Robust articulated upper body pose tracking under severe occlusions

Markos Sigalas[1,2], Maria Pateraki[1] and Panos Trahanias[1,2]

Abstract— **Articulated human body tracking is one of the most thoroughly examined, yet still challenging, tasks in Human Robot Interaction. The emergence of low-cost real-time depth cameras has greatly pushed forward the state of the art in the field. Nevertheless, the overall performance in complex, real life scenarios is an open-ended problem, mainly due to the high-dimensionality of the problem, the common presence of severe occlusions in the observed scene data, and errors in the segmentation and pose initialization processes.**

In this paper we propose a novel model-based approach for markerless pose detection and tracking of the articulated upper body of multiple users in RGB-D sequences. The main contribution of our work lies in the introduction and further development of a virtual *User Top View*, a hypothesized view aligned to the main torso axis of each user, to robustly estimate the 3D torso pose even under severe intra- and inter-personal occlusions, exempting at the same time the requirement of arbitrary initialization. The extracted 3D torso pose, along with a human arm kinematic model, gives rise to the generation of arms hypotheses, tracked via Particle Filters, and for which ordered rendering is used to detect possible occlusions and collisions.

Experimental results in realistic scenarios, as well as comparative tests against the NiTE™ user generator middleware using ground truth data, validate the effectiveness of the proposed method.

I. INTRODUCTION

Markerless articulated body tracking constitutes a challenging and highly important task in Robotic Vision, targeting a variety of applications. The latter include complex Human Robot Interaction (HRI) tasks [1], such as user interaction with robotic guides or service robots [2], and application scenarios in relevant sectors, such as gaming and augmented reality. Recently, the introduction of low-cost real-time depth (RGB-D) cameras, such as the Kinect™ sensor [3], has significantly facilitated the task at hand, giving rise to fast and accurate pose recovery approaches and, thus, pushing forward the state-of-the-art (e.g. [4], [5], [6]). Nevertheless, and despite the fact that the majority of the recently developed approaches are quite effective in controlled or semi-controlled environments, in more complex cases, involving multiple users moving and (inter-)acting arbitrarily, performance may be limited.

In a large number of body pose tracking methodologies, there is the inherent requirement of an initialization phase. This can be done either explicitly, by having the user stand at a specific pose (e.g. T-pose), or implicitly, requiring a certain amount of frames to register the user. However, in real life scenarios where users move, act and interact freely, an initialization phase is not always possible. In such cases of naturalistic interactions, the problem of frequently occurring intra- and inter-person occlusions, namely occlusions imposed across body parts of the same user or across different users, respectively, may additionally deteriorate performance. In fact, estimating the human pose when a person is partially or heavily occluded in the scene remains challenging and is of utmost importance in real-world applications [7].

In this work, we present a novel markerless articulated upper body tracking methodology, able to overcome key limitations imposed in complex, free-form interaction scenarios. We mainly focus in cases where multiple users enter, exit or move freely across the scene and independently act and interact. Within this context, we are interested in estimating the upper body configuration, including torso and arms, under the assumption that no initialization phase is possible and that the pose recovery and tracking should remain unaffected from partial intra- and inter-person occlusions.

The employed upper body model consists of five parts; the torso and the two (left-right) upper and forearms [8]. The torso is considered as a rigid body with 6 Degrees of Freedom, represented by an elliptic cylinder, while the arms are modeled using a kinematic model similar to the one presented in [9]; the arm parts (upper and forearm) are both represented as cylinders.

A brief overview of the developed methodology is in order. Initially, we detect users by classifying skin-colored blobs into faces and palms. Based on the detected face locations, we perform ordered-based segmentation of each user from the rest of the scene. Next, each user is evaluated according to the depth ordering, against other users *in front* of him, for possible occlusions. We then estimate the *User Top View (UTV)*, a hypothesized view aligned to the main axis of a users torso, based on the *minimum projection ratio criterion*. *UTV* is used to determine the torso pose and, effectively, the location of the two shoulders. Shoulder locations, in conjunction with the detected palms and a set of anthropometric proportions, are used to generate a set of configuration hypotheses for each arm, tracked by a separate particle filter. Ray tracing is used to render each of the body parts (particles) and to detect and handle occlusions or collisions with prior evaluated users and parts. Possible detected occlusions and collisions are further used, together with the kinematic model workspace, to constraint the hypotheses space. Finally, a hypothesized depth map is generated for each arm configuration, which is compared against the observed one in order to evaluate each hypothesis,

*This work has been partially supported by the EU Information Society Technologies research project James (FP7-045388)

[1]Institute of Computer Science, Foundation for Research and Technology - Hellas, Heraklion, Crete, Greece

[2]Department of Computer Science, University of Crete, Greece

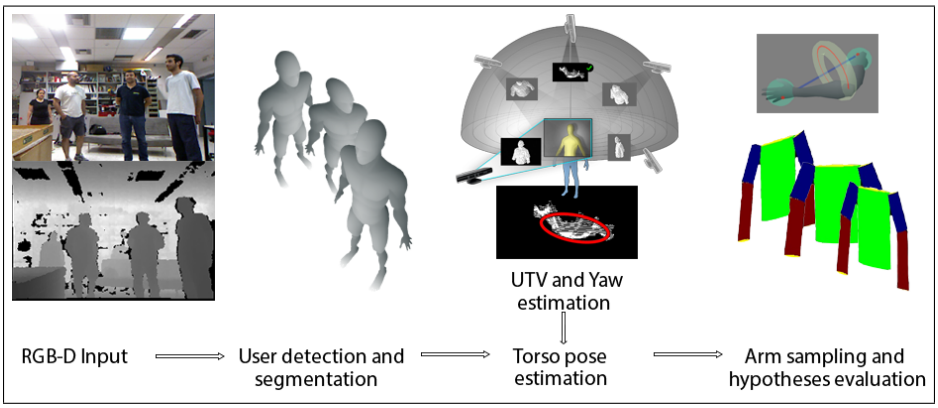


Fig. 1. Schematic overview of the proposed methodology.

resulting to a full upper body pose estimation.

To evaluate our approach we performed extensive experiments using realistic interaction scenarios in different environments, with varying number of users. We further compared our method against the NiTE™ [6], [10] skeletonization tool, using ground truth derived from marker-based sequences. Both qualitative and quantitative results are presented that attest on the effectiveness and performance of the proposed methodology.

Related work

State-of-the-art body pose estimation and tracking approaches are thoroughly reviewed in [11], [12], [13], while the emergence of real-time depth sensors has stimulated new research [7]. The most widely used approach is the learning-based approach of Shotton [5] used by Microsoft in its Kinect for Windows SDK. The method utilizes Random Forests, which are employed in order to perform a per pixel body part classification, using a large synthetic training dataset containing various body configurations. Extensions of this work [14], [15] have managed to enhance the initial machine learning method by providing faster and more accurate results. The OpenNI framework [6] also includes a body pose algorithm in its NiTE [10] middleware. To the best of our knowledge, the NiTE algorithm has not been published, but its functionality and performance approach the ones of the algorithm of Shotton [5].

Similarly to the above, Ye et al. [16] and Shen et al. [17] claim to provide accurate pose estimation in cases where occlusions are present. The authors in [16] utilise the coherent drift point (CDP) algorithm to solve non-rigid point registration; in [17] the authors use an exemplar-based method to learn an inhomogeneous systematic bias for body pose correction and tagging. Hernandez et al. [18] use Graph Cuts optimization to classify pixels to seven body parts, while Baak et al. [19] follow a data-driven hybrid strategy, combining local optimization and global retrieval techniques. In the same context, Probabilistic Graphical Models [20] as well as hybrid approaches, such as Connected Poselets [21], have also been used to infer the body pose.

Grest et al. [22] use Iterative Closest Point to extract and track the skeleton while Zhu and Fujimura [23] build heuristic detectors to locate upper body parts (head, torso, arms). Similarly to the latter, Jain et al. [24] locate the upper body by sliding template matching and use distance transform analysis to infer the pose of the arms. Moreover, in [25], vertices are classified and segmented into different body parts, while, Plagemann et al. [26] build a 3D mesh in order to form geodesic maps for the detection of the head, hand and foot.

Evidently, a large body of research deals with the problem of articulated body pose extraction and tracking. However, limiting factors are still present, especially when dealing with complex, realistic interaction scenarios. One such limitation regards the inherent requirement for an initialization period, either explicitly, demanding a specific predefined pose [22], [23] or implicitly, by registering and tracking the user over a time-window [5], [10]. Moreover, learning-based approaches, as [5], [10], are characterized by the absence of a kinematics coherence in the provided poses. Even more importantly, a serious drawback of most of the state-of-the-art approaches is the limited ability to cope with instances of severe occlusions, and hence the inferior performance in such cases. Although some works have attempted to address self-imposed occlusions [16], coping with inter-person occlusions remains problematic [12].

Contributions

In this paper, we address the above-mentioned shortcomings and propose a methodology for robustly and accurately inferring the upper body pose of multiple users, which move, act and interact freely in naturalistic scenarios. Our main contribution is the introduction and development of the *User Top View (UTV)* as a robust indicator of the 3D body pose. Additionally, ordered rendering of each user together with the employed kinematic arm model, facilitate robust and effective handling of collisions and occlusions across different parts and/or different users.

II. Methodology

A schematic overview of the proposed methodology is depicted in Fig. 1. Once users are detected and segmented

from the scene, the depth of each face centroid is utilized to determine the ordering of user evaluation, and provide an initial estimation about existing occlusions. Then, the 3D configuration of the torso is derived from the estimated *UTV*, which further steers estimation of the shoulder joints. Given the location of shoulders and the corresponding detected palms, we generate a set of arm-hypotheses according to the employed kinematic model. Each arm hypothesis is rendered and checked against possible occlusions and collisions. Finally, the generated depth map is compared to the observed depth map and used to evaluate each hypothesis. Overall, the proposed methodology involves four steps, which are executed iteratively while tracking multiple users:

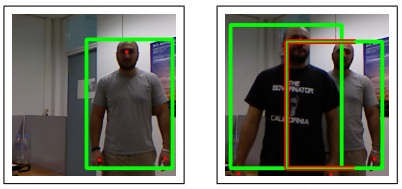
- **Agent segmentation and ordering**. Based on face and palm detection, users are detected, ordered, segmented from the scene and checked for possible occlusions. Each user is subsequently processed according to the resulting depth order.
- **Torso pose estimation**. *UTV* is used to estimate the torso configuration for each user. 2D ellipse approximation, on the hypothesized view projections, is used to derive the orientation of the torso's main axis and also locate the shoulder joints.
- **Arm hypotheses generation**. Given the shoulder positions and the detected user palms, a set of arm hypotheses is generated, constrained by the kinematic model.
- **User rendering and hypothesis evaluation**. Ray casting is used to render the body parts, and further constraint the arm hypotheses. For each configuration, a depth map is generated, which is compared in turn against the actual depth map in order to evaluate the hypothesis.

Throughout the process, specific quantitative parameters regarding the human-body are used, namely the torso size and the lengths of each arm part (upper and forearm). Since an initialization phase isn't available, we rely on established adult anthropometric proportions [27], [28] to set these parameters relative to the human-body height; the latter is readily available as a by-product of detected 3D face position.

A. Agent segmentation and ordering

The first step of our methodology involves the detection and segmentation of the users, as well as ordering them w.r.t. their depth. To detect users, we rely on face and palm classification based on human skin color tracking, as described in [29]. The 3D face location, besides facilitating user detection, has a twofold use. As mentioned above, it gives rise to an approximation of the anthropometric parameters (lengths) of the model. Moreover, it is readily exploited to obtain an ordering of the users according to their distance (depth) from the camera.

User segmentation is performed via a standard connected components algorithm operating onto the depth channel, thus providing the corresponding point cloud. The segmented area of each user is tracked using the associated bounding box, as shown by the green rectangle in Fig. 2(a). Whenever two, or more, bounding boxes collide (overlapping one another or being adjacent) it is assumed as an indicator for possible intra-person occlusion (red outline in Fig. 2(b)). This information is further utilized in the following steps.



(a) (b)

Fig. 2. (a) Bounding box of the segmented area. (b) Overlapping area bounding boxes as an indicator for possible occlusion.

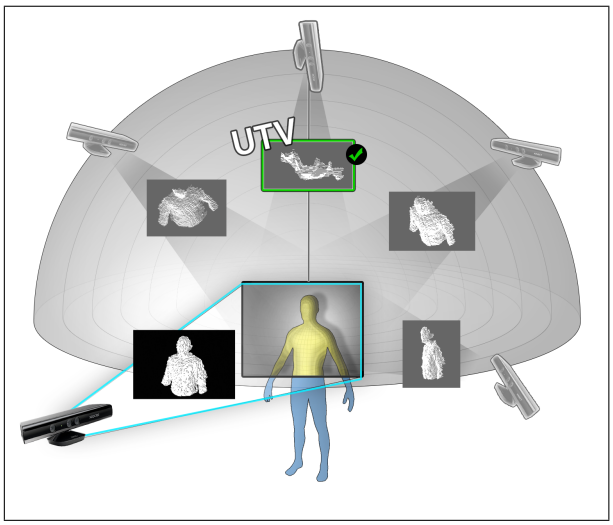
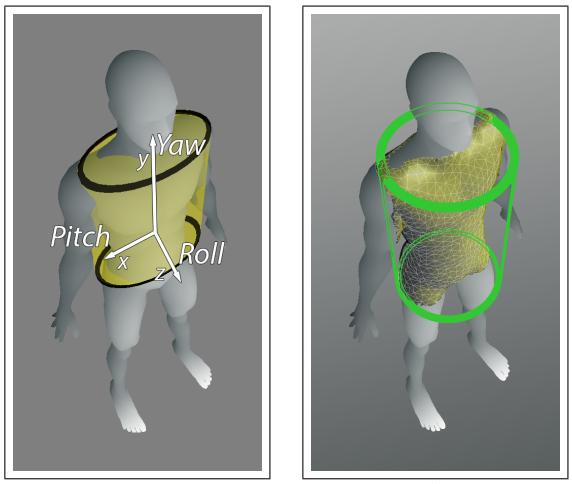


Fig. 3. Virtual camera moving on a semi-sphere. For each view, the corresponding re-projection is depicted. The hypothesized view that minimizes the projection ratio criterion, is considered to be the *UTV*.

B. Torso pose estimation

Given the location of the face and the corresponding segmented point cloud of each user, we subsequently obtain a top-view re-projection of the point cloud. The sought virtual top-view camera, termed *User Top View* (*UTV*), has its optical axis aligned to the user's torso axis and its formulation constitutes a main contribution of this work. Based on the estimated *UTV*, we select the 3D points of the torso area which, in turn, are used to derive the 3D pose of the torso.

To achieve this, we assume a virtual camera allowed to move on a semi-sphere above the user, as depicted in Fig. 3; the optical axis of the virtual camera remains normal to the semi-sphere surface for all assumed camera positions. For each such position, we select points belonging to the *hypothesized torso* and re-project them on the virtual image plane. In order to delineate the area that contains the torso points, we represent the user-torso as an elliptic cylinder, the size of which is determined by the anthropometric measurements. In Fig. 4(a), the torso representation is illustrated, along with



(a) (b)

Fig. 4. (a) Elliptic cylinder model for the torso. (b) Circular cylinder casting to select points and compute the ratio criterion. Yellow points represent the point lying inside the cylinder, namely the P_{area}.

the orientations about the x, y and z axis, denoted as pitch, yaw and roll, respectively. However, since the orientation about the torso axis, namely the yaw, is unknown, we relax this parameter and initially compute only the pitch and roll angles, by selecting points which lie inside a **circular** cylinder, the axis of which is aligned to the virtual camera's optical axis, as illustrated in Fig. 4(b).

To evaluate the virtual views, for each top-view re-projection of a user's body, we employ the *projection ratio criterion* given as:

$$P_{ratio} = \frac{P_{proj} * P_{cyl}}{P_{area}^2} \tag{1}$$

where P_{cyl} is the surface area of the hypothesized cylinder, P_{area} is the total number of 3D points inside the cylinder and P_{proj} is the number of point projections on the image plane. In other words, assuming that the torso is the biggest part of a human body, (1) demands for the minimum number of point projections and the maximum body area (3D points) coverage simultaneously.

Evidently, the virtual view with the minimum P_{ratio} is the one obtained when the virtual camera on top of the user has its optical axis coinciding with the main axis of the human torso (see Fig. 3). This plausible and intuitive assumption has also been experimentally verified in our work; we conducted a series of experiments involving users that assumed various body configurations. In all cases, ground truth data was available, using markers attached to the users. Given the torso orientation (provided by the markers), the virtual camera assumed many densely-placed locations on a semi-sphere on top of the user and with axis of symmetry the main torso axis. For all such camera-configurations the re-projection of the point cloud onto the camera image plane has been estimated. Fig. 5 shows the P_{ratio} in comparison against ground truth data, averaging over all experiments. Assuming that ground truth lies at $(0, 0)$, where the "roll angle" and

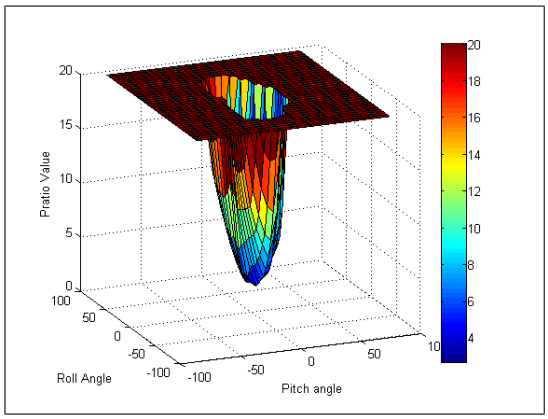


Fig. 5. Plot demonstrating the value of P_{ratio} for various views. Each axis represents difference between the ground truth and the estimated virtual view orientation. P_{ratio} becomes minimum when it coincides with the ground truth.

"pitch angle" axes depict the difference between the hypothesis and the corresponding ground truth orientation, we can observe that the projection ratio drops as the hypothesized view approaches *UTV*, and becomes minimum at *UTV*.

Occlusion handling: During agent segmentation, in II-A, we determined whether there exist possible occlusions across users. If this is the case for the currently examined user, the above presented ratios are adjusted according to the percentage of occlusion P_{occl} of the virtual cylinder. For the calculation of P_{occl}, the hypothesized circular cylinder is projected onto the image plane of the actual camera. Projected points that lie inside the bounding boxes of the users "in front" of the currently examined one, are deemed as occluded points of the user and contribute to the calculation of P_{occl}; the latter is the ratio of occluded points over the total number of cylinder projected points. In turn, P_{ratio} becomes

$$P_{ratio} = (1 - P_{occl}) * \frac{P_{proj} * P_{cyl}}{P_{area}^2} \tag{2}$$

P_{occl} can be considered as a normalization factor, aiming to compensate for the ratio changes caused by the occlusions. By adjusting the *projection ratio criterion*, we favor occluded hypotheses which would satisfy the initial projection ratio criterion (1), if they weren't occluded. Effectively, the *normalized* projection criterion guarantees high detection rates and accuracy even in cases of severe occlusions.

UTV *refinement:* The above provides a universal and readily implementable criterion for estimating *UTV*. In practice though, since the torso axis is not available, the semi-sphere for placing the virtual camera can not be accurately defined. In our implementation we overcome this by using the centroid of the (already available) user's face and placing the semi-sphere accordingly. The latter may introduce small errors in the estimation of *UTV*, mostly due to varying head positions and orientations, affecting the center of the camera pivot.

To cope with this we employ a *UTV*-refinement step. More specifically, we consider a small neighbourhood, around the computed *UTV*, that is deemed to contain the actual *UTV*.

TABLE I

DENAVIT-HARTENBERG PARAMETERS FOR THE 4-DOF MODEL OF THE HUMAN ARM EMPLOYED IN OUR APPROACH.

i	θ_i	α_i	a_i	d_i	range
1	θ_1	$-90°$	0	0	$-90° \dots 90°$
2	$\theta_2 - 90°$	$-90°$	0	0	$-90° \dots 90°$
3	$\theta_3 + 90°$	$+90°$	0	l_u	$-230° \dots 90°$
4	θ_4	$-90°$	0	0	$0° \dots 145°$
5	$\theta_5 = 0$	$+90°$	0	l_f	$0° \dots 350°$

However, the virtual camera pivot is an approximation of the neck centroid instead of the that of the face. The neck is approximated by examining *narrow slices* along the virtual camera axis. For each of the resulting virtual cameras, P_{ratio} is recalculated and, thus, *UTV* estimation is refined.

Yaw calculation: Based on the derived *UTV*, torso orientation is readily available as the orientation assumed by the *UTV*-axis. More specifically, the *UTV*-axis defines the two degrees-of-freedom for the torso orientation. The third degree-of-freedom, namely torso rotation around the *UTV*-axis, is obtained by fitting a 2D ellipse on the torso points re-projected on *UTV* with center the previously approximated neck. Due to the nature of data, typical ellipse surface fitting would provide poor results in cases of occlusions and significant torso rotation angle. Instead, we fit a line on the re-projected points, and demand for the ellipse major axis to be parallel to that line. By doing so, torso orientation remains unaffected from arbitrary movements and/or the presence of occlusions. The two ellipse extrema are, finally, considered as the locations of the corresponding user shoulders.

C. Arm hypotheses generation

Each arm is modeled using the 4-DoF kinematic model shown in Fig. 6(a), similar to the one presented in [9], and described by the Denavit-Hartenberg parameters shown in Table II-C. Angles $\theta_1, \theta_2, \theta_3$ refer to the 3 DoFs of the shoulder joint, while angle θ_4 refers to the DoF of the elbow joint. l_u and l_f refer to the lengths of the upper and forearm, respectively. Since we are not interested in the orientation of the palm, we practically discard the fifth angle by setting $\theta_5 = 0$.

Typically, given the location of the shoulder and palm, one has to solve the inverse kinematic equations in order to come up with the possible positions of the elbow. However, this involves performing a series of expensive calculations, which, in our case, could be prohibiting for real time execution.

Instead we use a two-way mapping between arm configuration and spherical coordinates, as depicted in Fig. 6(b). Given a shoulder joint, the elbow is allowed to move freely along the surface of a sphere (more accurately, along a part of the sphere surface, constrained by the kinematic workspace) with center the shoulder itself and radius the length of the upper arm, l_u. Equivalently, the palm moves along (part of) the surface of a sphere with center the elbow location and radius the length of the forearm, l_f. This mapping is performed offline and facilitates switching rapidly between the two reference systems, i.e. kinematic $(\theta_1 \dots \theta_4)$ and spherical (ω, ϕ), without excessive computations.

To cope with possible data noise or errors introduced from previous computations, the detected shoulders and palms are represented as 3D gaussian distributions, with mean value μ the location of the corresponding joint. In the case of shoulders, the standard deviation σ is the shoulder joint size, derived from the anthropometric proportions. In the case of palms, σ is inversely proportional to the confidence of the skin classifier, according to a predefined scale factor. A visual representation of the sampling procedure is given in Fig. 6(c).

In order to generate arm hypotheses we sample from the corresponding distributions, to end up with a hypothetical shoulder-palm pair. Given the 3D location of the shoulder $\vec{S}$ and the palm $\vec{P}$, the elbow is constrained to lie on a 3D circle around the $\vec{SP}$ vector with center $\vec{C}$ (red disc in Fig. 6(c)), given by:

$$\vec{C} = \vec{S} + \frac{i}{l}(\vec{SP}) \tag{3}$$

where

$$i = \frac{l^2 + {l_u}^2 - {l_f}^2}{2l} \tag{4}$$

$l = \|\vec{SP}\|$, l_u is the upper arm length and l_f the forearm length. Given two unit vectors, $\vec{v_1}$ and $\vec{v_2}$, perpendicular to $\vec{SP}$ and to each other, we estimate possible elbow positions $\vec{E_k}$ as:

$$\vec{E_k} = \vec{C} + r(\vec{v_1}\cos(\theta_k) + \vec{v_2}\sin(\theta_k)) \tag{5}$$

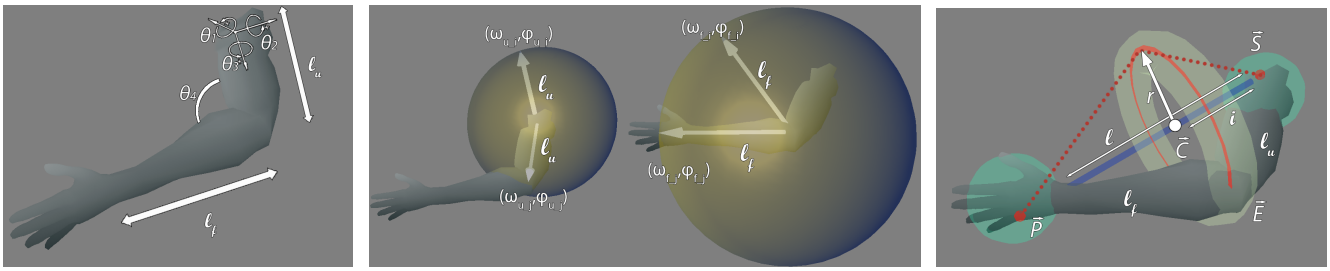
where $r = \sqrt{{l_u}^2 - i^2}$ is the circle radius and θ_k is the angle about the $\vec{SP}$, constrained by the kinematic model.

The above arm hypothesis sampling procedure is illustrated in Fig. 6(b). Shoulder and palm pairs are sampled from the corresponding distributions, illustrated as spheres. For each pair, a 3D circle (depicted by red color) about the shoulder-palm axis defines the elbow legal locations. In total, the elbow workspace forms a doughnut-like area, constrained by the kinematic model, from which samples are drawn uniformly, to produce the arm hypotheses set. For each arm, the corresponding hypotheses are tracked over time by means of a separate particle filter, ensuring temporal and spatial consistency.

D. User rendering and hypothesis evaluation

Each particle is in turn rendered in order to generate a hypothetical depth map, which, compared to the observed depth map, is used to evaluate the hypothesis. Arm rendering follows the user depth ordering, discussed in II-A. This means that the body parts of already examined users (users "in front" of the current one) are already rendered in the scene. Therefore, each arm hypothesis is also checked against occlusions and collisions with existing parts, of the same or different users.

Body parts rendering is done using ray tracing and the projective geometry of quadrics as described in [30]. Briefly,



(a) (b) (c)

Fig. 6. (a) Kinematic model of the arm. (b) Mapping between kinematic angles and spherical coordinates. (c) Shoulder-Palm pair sampling. Shoulders and palms are drawn from normal distributions. The elbow distribution have a torus-like shape, from which we draw samples uniformly.

we model the torso as an elliptical cylinder and the arm parts (upper and forearm) as circular cylinders. Cylinders, here aligned with the y-axis, can be represented in homogeneous coordinates as a symmetric 4×4 matrix Q:

$$Q = \begin{bmatrix} \frac{1}{a^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{b^2} & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \tag{6}$$

where a and b are the semi-major and semi-minor axes, respectively. In case of circular cylinders (e.g. arms) $a = b$. Their surface is defined by the points X which satisfy the equation:

$$X^T Q X = 0 \tag{7}$$

Additionally, a pair of planes $\pi 0, \pi 1$, parallel to the xz-plane can be described by

$$Q\pi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -\frac{(y_0+y_1)}{2} \\ 0 & 0 & 0 & 0 \\ 0 & -\frac{(y_0+y_1)}{2} & 0 & y_0 y_1 \end{bmatrix} \tag{8}$$

In order to render the hypothesized scene, we cast a ray for each image pixel and find its intersections -if any- with the existing quadrics. The camera center 0 and a point x in image coordinates, define a ray $X(t) = [x, t]^T$ in 3D space, calculated using the camera intrinsic parameters. The point of intersection of the ray with the quadric can be found by substituting X by $X(t)$ in (7)

$$X(t)^T Q X(t) = 0 \tag{9}$$

and solving for t. In case the ray intersects the quadric, (9) has two solutions, while in case the ray is tangent to the quadric (9) has a unique solution. Apparently, if the ray and the quadric do not intersect, (9) does not have a real solution. In case of a truncated quadric (matrix Q describes an infinite cylinder), the following condition should hold true:

$$X(t)^T Q\pi X(t) \geq 0 \tag{10}$$

so that the ray intersects the quadric within its boundaries. If a ray intersects *earlier* another body part, the corresponding point is considered occluded, whereas, if the ray-quadric intersection happens within another rendered quadric, the current body part is considered as collided. The percentage of per body-part occluded and collided points is used to further constrain the hypotheses set, by eliminating parts with high occlusion and collision ratios, respectively.

For each rendered hypothesis, the corresponding depth map is generated and compared against the observed one. The discrepancy between the hypothetical and observed depth maps is used as the metric to evaluate the particle. A predefined number of particles with the best score (i.e. minimum depth discrepancy) are propagated to the next frame where the sampling procedure is repeated. By tracking multiple hypotheses simultaneously we achieve a twofold result, that is (a) cope with noise and loss of data and (b) deal effectively with inter- and intra-person occlusions.

III. RESULTS

In order to assess the performance of the proposed methodology, we conducted a series of experiments of varying difficulty level. The experiments took place in various areas of an indoor environment and involved single or multiple persons, moving, acting and interacting freely in the scene. For evaluation purposes the experiments were organized in two categories. The first category concerns interaction scenarios with one or two users and the employment of marker-based ground truth data for obtaining quantitative results. The second category refers to experiments for providing illustrative results in test cases involving up to three users.

Quantitative results: Six interaction sequences, containing a total of more than 3000 frames, have been acquired and processed in order to quantitatively assess our methodology. As explained above, marker-based ground truth data have been used for the task at hand. In three out of the six sequences a single user is present, while in the other three, two users are present to facilitate occlusions among them. In the experiments with two users, one person assumed a "dummy role", that is to partially occlude the second person. The latter was the subject that was tracked and further utilized to obtain quantitative assessment figures.

In order to obtain ground truth data, prominent color markers have been attached on both sides of a user, and a two-camera setup has been employed to cope with cases of severe occlusions. More specifically, the setup consisted of two Kinect™ sensors, one facing the interacting users

and the other being placed behind them. Interference among the two sensors is avoided by having the second (rear) camera at a certain height, overlooking the scene. With this configuration, either the front- or the back-side markers are visible at all times. Example snapshots with one or two users are illustrated in Fig. 7, where both (frontal and rear) views are shown, together with the estimated 3D upper body pose. As explained above, in the rightmost example the tracking result illustrates only the occluded user.

For the actual quantitative assessment, each sequence has been processed by our methodology and the NiTE™ skeletonization module. The pose estimation, by means of joint-angles, provided by each method has been compared against the derived ground truth. Table II provides the obtained statistics, namely the mean angular error (averaged difference between the actual and the estimated angle) μE and its standard deviation σE. In order to better highlight the performance of the proposed methodology in the presence of occlusions, results in Table II are divided in two parts: (a) overall results for the six sequences, and (b) results for specific parts of the sequences, where inter-person occlusions are present. The latter have been manually detected and serve as the means to illustrate the effectiveness of *UTV*-based pose estimation in such cases. As can be observed, overall our methodology compares favourably to NiTE™ both in μE and σE. More importantly, results in cases of occluded interacting users demonstrate accurate performance of the proposed methodology and its superiority over NiTE™.

TABLE II

COMPARATIVE ASSESSMENT RESULTS. μE=MEAN ANGULAR ERROR THROUGHOUT THE SEQUENCE, σE= STANDARD DEVIATION OF ERROR.

	Overall		Under Occlusions	
	μE	σE	μE	σE
Proposed Methodology	**7.40** °	**3.70** °	**11.39** °	**6.48** °
NiTE™	8.03 °	4.20 °	14.14 °	7.57 °

Qualitative results: Additionally to the experiments that resulted in quantitative and comparative results, we extensively evaluated our methodology in numerous scenarios of varying difficulty. Throughout these tests, up to three users were involved, and in many cases user-occlusions were present. Illustrative instances from the named experiments are presented in Fig. 8 (single user case) and Fig. 9 (multiple users, featuring user-occlusions). In addition to results obtained by our methodology (middle image of each snapshot in blue background), Fig. 9 illustrates the performance of the NiTE's middleware (rightmost image in gray background) in the presence of occlusions. As can be observed, the proposed methodology succeeded in effectively tracking the pose of the upper body in all cases. On the contrary, the performance of NiTE™ deteriorates in cases of occluded users, either by providing erroneous estimations, as in the case of the left instance, or by completely losing track of occluded users, such as the middle user in the case of the right instance.

IV. DISCUSSION

In this paper we presented a robust, model-based methodology for human-torso 3D pose extraction. An explicit initialization/calibration phase is avoided, since it is not an option in complex, real life HRI scenarios. Moreover, we achieve pose-recovery in realistic interaction scenarios, even in the presence of severe occlusions. The latter is our major advantage compared to the state-of-the art, and is also supported by the obtained quantitative and qualitative results. This has been achieved by the introduction and formulation of *UTV*, which constitutes a main contribution of our work.

Our planned future work involves an immediate and a long-term goal. The former regards the extension of our methodology to cope with full-body pose recovery, including legs. This is a plausible scenario given that the employed user-modeling is directly amenable to extensions and/or additions. The latter addresses the study of more complex and involved interactions among users, a case that challenges most contemporary approaches to pose-recovery.

REFERENCES

[1] M. Pateraki, H. Baltzakis, and P. Trahanias, "Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation," *CVIU*, vol. 120, pp. 1–13, 2014.

[2] A. F. Foka and P. E. Trahanias, "Probabilistic autonomous robot navigation in dynamic environments with human motion prediction," *I.J. of Social Robotics*, vol. 2, no. 1, pp. 79–94, 2010.

[3] Microsoft kinect for xbox 360. Redmond WA. [Online]. Available: http://www.xbox.com/en-US/kinect

[4] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 755–762.

[5] 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.

[6] OpenNI. [Online]. Available: http://www.openni.org

[7] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, 2013.

[8] M. Sigalas, H. Baltzakis, and P. Trahanias, "Gesture recognition based on arm tracking for human-robot interaction," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 5424–5429.

[9] T. Asfour and R. Dillmann, "Human-like motion of a humanoid robot arm based on a closed-form solution of the inverse kinematics problem," in *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, vol. 2. IEEE, 2003, pp. 1407–1412.

[10] NiTE. [Online]. Available: http://www.openni.org/files/nite

[11] 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*, vol. 104, no. 2, pp. 90–126, 2006.

[12] R. Poppe, "Vision-based human motion analysis: An overview," *CVIU*, vol. 108, no. 1, pp. 4–18, 2007.

[13] S. Escalera, "Human behavior analysis from depth maps," in *Articulated Motion and Deformable Objects*, ser. Lecture Notes in Computer Science, F. Perales, R. Fisher, and T. Moeslund, Eds. Springer Berlin / Heidelberg, 2012, vol. 7378, pp. 282–292.

[14] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *2011 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 415–422.

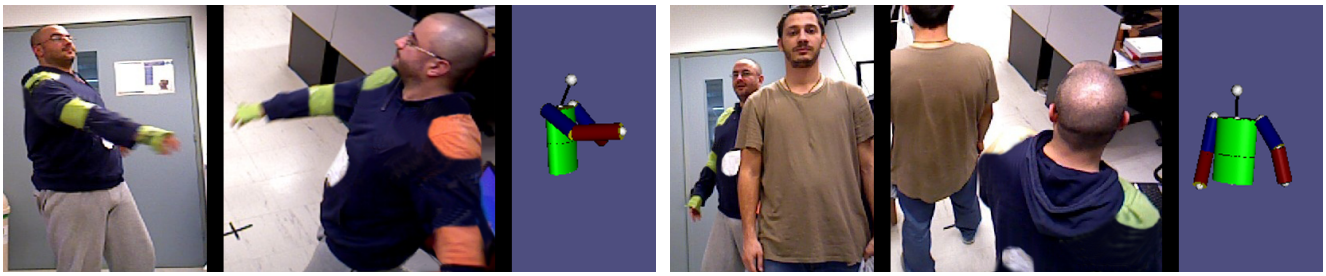


Fig. 7. Two snapshots from the experiments conducted to obtain quantitative results. Each snapshot shows three images: left - image captured from the front camera; middle - image captured from the rear camera; right - extracted 3D pose.

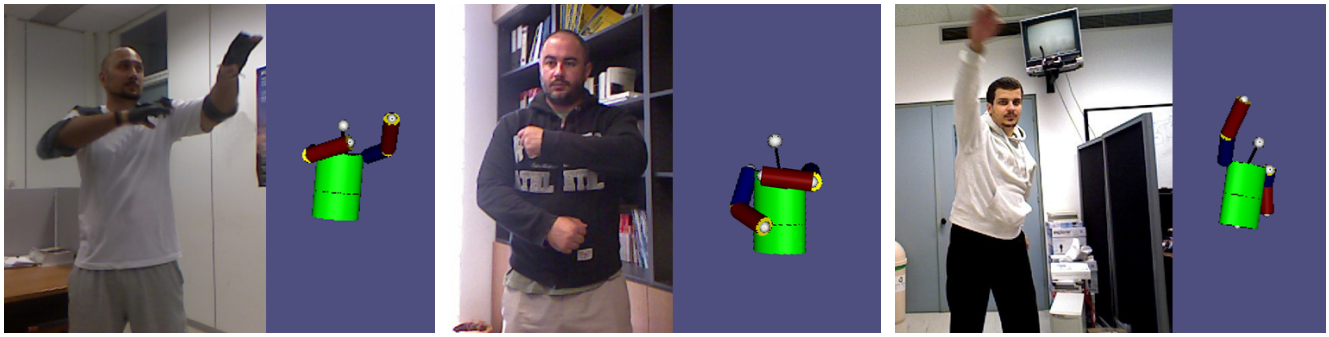


Fig. 8. Single user in the scene performing various poses containing self-occlusions.

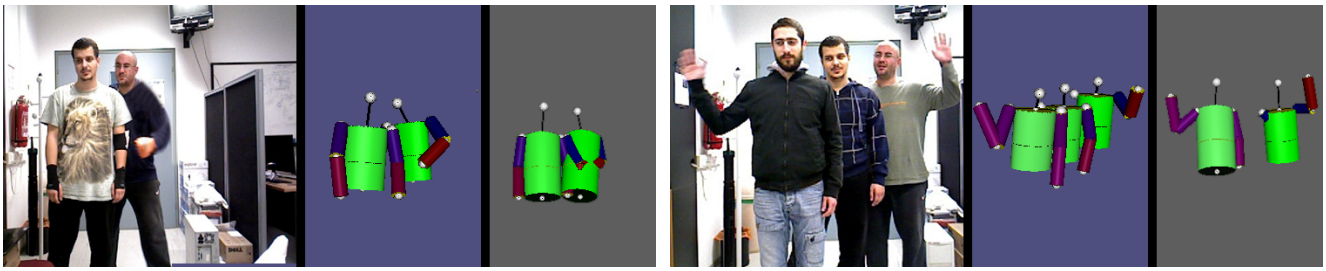


Fig. 9. Body pose tracking for multiple users. The performance of our methodology (3D pose in blue background) remains unaffected in cases of occluded users. On the contrary, the performance of NiTE^{TM} (in gray background) severely deteriorates.

[15] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 103–110.

[16] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3d pose estimation from a single depth image," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 731–738.

[17] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, and Z. Tu, "Exemplar-based human action pose correction and tagging," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1784–1791.

[18] A. Hernandez-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, and S. Escalera, "Graph cuts optimization for multi-limb human segmentation in depth maps," in *Proc. CVPR*, 2012.

[19] 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*. Springer, 2013, pp. 71–98.

[20] 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)*, 2011.

[21] B. Holt, E.-J. Ong, H. Cooper, and R. Bowden, "Putting the pieces together: Connected poselets for human pose estimation." in *Proc. International Conference on Computer Vision*, 2011.

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

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

[24] H. P. Jain, A. Subramanian, S. Das, and A. Mittal, "Real-time upper-body human pose estimation using a depth camera," in *Computer Vision/Computer Graphics Collaboration Techniques*. Springer, 2011.

[25] E. Kalogerakis, A. Hertzmann, and K. Singh, "Learning 3d mesh segmentation and labeling," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, p. 102, 2010.

[26] 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*. IEEE, 2010, pp. 3108–3113.

[27] E. Churchill, J. McConville, L. Laubach, P. Erskine, K. Downing, and T. Churchill, *Anthropometric Source Book. Volume II: Handbook of Anthropometric Data*, 1978.

[28] B. L. Webber, C. B. Phillips, and N. I. Badler, "Simulating humans: Computer graphics, animation, and control," *Center for Human Modeling and Simulation*, p. 68, 1993.

[29] H. Baltzakis, M. Pateraki, and P. Trahanias, "Visual tracking of hands, faces and facial features of multiple persons," *Machine Vision and Applications*, 2012, 10.1007/s00138-012-0409-5. [Online]. Available: http://dx.doi.org/10.1007/s00138-012-0409-5

[30] B. Stenger, A. Thayananthan, P. H. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter," *IEEE Transactions on PAMI*, vol. 28, no. 9, pp. 1372–1384, 2006.