in [2]. Information about the spatial distribution of the pixels of each tracked object (i.e. its shape) is passed on from frame to frame using the object's current dynamics, as estimated by the Kalman filter. The density of the propagated pixel hypotheses provides the metric, which is used in order to associate observed skin-colored pixels with existing object tracks in a way that is aware of each object's shape and the uncertainty associated with its track (Fig. 2(b)). Finally, blobs are further classified into face and hands (Fig. 2(c)), and the belief about their class is maintained and continuously updated. For this purpose we employ an incremental probabilistic classifier, as in [3], using as input the speed, orientation, location and contour shape of the tracked skin-colored blobs. This classifier permits identification of hands and faces of multiple humans and is able to maintain hypotheses even in cases of partial occlusions. Based on the detected face, we extract the user's silhouette by depth thresholding using connected components labeling and the face centroid projected onto the point cloud as seed point. The silhouette is further refined via a cubic spline fitting to secure piecewise continuity.

### 2.2. Shoulder joint approximation

Given the location of the face, we select sets of points on the RGB silhouette, delineating possible shoulder areas. Selection is based on pose and scale invariant features satisfying certain geometric constraints. Naturally, the shoulder area on the silhouette is characterized by two body parts:

- *acromial or shoulder point.* That is the upper part of

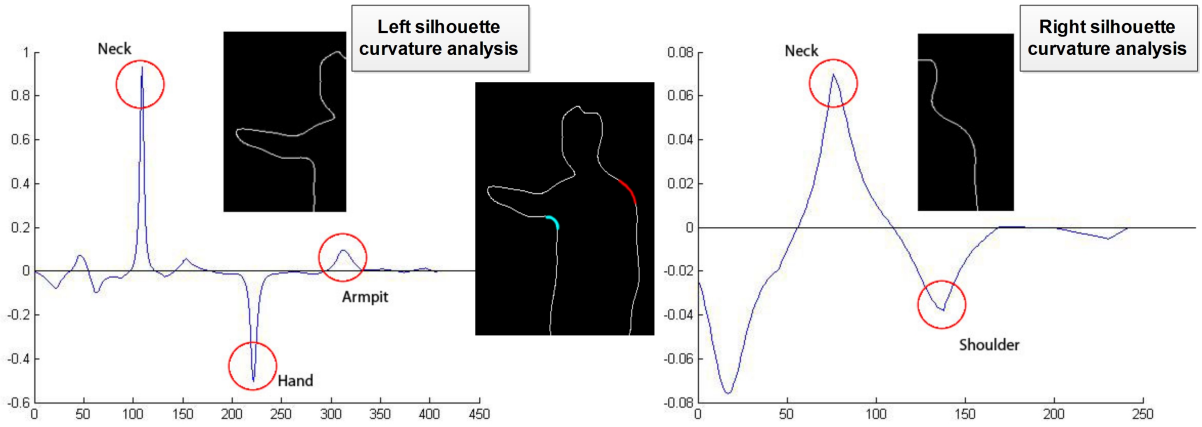
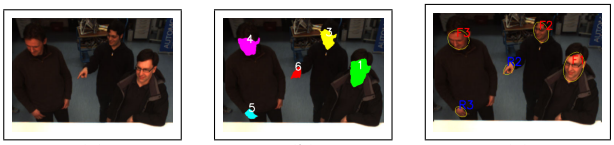



Figure 3. Curvature analysis for left and right parts of the user's silhouette.



(a) (b) (c)

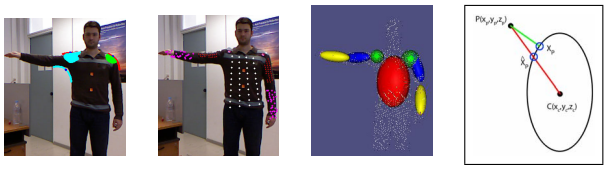
Figure 2. User tracking using skin colored blobs detection and tracking. (a) original RGB image, (b) skin colored hypotheses, (c) hypotheses classification into faces and hands.

the shoulder (red points in Fig. 3) and is robustly detectable for all configurations where the elbow is below the shoulder.

- *axillary or armpit*. That is the area "below" the shoulder (light blue points in Fig. 3). Similarly to *acromial*, the armpit is visible in most of the shoulder-elbow configurations and is scale and pose invariant.

Curvature analysis of the silhouette's spline results with the location of the aforementioned body parts, as depicted in Fig. 3. In order to limit our search space, we first locate an approximation to the neck, which exhibits the highest positive curvature response. This point steers the detection of either the shoulder, which exhibits high negative response, or the armpits with significant positive curvature. Anthropometric measurements, can be used in order to eliminate possible outliers, such as the hand on the left silhouette part of Fig. 3.

Given the extracted silhouette segments, a connected components procedure is performed to select the enclosing 2D shoulder area (Fig. 4(a)) which, in turn, provides the 3D point cloud around the shoulder joint. A standard least squares fitting algorithm is used to fit a sphere on the resulting point cloud (green spheres in Fig. 4(c)), which approximates the position of the shoulder joint. The estimated size of the shoulder joint is used to bound the radius of the sphere within certain limits and, thus, facilitate and speed up convergence.



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

Figure 4. (a) Selected shoulder area for sphere fitting. (b) Selected points along user's body and arms for ellipsoid fitting. (c) Approximated shoulder joints (green spheres), upper body (red ellipsoid), upper and forearms (blue and yellow ellipsoids respectively). (d) "Pseudo-distance" for ellipsoid fitting.

## Shoulder evaluation and tracking

Given variations of the human appearance and/or artifacts in the user's silhouette, multiple shoulder detections (shoulder candidates) are possible. Since arbitrary thresholding or elimination of certain candidates is not robust, we evaluate the shoulder candidates in pairs (left - right shoulder), against criteria imposed by the user's anthropometry. Such robust criteria in the studied case are the torso width and the distance between the face centroid and the midpoint of the two shoulders. In all our experiments, we set the former to be at approximately 0.25 of the estimated height, and the latter at approximately 0.18 of the estimated height.

The candidate pair with the highest score represents the selected shoulder joint locations. Subsequently, each shoulder is independently tracked over time by means of an Extended Kalman filter, thus providing a smooth trajectory of shoulder 3D locations.

## 2.3. Body pose estimation

In order to robustly utilize information in the 3D point cloud of the human body, we model the human torso using an arbitrarily oriented 3D ellipsoid. Given the approximate 3D location of the two shoulders, we sample points along the user's body (white dots in Fig. 4(b)) and attempt to fit a

3D ellipsoid on the resulting point cloud with a non-linear regression algorithm (Fig. 4(c)). In order to ensure sampling of points which indeed lie on the user's body, connected components with respect to spatial criteria (both 2D and 3D) are imposed on the selection mechanism. Such a representation adequately describes the morphology of the body and provides robust information about the full upper body pose of the user. However, fitting an ellipsoid to a point cloud requires extensive computations, which need to be relaxed for real time implementation. For this purpose, we adopt a custom optimization algorithm, inspired by the gradient descent technique for the typical least squares fitting method, which greatly accelerates computations without jeopardizing fitting accuracy.

An arbitrarily oriented 3D ellipsoid can be parametrically expressed as:

$$(X-C)^T R^T A R (X-C) = 1, \tag{1}$$

where $X = [x, y, z]^T$ is a 3D point on the ellipsoid's surface, $C = [x_c, y_c, z_c]^T$ is the center of the ellipsoid, $R$ is a 3x3 rotation matrix, denoting the 3D orientation of the ellipsoid, and $A$ is a 3x3 diagonal matrix with:

$$A = \begin{bmatrix} 1/a^2 & 0 & 0 \\ 0 & 1/b^2 & 0 \\ 0 & 0 & 1/c^2 \end{bmatrix}, \tag{2}$$

where $a, b$ and $c$ denote the three semi-axes of the ellipsoid.

The objective function to minimize, in order to fit an arbitrary ellipsoid on a given set of 3D data points, is given by:

$$E = \sum (X_P - P)^2, \tag{3}$$

which is the sum of squared distances between the point cloud and the ellipsoid (see Fig. 4(d)). However, estimating the closest point $X_P$ on the surface of the ellipsoid to a given 3D point $P$ of the point cloud, is a computationally demanding task. For this reason, we instead estimate a *"pseudo-distance"* given by the intersection of the ellipsoid and the $\vec{PC}$ ray (denoted as $\hat{X_P}$ in Fig. 4(d)) which is then used in the minimization process. $\hat{X_P}$ is easily computed by:

$$\hat{X_P} = P + t(C - P), \tag{4}$$

where $0 \leq t \leq 1$. Therefore, in order to estimate $\hat{X_P}$, we need to solve for $t$. By substituting $\hat{X_P}$ in Eq.( 1) we end up with:

$$t = 1 \pm \sqrt{1/K}, \tag{5}$$

where $K = (P-C)^T R^T A R (P-C)$. This formulation gives rise to an extremely efficient implementation, appropriate for real-time performance.

As seen above, an ellipsoid is described by 9 degrees of freedom (center, radii and orientation). Partial derivation of the objective function of Eq. 3, with respect to each of these variables, requires computation of rather complex analytical expressions, including hundreds of terms each. To overcome this, we resort to a finite differences approach, using binary search in the state space, with an adaptive step strategy. This technique resembles the gradient descent optimization, drastically reducing computational costs. Given an initial estimation, we move along the dimensions which reduce the response of Eq. 3, until no further improvement is possible or a predetermined termination condition is met. The location and orientation of the shoulder joints and the approximated torso size provide the algorithm with strong initial parameter values. This significantly enhances and speeds up convergence within only a small number of iterations (approx. 10), as has been experimentally verified.

### Towards arm tracking

The above described methodology can be extended in order to estimate and track the pose of the full upper body, including the arms. Although not examined thoroughly, early experiments indicate the plausibility of such a generalization, as verified by promising results. Similarly to the body, arm parts (upper and forearm) are modeled as ellipsoids, the size of which is approximated using anthropometric constraints w.r.t. the user height. As before, a connected components procedure is employed to select areas belonging to each arm part accordingly. Spatial information and self-exclusion rules combined with information on skin blobs classified as hands, from 2.1, are used in order to end up with the arms point clouds, on which the relevant ellipsoid is fitted, as seen in Fig. 4 (b, c).

## 3. Results

The method presented in this paper extracts information about the upper body pose of users which enter, leave, move or act freely in the scene. In order to evaluate our approach we conducted a series of experiments, with varying difficulty levels. Our experiments involve both single and multiple users, assuming several poses, in various relative to the camera positions and orientations, mostly in cluttered environments. In six of these experiments, we have attached markers on the shoulders of the users, in order to obtain ground truth and evaluate quantitatively our method. Additionally, a comparison against the widely used skeletonization module of OpenNI has been performed, using again the ground truth data. As will be verified, our method successfully managed to extract robust information about the user's upper body pose, coping with variations in scale, posture, distance from the camera, clothing and sex across users. Furthermore, the fact that no explicit initialization is required, results to high detection rates and also facilitates its employment in real life scenarios.

| | Ours | OpenNI | Ours | OpenNI | Ours | OpenNI |
|---|---|---|---|---|---|---|
| Test case | DP | | $\mu E$ | | $\sigma E$ | |
| 1 | **99.43%** | 88.07% | **6.73°** | 9.98° | **6.64°** | 8.31° |
| 2 | **97.06%** | 25.35% | **6.69°** | 12.31° | **8.40°** | 8.39° |
| 3 | **92.30%** | 83.81% | **4.88°** | 7.91° | **3.15°** | 7.13° |
| 4 | **99.47%** | 77.02% | **6.70°** | 10.42° | **4.49°** | 6.11° |
| 5 | **98.83%** | 0% | **3.48°** | – – – | **2.94°** | – – – |
| 6 | **93.76%** | 0% | **8.92°** | – – – | **8.80°** | – – – |
| Average | **96.80%** | 68.56% | **6.23°** | 10.15° | **5.73°** | 7.48° |

Table 1. Comparative statistics. DP= percentage of frames where body orientation was estimated, $\mu E$=mean orientation error throughout the sequence, $\sigma E$= standard deviation of error. Columns in bold-font refer to the proposed method, whereas normal-font refer to OpenNI.

## Quantitative Results

As already mentioned, we conducted a series of experiments in order to extract quantitative information about the performance of our method. More precisely, we tested 6 sequences (summing to a total of more than 5000 frames) of single users performing a variety of poses, in an office environment. We used prominent colored markers, on the user's clothing, in order to unambiguously detect the actual location of the shoulders and provide ground truth for the body orientation. Additionally, we also tested the skeletonization module of the OpenNI against the ground truth, and compared the results with those of our methodology, considering the body orientation (angular) error.

Fig.5 shows a variety of indicative resulting images from the ground truth sequences. The user is roughly turned to the camera and performs a series of poses, by raising either or both hands and rotating, bending or stretching his body with respect to the camera. The actual (ground truth) orientation (in degrees) and the estimated ones by the two methods are superimposed on the images at the upper part of each one. Additionally, the thick white arrow depicts the actual orientation, while the green and red ones illustrate the estimation of our methodology and OpenNI, respectively.

The ground truth sequences were used in order to derive a set of statistics which attest for the robustness and effectiveness of our methodology. For each sequence, we gathered three types of information: percentage of frames where an estimation has been provided (DP), mean error of the estimated from the actual orientation ($\mu E$) and standard deviation of the error ($\sigma E$). The average values shown, are calculated over only the frames where an estimation has been derived. As can be observed in Table 1, our method performed significantly better in all cases, achieving very small average error. Additionally, the fact that no initialization is required results to fast pose estimation and high detection rates ($> 90\%$). On the other hand, OpenNI couldn't effectively cope with the initialization problem, which consequently led to either low detection rates (e.g. $25\%$ in test case 2) or no detection at all for a whole sequence (e.g. test cases 5-6 in Table 1). Some of these cases are depicted in Fig.6.

## Qualitative Results

In addition to the quantitative experiments, we extensively tested our method in numerous scenarios, by modyfing the environment, the number of users in the scene and the type of interaction. On top of that, we also tested our method on a robotic bar-tending simulating scenario (rightmost image in Fig. 7), serving as attention seeking detector and providing input for the robot's social planner about whether or not a user needs to be served. As illustrated in Fig. 7, the proposed approach could effectively extract the body pose in all cases, dealing with both male and female users, of varying postures and proportions, performing numerous acts. Moreover, with a rather straightforward C++ implementation, with no multi-threading or GPU programming, on a standard commercial PC (Intel i7 3.5 GHz with 8 GBs or RAM) we managed to achieve real-time execution, reaching a speed of approximately 25 frames per second for a single user and 18 frames per second for two users in the scene.

### Results on arm pose tracking

Fig. 8 depicts initial results on arm pose recovery and tracking. Although still in a preliminary phase, the results are very promising, as our system was able to estimate difficult arm configurations, such as the one illustrated in the middle image, where the arms are in front of the body.

## 4. Discussion

In this paper we presented a fast and robust methodology for human-torso 3D pose extraction. Contrary to many state-of-the-art approaches, our method avoids an explicit pose initialization phase, which is not possible in complex, real-life scenarios. Extensive testing in a large variety of set-ups, involving multiple users, arbitrarily moving in the scene, has proved the effectiveness of the pro-

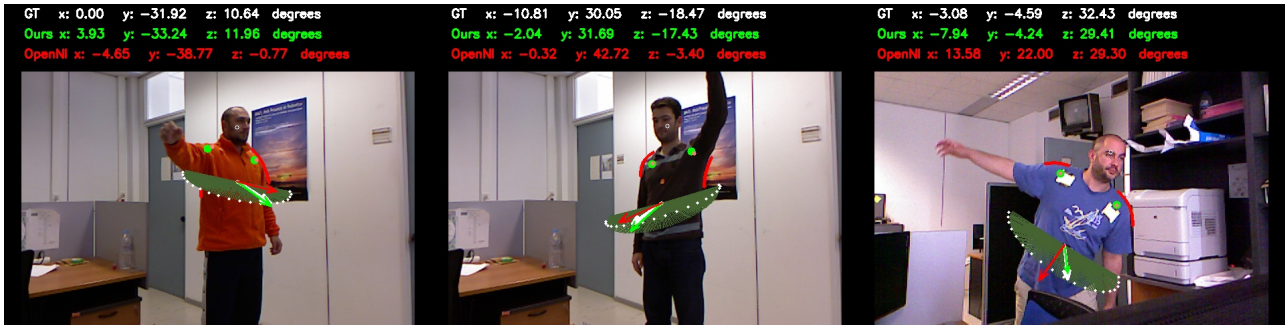



Figure 5. Comparison with OpenNI skeletonization module. In each image, the orientation (in degrees), ground truth and the estimation of each method is shown at the upper part of each image. The thick white arrow depicts the ground truth orientation, the green one depicts the estimated orientation of our methodology, and the red arrow depicts the orientation estimated by OpenNI.

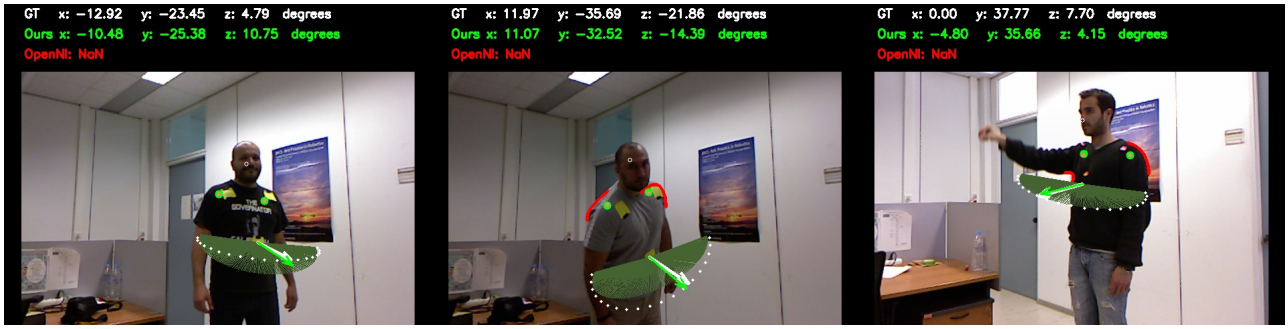



Figure 6. Successful body pose estimation compared to OpenNI failure to register the user.

posed approach. More importantly, quantitative experimentation against ground truth data, and comparative results with those of OpenNI have revealed superior and robust performance in demanding scenarios.

Our planned future work regards extension of the method to cope with full-body pose recovery. The current body pose tracking results, as well as the promising arm tracking preliminary experiments, attest to the appropriateness of the method for the task at hand. Nevertheless, important issues, such as occlusion or interaction (between agents) handling are still under investigation and will be the immediate target of our future research.

## References

[1] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Consumer Depth Cameras for Computer Vision*, pages 71–98. Springer, 2013. 2

[2] H. Baltzakis and A. Argyros. Propagation of pixel hypotheses for multiple objects tracking. *Advances in Visual Computing*, pages 140–149, 2009. 3

[3] H. Baltzakis, M. Pateraki, and P. Trahanias. Visual tracking of hands, faces and facial features of multiple persons. *Machine Vision and Applications*, pages 1–17, 2012. 10.1007/s00138-012-0409-5. 3

[4] J. Charles and M. Everingham. Learning shape models for monocular human pose estimation from the microsoft xbox kinect. In *Proc. International Conference on Computer Vision (ICCV)*, page 12021208, 2011. 1, 2

[5] M. Corp. Kinect for xbox 360. 1

[6] S. Escalera. Human behavior analysis from depth maps. In *Articulated Motion and Deformable Objects*, volume 7378 of *Lecture Notes in Computer Science*, pages 282–292. Springer Berlin / Heidelberg, 2012. 2

[7] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *International journal of computer vision*, 87(1-2):75–92, 2010. 2

[8] A. Gaschler, K. Huth, M. Giuliani, I. Kessler, J. de Ruiter, and A. Knoll. Modelling state of interaction from head poses for social human-robot interaction. In *Proceedings of the Gaze in Human-Robot Interaction Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012)*, Boston, USA, 2012. 1

[9] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: A multi-view approach. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'96,*, 1996. 1

[10] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *IEEE Intl. Conference on Computer Vision (ICCV)*, pages 415–422, 2011. 1, 2

[11] D. Grest, J. Woetzel, and R. Koch. Nonlinear body pose estimation from depth images. *Pattern Recognition*, pages 285–292, 2005. 2

[12] A. Hernandez-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, and S. Escalera. Graph cuts optimiza-

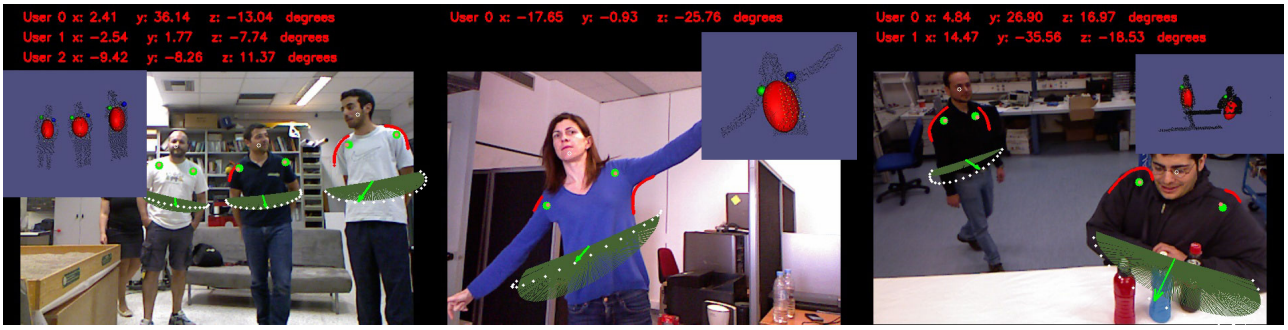



Figure 7. Qualitative results of our body orientation estimation methodology. Body orientation is superimposed on each image.


Figure 8. Preliminary results on arm tracking.

tion for multi-limb human segmentation in depth maps. In *Proc. Computer Vision and Pattern Recognition*, 2012. 2

[13] B. Holt, E.-J. Ong, H. Cooper, and R. Bowden. Putting the pieces together: Connected poselets for human pose estimation. In *Intl. Conference on Computer Vision*, 2011. 2

[14] S. Ilic and P. Fua. Generic deformable implicit mesh models for automated reconstruction. In *First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003. HLK 2003.* IEEE, 2003. 2

[15] V. John, E. Trucco, and S. Ivekovic. Markerless human articulated tracking using hierarchical particle swarm optimisation. *Image and Vision Computing*, 2010. 2

[16] E. Kalogerakis, A. Hertzmann, and K. Singh. Learning 3d mesh segmentation and labeling. *ACM Transactions on Graphics (TOG)*, 29(4):102, 2010. 2

[17] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, pages 90–126, 2006. 2

[18] D. Ognibene, E. Chinellato, M. Sarabia, and Y. Demiris. Towards contextual action recognition and target localization with active allocation of attention. *Biomimetic and Biohybrid Systems*, pages 192–203, 2012. 1

[19] OpenNI. 1, 2

[20] O. Ozturk, T. Yamasaki, and K. Aizawa. Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference*, pages 1020–1027. IEEE, 2009. 1

[21] M. Pateraki, H. Baltzakis, and P. Trahanias. Using dempster's rule of combination to robustly estimate pointed targets. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1218–1225. IEEE, 2012. 1

[22] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun. Real-time identification and localization of body parts from depth images. In *Robotics and Automation (ICRA), 2010 IEEE International Conference*, pages 3108–3113. IEEE, 2010. 2

[23] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1):4–18, 2007. 2

[24] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *In Proc. Computer Vision and Pattern Recognition*, 2011. 1, 2

[25] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–421. IEEE, 2004. 2

[26] C. Stoll, N. Hasler, J. Gall, H. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 951–958. IEEE, 2011. 2

[27] X. Wei, P. Zhang, and J. Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics (TOG)*, 31(6):188, 2012. 2

[28] Y. Zhu and K. Fujimura. Constrained optimization for human pose estimation from depth sequences. *Computer Vision–ACCV 2007*, pages 408–418, 2007. 2

[29] Z. Zivkovic. Improved adaptive gausian mixture model for background subtraction. In *Proc. of the International Conference on Pattern Recognition*, 2004. 3