Full-Body Pose Tracking—The Top View Reprojection Approach

Markos Sigalas, Maria Pateraki, and Panos Trahanias

Abstract—Recent introduction of low-cost depth cameras triggered a number of interesting works, pushing forward the state-of-the-art in human body pose extraction and tracking. However, despite the remarkable progress, many of the contemporary methods cope inadequately with complex scenarios, involving multiple interacting users, under the presence of severe inter- and intra-occlusions. In this work, we present a model-based approach for markerless articulated full body pose extraction and tracking in RGB-D sequences. A cylinder-based model is employed to represent the human body. For each body part a set of hypotheses is generated and tracked over time by a Particle Filter. To evaluate each hypothesis, we employ a novel metric that considers the *reprojected Top View* of the corresponding body part. The latter, in conjunction with depth information, effectively copes with difficult and ambiguous cases, such as severe occlusions. For evaluation purposes, we conducted several series of experiments using data from a public human action database, as well as own-collected data involving varying number of interacting users. The performance of the proposed method has been further compared against that of the Microsoft's Kinect SDK and NiTE™ using ground truth information. The results obtained attest for the effectiveness of our approach.

Index Terms—Human pose estimation, model-based, tracking, particle filtering

1 Introduction

Human pose estimation and tracking is an important and challenging topic, encountered in a wide range of applications, including human-robot interaction, social robots, gaming, virtual reality and medical applications. The emergence of low cost real-time depth cameras, such as the Kinect sensor [1], led to numerous important approaches to the pose extraction and tracking problems, significantly pushing forward the state of the art (e.g., [2], [3], [4]). Nevertheless, despite rather accurate performance in controlled or semi-controlled scenarios, coping with highly complex cases involving multiple users remains a challenging problem. The difference in the visual appearance and body shape of humans, the variability in lighting and environment background clutter as well as occlusions are a few reasons which set the problem as non-trivial. Also, the inherent requirement, presented in a number of pose tracking methodologies, for an explicit or implicit initialization phase, influences performance in realistic scenarios, where initialization is usually not possible. Moreover, in such cases, the fact that multiple users arbitrarily move and interact with each other, results to the frequent appearance of intra- and inter-person occlusions, namely occlusions across body parts of the same user or across different users, respectively. This, along with the anthropometric and kinematic inconsistencies encountered in many of the non-model-based approaches, results to erroneous pose extraction and, thus, may severely deteriorate performance.

The objective of our work is to achieve accurate body pose estimation and tracking in realistic complex scenarios in the presence of severe occlusions. The main contribution of our approach is the introduction of the *Top View Reprojection* (TVR) concept in a model-based framework to uniformly treat the body parts representing the human body. The employed body model consists of an ordered structure of 11 body parts, represented as cylinders. An appropriate metric, formulated as a scoring function, is associated with each assumed pose, and pose estimation for each body part is derived through function minimization.

The current work builds on and significantly extends the TVR formulation for human torso pose estimation, initially presented in [5]. The estimation of the full-body pose entails greater complexity due to a number of reasons. Segmentation is more involved when dealing with several body parts of different sizes and occlusions may also affect the overall performance, by creating local minima or even shifting the global minimum. Moreover, collisions across different users or body parts of a user should also be taken into consideration. Consequently, in this work the formulation of the scoring function has incorporated the above aspects in order to effectively cope with different body parts and challenging cases of severe occlusions and frequent interactions among users. Accordingly, the TVR framework facilitates a generic treatment to the task of estimating the pose of the articulated human body. In that sense, all body parts are uniformly processed, eliminating the need of intermediate body part descriptors. Moreover, it allows for high tolerance on human body diversity, caused by the users' gender, anthropometric characteristics or posture, and most importantly for credibly treating occlusions within the TVR scoring function.

For evaluation purposes, we conducted an extensive series of experiments using (a) a public benchmark involving

• *The authors are with the Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH) and the Department of Computer Science, University of Crete, Heraklion, Crete, Greece. E-mail: {msigalas, pateraki, trahania}@ics.forth.gr.*

Manuscript received 1 Mar. 2015; revised 31 Oct. 2015; accepted 9 Nov. 2015. Date of publication 0 . 0000; date of current version 0 . 0000. Recommended for acceptance by S. Escalera, J. Gonzàlex, X. Baró, and J. Shotton. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2015.2502582

individual users performing different actions and (b) own-captured data involving varying number of users acting and interacting arbitrarily with each other. Ground truth information has been employed to compare our method against the skeletonization tool of Microsoft's Kinect SDK [3] and NiTE [6]. Both illustrative and comparative results attest for the effectiveness and accurate performance of our methodology. In particular, it has been experimentally verified that the proposed, TVR-based approach copes commendably with cases of severe occlusions, either across body parts of the same user, or across different users. This constitutes a major advantage of the proposed approach compared to the state-of-the art and verifies its suitability in human pose estimation in realistic scenarios.

2 Related Work

Markerless body pose estimation has been an active research area in computer vision for decades. Accordingly, numerous image-based methods have been developed and interested readers are referred to Moeslund et al. [7], [8] and Poppe [9] for comprehensive literature surveys. However, among various challenges, the issue of occlusions is very difficult to deal with image-based methods. Recently, the release of the Microsoft Kinect sensor [1] has led the growing trend to use depth information, which provides a robust cue in dynamic scenes, surpassing the inherent ambiguity in image data. An overview of state-of-the-art body pose estimation and tracking approaches using depth information can be found in [10], [11], [12]. The learning-based approach of Shotton et al. [3], implemented in the Microsoft Kinect SDK [13], is probably the most widely used method for body pose extraction in RGB-D sequences. A large corpus of synthetic data, containing various realistic body configurations, is employed to train a Random Forest classifier, in order to assign each pixel to the corresponding body part. Extensions of this work [4], [14], [15] managed to provide more accurate solutions and alleviate some inconsistencies imposed by the absence of a kinematic/anthropometric model. The Kinect SDK method [3] has been mainly tuned for a large living room with a stationary narrow field-of-view camera, and limited user occlusions and operates satisfactorily for its intended use. Accordingly, the scalability of the method in more demanding environments, with different clutter, occlusions and interactions among multiple humans is questionable, as well as how much data would be required for such generalization.

When dealing with occlusions, a number of relevant depth-based methods have recently appeared. Ye et al. [16] match the input depth map with a set of pre-captured motion exemplars to generate a body configuration estimation. They further rely on the coherent drift point (CDP) algorithm to solve non-rigid point registration and provide accurate pose estimation under the presence of occlusions. Yet, the authors state that failures exist in cases when there is no similar pose in the database for which a reasonable point correspondence can be estimated using CDP. Similarly, in [17], an exemplar-based method is used to learn an inhomogeneous systematic bias for body pose correction and tagging. Graph Cuts [18] have showed good performance in segmenting different parts of single users. Pictorial Structure Models [19], as well as hybrid approaches combining local pose optimization and global retrieval techniques, such as the one presented in [20], have also been successfully utilized to infer the body pose, addressing the issue of self-occlusions. Inspired by Pictorial Structures, Kiefel and Gehler [21], proposed the Field of Parts formulation which also models the absence of the human body, in an attempt to cope with possible occlusions. However, failure cases in most of them are usually cases of extreme occlusions [19], [20] or occlusions between different users [21].

Local optimization for the estimation of body parts is frequently employed for body pose estimation. In [18], pixels are classified into seven body parts, while, Siddiqui and Medioni [22] build different detectors for the head, forearm and hand. Similarly, Zhu and Fujimura [23] build heuristic detectors to locate upper body parts (head, torso, arms), yet there are difficulties in recovering the arms from tracking failures. Zhu and Fujimura [24] use a part-based approach with bayesian inference to effectively address self-occlusions and recover from tracking failures. Kalogerakis et al. [25] use a data-driven Conditional Random Field (CRF) model to segment and label eight body parts from the 3D mesh. 3D mesh is also utilized in [26] and [27] along with geodesic maps, in order to detect the head, hands and feet. A combination of CRF and random forests has been proposed by Nowozin et al. [28], namely the Decision Tree Fields, in order to alleviate the inherent constraints of both techniques and end up with proper labeling. Similarly, Ramakrishna et al. [29] propose the Inference Machine framework, to recover the articulated body pose, which models the strong interaction among body parts and is able to cope with the high multi-modal appearance of each part.

Evidently, a large body of research deals with the problem of articulated body pose extraction and tracking. Despite the fact that most of the contemporary approaches perform well in usual cases, when dealing with complex, realistic interaction scenarios, limiting factors appear, affecting the overall effectiveness. One such drawback is the inherent requirement for an initialization period, either explicitly, demanding a specific predefined pose [23], [30], or implicitly, by registering and tracking the user over a time-window [3]. Additionally, the absence of a kinematic/anthropometric model leads to kinematic inconsistencies in the provided poses. Even more importantly, a serious limitation of most of the state-of-the-art approaches, is the poor capability 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.

3 Methodology Overview

In the current work we propose a full body pose estimation methodology, using input from RGB-D image sequences, which uniformly treats the body parts representing the human body. The overall method builds on the concept of *Top View Re-projection (TVR)*, initially presented in [5]. TVR is based on the observation that the visible area of any object varies according to the point of view of the camera and in

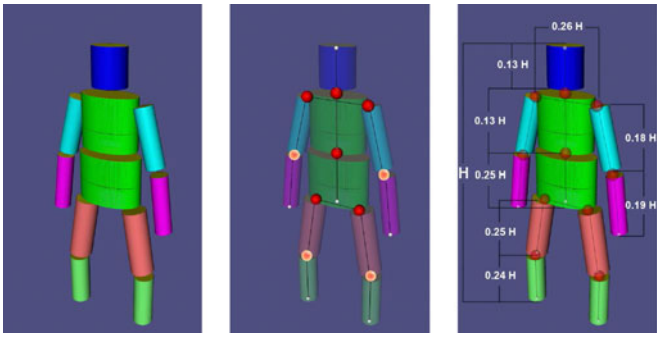

Fig. 1. (a) 11-part body model: head, upper and lower torso, upper and forearms and upper and lower legs. Body parts are modeled either as elliptical (upper and lower torso) or circular (rest 9 parts) cylinders. (b) Joints of the human body: Red spheres represent *spherical* joints while yellow spheres represent *hinge* joints. (c) Anthropometric measures of an adult human, expressed proportionally to the human's height *H*.

the case of cylindrical objects becomes minimum when the view axis of the camera is aligned with the major axis of the object, resulting in a unique object pose. We exploit the TVR minimum property for cylindrical objects in the pose estimation domain, by following a model-based approach, assuming the human body to comprise of cylindrical segments, corresponding to the different body parts, as well as spherical joints. The cylindrical modeling of the human body in conjunction with the TVR property, facilitate the formulation of a reliable scoring function, able to render the pose even in cases of occluded cylindrical objects (body parts). The employed scoring function is uniformly defined, in the sense that it is not dependent on the specific body part, offering a robust general-purpose metric for the estimation of the underlying pose.

For each body part a number of pose hypotheses is sampled from its configuration space, respecting the underlying kinematic model. Each of the pose hypothesis is evaluated against the TVR scoring function and the hypothesis with the best score yields for the assumed pose and the location of the connected joints. The full body pose is consequently extracted from the estimated poses of the individual body parts, processed in a top-to-bottom order and enforcing a local optimization schema.

In the following two sections, the various steps of the proposed approach are described in detail. The employed body model and the TVR-based scoring function, constituting a main contribution of the current work, are presented in the next Section 4. Subsequently, the overall framework for body pose recovery, integrating the aforementioned scoring function and the sampling/update of pose hypotheses of each body part, is discussed in Section 5.

4 Human Model and Scoring Function

4.1 Body Modeling

The human body is represented by the part-based, kinematic structure of Fig. 1, consisting of 11 body parts; namely the head, the upper and lower torso, the left and right upper and forearms and the left and right upper and lower legs (Fig. 1a). Each part is modeled as a cylinder, either elliptical, in the case of upper and lower torso, or circular for the rest of the body parts. These are connected to each other with 10 joints in total, labelled as (Fig. 1b): *spherical* joints, shown as red spheres, and *hinge* joints, shown as yellow spheres. Joints of the first category have three Degrees of Freedom (DoF) referring to the joints of the neck, the torso, the two shoulders and the two hips. The rest of joints, namely the two elbows and the two knees, belong to the second category and have one DoF. Thus, a 22 DoF kinematic model is established, consisting of 11 cylindrical parts, similar to the one presented in [31].

TABLE 1
Joint Angular Limits of the Employed Model

Joint	x-angle (deg)	y-angle (deg)	z-angle (deg)
Neck	$[-30°, 60°]$	–	$[-45°, 45°]$
Torso	$[-30°, 45°]$	$[-30°, 30°]$	$[-30°, 30°]$
L. shoulder	$[-60°, 180°]$	$[-90°, 90°]$	$[-10°, 180°]$
R. shoulder	$[-60°, 180°]$	$[-90°, 90°]$	$[-180°, 10°]$
L. hip	$[-45°, 100°]$	$[-90°, 90°]$	$[-10°, 90°]$
R. hip	$[-45°, 100°]$	$[-90°, 90°]$	$[-90°, 10°]$
Elbows	$[-10°, 160°]$	–	–
Knees	$[-100°, 10°]$	–	–

The employed model is used to constrain the workspace of each body part via linear and angular limits and distinguish between valid and invalid poses. Table 1 presents the angular limits used in our implementation, which are experimentally established [32]. To dynamically infer the size of each body part, we also employ established anthropometric measurements for the human body [32], [33]. Assuming an adult human, individual body part sizes can be expressed proportionally to the user's height H, as illustrated in Fig. 1c. The anthropometric model of Fig. 1c, has been experimentally established and is able to handle proportional variations among users -of same or different gender- without affecting the overall performance.

4.2 Top View Re-Projection—TVR

The idea behind the *Top View Re-projection* stems from the natural observation that the visible area of any object, i.e., its projection onto the image plane, varies according to the point of view of the camera and becomes minimum at certain views, depending on the object's shape. This is quantitatively formulated by introducing the *reprojection ratio* f_{reproj} of an object, namely the ratio of the number of reprojected (visible) points N_{Pr} to the total number of 3D points of the object N_{3D}

$$f_{reproj} = \frac{N_{Pr}}{N_{3D}}. \quad (1)$$

Interestingly, in the case of cylindrical objects, the reprojection ratio has two minima, the top and bottom view, when the view axis of the camera is aligned to the major axis of the cylinder, resulting to a unique object pose. This property is illustrated in Fig. 2. Given a simulated 3D point cloud representing a cylinder of known size (either circular or elliptical) we generate multiple views, by rotating the cylinder around the x-axis, with and without occlusion (bottom row and top row, respectively). For each view of the cylinder, the corresponding ratio of the reprojected points to the total number of 3D points in the point cloud is computed and is

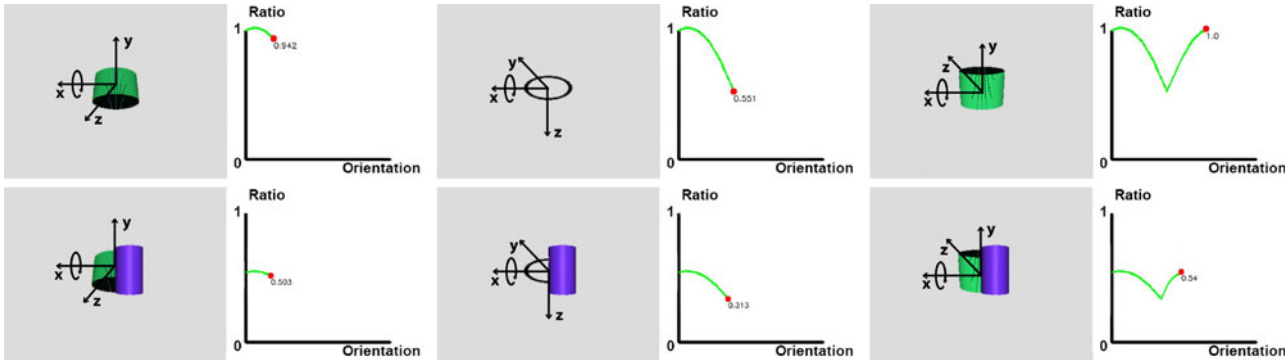

Fig. 2. Different views of a simulated 3D point cloud of an elliptic cylinder which rotates around the *x*-axis without occlusions (top row) and while occluded by another object (bottom row) and the respective plot of the re-projection ratio. In both cases, the re-projection ratio becomes minimum at the cylinder's top and bottom view, as shown in the middle column.

shown in the respective 2D graph of each image. As can be observed, f_{reproj} varies according to the view and becomes minimum at the cylinder's top and bottom view, as shown in the middle column of Fig. 2.

This property of cylindrical objects can therefore be exploited to derive the pose of a human body part, assuming the latter is modeled by a cylinder. Since both the top and bottom view give rise to a unique object pose, we restrict to the estimation of the cylinder's *Top View* to derive the pose of a human body part. With this in mind, we utilized the TVR concept to estimate the torso pose in our previous work [5]. Given the organized 3D point cloud of the user, captured by a single RGB-D camera, we generate multiple virtual (hypothetical) views of it, as if the camera is moving along the surface of a semi-sphere around the user according to Fig. 3. Each virtual view is a hypothesis for the Top View of the observed torso (or the cylinder representing the torso) and, consequently, a hypothesis for the user's torso configuration as in Fig. 4. Each hypothesis is scored based on the corresponding reprojection ratio (f_{reproj}) and the hypothesis with the best score is denoted as the object's *Top View*. Thus, the actual torso pose is directly extracted as a by-product of the estimated Top View.

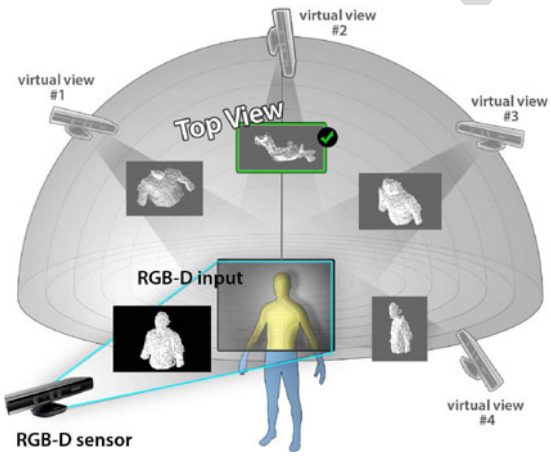

Fig. 3. Illustrative generated virtual views of the observed user. Reprojection ratio is minimized at the *Top View*, denoted with the green rectangle. The small images depict the reprojected points from the corresponding virtual view.

4.3 TVR Scoring Function

As explained above, in the case of simulated data (Fig. 2), f_{reproj} has a single, well defined, minimum at the object's Top View. Consequently, the torso pose can be estimated based only on the reprojection ratio [5], since the respective body part occupies a significant portion of the user's area and can be easily segmented. Due to the fact that the TVR approach is mainly based on the minimum reprojection ratio of cylindrical objects, the estimation of the full body pose is greatly facilitated by the employed cylinder-based model (Fig. 1a). However, the estimation of the full-body pose entails greater complexity due to the following reasons. First, we seek the best out of a sampled set of hypothesized 3D cylinders, namely the one that is best aligned to the point cloud of the segmented body part and for which the reprojection ratio becomes minimum. In this aspect, segmentation is less trivial when dealing with several body parts of different sizes. Occlusions may also affect the overall performance, namely alter the ratio's behavior, by creating several local minima, valleys and plateaux or even by shifting the global minimum. Moreover, collisions across different body parts or different users should also be taken into consideration, thus penalizing or exempting collided hypotheses. Consequently, the formulation of a generic scoring function, aiming to select the best hypothesis, must incorporate the above aspects in order to effectively cope with different body parts and challenging cases of severe occlusions and frequent interactions among users. Towards this goal the scoring function f_{TVR} is formulated as:

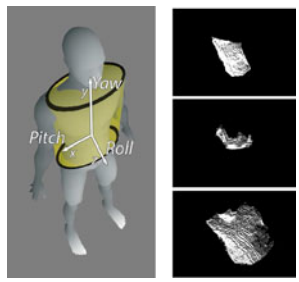

Fig. 4. (a) Elliptic cylinder model of the human torso. (b) Reprojections of the point cloud from three indicative hypothetical views.

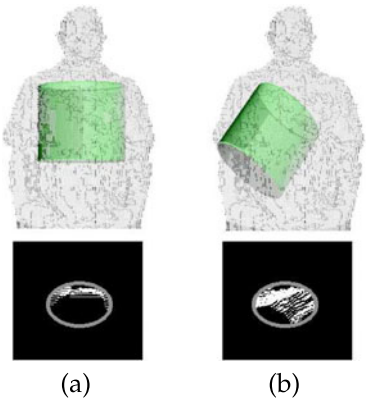

Fig. 5. Computation of f_{reproj} for two different upper torso pose hypotheses (top) and their corresponding top-views with the reprojected points (bottom). (a) indicates the hypothesis with the minimum score and (b) an erroneous hypothesis.

$$f_{TVR} = f_{reproj} \times f_{align} \times f_{discr}, \tag{2}$$

where f_{reproj} is the reprojection ratio, f_{align} represents an alignment term that penalizes misalignments between hypotheses and observation data, and f_{discr} represents a discrepancy term that compensates for erroneous minima caused by occlusions and/or collisions across body parts.

4.3.1 Reprojection Ratio f_{reproj}

Let a hypothesized cylinder representing a specific body part, and consider also the segmented 3D point cloud of an observed human body (Section 5.1). The reprojection ratio f_{reproj} is computed as the ratio of the number of reprojected points N_{Pr} lying inside the hypothesized cylinder to the total number of 3D points lying inside the hypothesized cylinder N_{3D}, as illustrated in Fig. 5. In this way each hypothesis controls the segmentation of the respective body part via the minimum reprojection ratio property and the one with the best (minimum) score yields for the best pose. Essentially, segmentation proceeds by assigning to the body part under consideration those 3D points that lie in the interior of the current hypothesized cylinder, and the selected 3D points are then excluded from the processing of subsequent body parts.

4.3.2 Alignment Term f_{align}

The alignment term, as its name implies, is responsible for the alignment between a hypothesis and actual observation. While the *reprojection ratio* remains a strong indicator for the Top View and respective body part pose, there are cases where f_{reproj} favors certain pose hypotheses due to the small number of 3D points lying inside the hypothesized 3D cylinder, although they do not align to the observation. f_{align} is, therefore, introduced to penalize those hypotheses, and is computed as:

$$f_{align} = \frac{N_{cyl}}{N_{3D}}, \tag{3}$$

where N_{cyl} is the number of visible 3D points of each hypothesized 3D cylinder (computed during the hypothesis rendering step, described in Section 4.3.4), and N_{3D} is the number of 3D points that intersect with the hypothesized cylinder.

4.3.3 Discrepancy Term f_{discr}

The term f_{discr} integrates aspects of occlusions and collisions, aiming at:

a) Compensating for minima caused by occlusions, by favoring occluded hypotheses which would have better score if they were not occluded.
b) Favoring hypotheses with high overlapping areas between the segmented point cloud and the 3D cylinder.
c) Penalizing invalid hypotheses which collide with other body parts.

f_{discr} is consequently formulated as:

$$f_{discr} = f_{occl} \times f_{coll} \times f_{ovl}, \tag{4}$$

where f_{occl} is the occlusion factor, f_{coll} is the collision factor and f_{ovl} is the overlap factor. Their computation is described below (Section 4.3.5) in detail. As it assumes the rendering of cylinder hypotheses, we first present in the following paragraph the employed hypothesis rendering technique.

4.3.4 Hypothesis Rendering

To render the body parts hypotheses we use the ray tracing technique and the projective geometry of quadrics, described thoroughly in [5], [34]. This technique is perfectly suitable for the task at hand, due to the employed cylindrical modeling of body parts. Elliptical cylinders, in the direction of the *y*-axis, can be represented in homogeneous coordinates as a symmetric 4×4 matrix Q:

$$Q = \begin{bmatrix} \frac{1}{a^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{b^2} & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \tag{5}$$

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

$$X^T Q X = 0. \tag{6}$$

In order to render a hypothesized part, 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 eq. (6):

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

and solving for t. In case the ray intersects the quadric, eq. (7) has two solutions, while in case the ray is tangent to the quadric eq. (7) has a unique solution. Apparently, if the ray and the quadric do not intersect, eq. (7) does not have a real solution. In case of a truncated quadric (matrix Q describes an infinite cylinder), the following condition should also hold true:

$$X(t)^T Q \pi X(t) \geq 0 \tag{8}$$

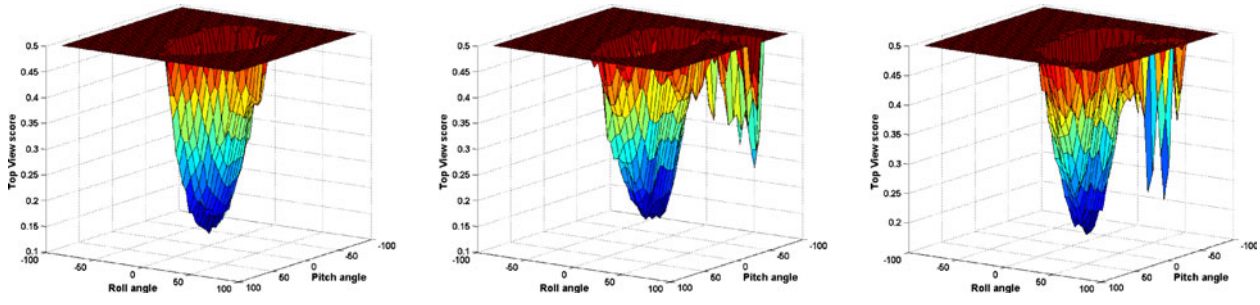

Fig. 6. Indicative graphs of f_{TVR} score under different percentages of occlusion compared to ground truth data: (a) no occlusion, (b) 30 percent and (c) 60 percent occlusion. The two *bottom* axes are the pitch and roll angles, namely the hypothesis configuration, and the vertical axis is the corresponding f_{TVR}. Pitch and roll angles are given as angular distance from the ground truth, which lies at (0,0), the center of the bottom grid. While additional local minima appear, f_{TVR} presents a strong, well defined, global minimum, at -or close to- the actual pose configuration.

so that the ray intersects the quadric within its boundaries. Q_π stands for the quadric representation of a pair of planes parallel to the xz-plane [34]. Ray tracing, as described above, is used to compute the number of visible 3D points, N_{cyl}, of each hypothesized cylinder as well as the number of point collisions N_{coll} by assuming depth overlapping across various hypothesized rendered body parts.

4.3.5 Pixel Characterization and Computation of Scoring Factors

The 3D rendering generates the depth map of the hypothetical 3D cylinder D_{cyl} which, in turn, is compared against the depth map of the segmented 3D points D_{3D} of the body part hypothesis and the original acquired depth map D_{all}. The later, D_{all}, is used to detect potential occluded pixels. Depth maps are superimposed and pixels are characterized as:

- *Inlier pixels* P_{in}. Pixels with valid depth in D_{cyl} and in D_{3D}, for which their absolute difference is below a certain threshold T, experimentally estimated.
- *Outlier pixels* P_{out}. Pixels with valid depth in D_{cyl} but not in D_{3D} and in D_{all}.
- *Occluded pixels* P_{occl}. Pixels with valid depth in D_{cyl} and in D_{all}, for which their difference is above the threshold T.

The normalization factor f_{ovl} is taken as the ratio of the number of outlier pixels to that of inliers. f_{coll} is the percentage of collided pixels -calculated in the rendering process as N_{coll}- and is used to penalize invalid hypotheses which collide with other body parts. Finally, the percentage of the estimated number of occluded pixels with respect to the number of 3D points of the hypothesized cylinder is used to calculate the occlusion factor f_{occl}. Accordingly, f_{occl}, f_{coll}, f_{ovl} are computed as:

$$\begin{aligned} f_{occl} &= 2 - \frac{P_{occl}}{N_{cyl}} \\ f_{coll} &= 1 + \frac{N_{coll}}{N_{cyl}} \\ f_{ovl} &= 1 + \frac{P_{out}}{P_{in}}, \end{aligned} \tag{9}$$

where N_{cyl} is the total number of 3D points of the hypothesized cylinder. Notice that f_{coll} and f_{ovl} are not allowed to assume zero values by the addition of the unity term, and f_{occl} assumes the $(2 - \frac{P_{occl}}{N_{cyl}})$ formula in order to favor larger values of the second term.

The derived formulation of the TVR score guarantees high detection rates and accuracy, even in cases of severe occlusions. This is illustrated in Fig. 6, where the score of several torso pose configurations, namely different pitch and roll angles, is compared against ground truth data for different occlusion scenarios, namely 0, 30 and 60 percent of occlusion. *Bottom* axes represent the pitch and roll angles and the vertical axis is the corresponding TVR score for each pitch/roll configuration. More precisely, the pitch and roll axes indicate the angular distance from the ground truth which lies at the center of the bottom grid, namely at (0,0). As observed, f_{TVR} maintains the global minimum property and hence renders the actual pose, even in the cases of occluded torso where the function's behavior is altered with the appearance of local minima.

5 Full Body Pose Estimation

Having formulated the TVR score function, we proceed in the current section with the formulation of the overall body pose recovery methodology. The two main steps of the latter are depicted in Fig. 7, initiating from a human-body segmentation and ordering step to the main pose recovery and tracking part. The *Human-body segmentation* step is used to detect all humans in the scene and segment their corresponding points clouds. This step also encompasses the estimation of a set of anthropometric measures for each segmented human body, based on the height of each user. The *Pose recovery and tracking* step utilizes the *TVR scoring function* for each body part; the latter are examined in a top-to-bottom order, starting from the head to recover the full body pose. The aforementioned two steps are further elaborated in the rest of this section.

5.1 Human-Body Segmentation

In order to segment each human-body from the scene, we follow the technique depicted in Fig. 8 and developed in our previous works [35], [36], [37]. Initially skin-colored blobs are detected in the RGB image. The geometric and motion characteristics of the skin-colored blobs are used in order to classify them into faces (Fig. 8b), which trigger the user's detection. Given the 3D location of the face, we can

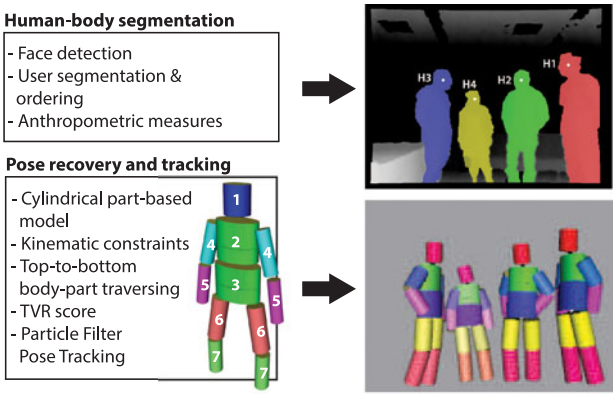

Fig. 7. Overview of the pose recovery methodology. The first step (top) concerns the human body segmentation and depth-based ordering and the second step (bottom) the pose recovery and tracking.

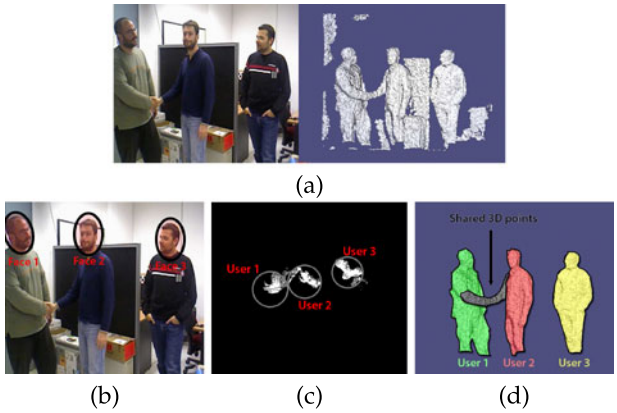

Fig. 8. (a) Indicative input from the RGB-D sensor. RGB image is depicted in the left part and the corresponding organized 3D point cloud of the depth channel is depicted in the right part. (b) Detected faces by skin-color classification on the RGB image. (c) Approximated reprojection of scene's overview, rotated by 90 degrees around the horizontal axis of the camera system. Users are ordered based on the detected face depth. (d) Point cloud of each human-body is segmented based on the scene overview. Grey area depicts the 3D points shared by users 1 and 2.

estimate the user's height and, thus, approximate the anthropometric parameters of the model. Additionally, users are ordered based on the face depth, starting from the one *closer* to the camera and moving *backwards*, and are processed accordingly.

The 3D face position steers the segmentation of the body point cloud, by employing a technique similar to the one presented in [38]. Assuming that the camera is roughly parallel to the floor, the 3D point cloud of the whole scene is rotated by 90 degrees around the horizontal axis of the camera system and reprojected on the image plane, providing an approximation of the scene's overview (Fig. 8c). Reprojected points within a specific radius from the face center projection are considered to belong to the current user. Finally, standard connected components labeling [39] is applied on the border points, in order to capture possibly extended limbs.

However, in a realistic scenario, where multiple users move, act and interact with each other freely, accurate segmentation is often impossible. This is because individual 3D point clouds of users may be adjacent or overlapping with each other, as in the handshake example of Fig. 8 where users 1 and 2 share the same subset of 3D points (Fig. 8d). In this type of cases of cluttered segmentation and collided hypotheses the *TVR* scoring function is able to retrieve robustly the body part hypothesis, exploiting also the underlying kinematic model in order to constrain valid hypotheses.

5.2 Pose Recovery and Tracking

Given the location of the face and the corresponding user's 3D point cloud, we employ a local optimization technique to infer the full body pose. Body parts are individually processed in a top-to-bottom sequential order. The pose of each body part indicates the location of the connecting joint which, in turn, is utilized for estimating the pose of the subsequent part. In other words, given the face location, we estimate the pose of the head which, in turn, is used to extract the position of the neck joint. The neck is used to recover the upper torso pose which yields for the location of the two shoulder joints and so forth. When the pose of a specific body part is recovered the respective segmented 3D points are subtracted from the user's point cloud and are not used in the subsequent processing of the remaining body parts.

The pose of each part is tracked over time by means of a Particle Filter (PF). Initially, multiple hypotheses are generated/sampled from the configuration space of the corresponding part, defined by the underlying kinematic model. Each of the hypotheses is evaluated against the *TVR* score, described in the previous section, and the hypothesis with the best score yields for the assumed pose and the location of the connecting joint(s). Finally, the set of *N-best* hypotheses are propagated to the next iteration and are used as seed for the next sampling step.

5.2.1 Hypothesis Sampling

At each frame, the location of the corresponding joint steers the generation of a set of body part configuration hypotheses, tracked and maintained over time by a PF. Based on the kinematic model presented in Section 4.1, each configuration particle (hypothesis) refers to the pitch, roll and yaw angles of the corresponding cylinder. Additionally, the underlying kinematic model is further utilized in order to eliminate invalid hypotheses.

The particles are generated and maintained using the Sampling/Importance Resampling (SIR) algorithm, initially introduced by Rubin et al. [40]. According to SIR, each particle m is assigned a weight w^m which is computed based on the Importance Sampling Principle as:

$$w_k^m = P(y_k|\theta_k^m)w_{k-1}^m, \tag{10}$$

which corresponds to the likelihood of the observations y for the particle m and the corresponding estimated configuration θ at time instant k.

5.2.2 Hypothesis Evaluation and Update

In our implementation, the likelihood $P(y_k|\theta_k)$ of each particle is determined by the TVR score of the corresponding hypothesis, as described in detail in Section 4.3. Clearly, the hypothesis which best satisfies the f_{TVR} yields for the pose of the corresponding body part and is used to calculate the

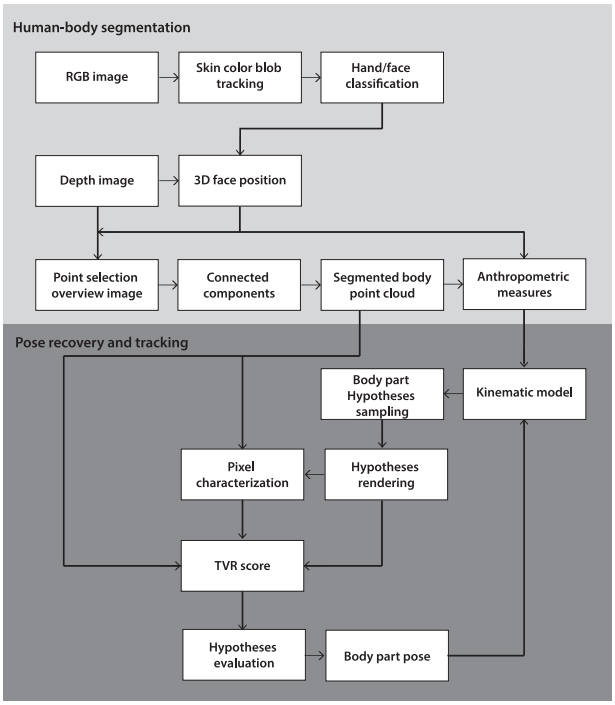

Fig. 9. Methodology pipeline and information flow between the subsystems.

location of the connecting joint(s). Additionally, to avoid degenerate situations in which large number of samples have weights close to zero, after a few iterations, SIR also includes a resampling step which ensures that unlikely samples are replaced with more likely ones. In that sense a predefined number of particles with the best score are propagated to the next frame where the sampling procedure is repeated.

In order to avoid excess computational costs, only the best hypothesis of a body part is utilized for deriving the location of the corresponding joint, and thus steering the pose estimation of the adjacent body part. This effectively keeps the number of hypotheses constant throughout the whole process.

5.3 System Block Diagram

A *Block Diagram* illustrating the overall methodology pipeline and information flow between the various system components, as described in the previous sections, is given in Fig. 9. Initially skin-color blobs, in the RGB image, are detected and classified into faces and/or hands. The detected skin-color faces are assigned to users and depth information is utilized to estimate each face's 3D location. A detected face is initially used to infer the user's body part sizes, based on anthropometric proportion measurements. Additionally, the derived face position steers the segmentation of each human-body from the scene following the technique described in Section 5.1.

Given the segmented 3D point cloud of the body, a local optimization technique is employed to recover the pose of each body part, in a top-to-bottom order starting from the head. The pose of each body part is tracked over time by means of a separate PF (Section 5.2). A set of hypotheses are generated at the beginning of each iteration, according to the constraints imposed by the employed kinematic model.

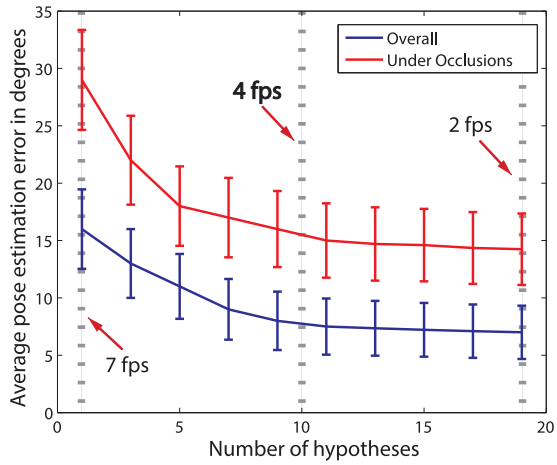

Fig. 10. Average error in body pose estimation as a function of the number of hypotheses tracked for each body part, along with the corresponding frame rate. Blue graph refers to the complete data set; red graph refers to the subset that included cases with significant occlusions.

Each hypothesis is then evaluated by the TVR scoring function, which finally leads to the selection of the hypothesis with the best TVR score as the resulting body part pose.

6 Results

The formulation of our TVR-based methodology, described in the previous sections, and more specifically the system presented in Fig. 9, has been implemented on a single core CPU system with 8 GB RAM. Prior the method's detailed assessment, we have performed different tests to set the value of specific parameters, namely the number of propagated hypotheses and the threshold T used in pixel characterization.

To assess the effect of tracking multiple hypotheses in terms of average error in pose estimation and computational complexity, we conducted a series of experiments where tracking was performed with a varying number of hypotheses. More specifically, the number of hypotheses that was tested was in the range of 1-19. In each case, the average pose error (in degrees) was derived for data with and without occlusions. The error graphs for the two cases are presented in Fig. 10, along with the corresponding computation times expressed as processed frames per second (fps). As can be observed, in both cases the error drops with the number of hypotheses, approximately assuming half of its value when the number of hypotheses becomes 10. It is also evident that increasing the number of hypotheses beyond 10 has a minimal effect on the pose error. Given that the error is notably larger in cases where occlusions are present, we may deduce that the effect of multi-hypothesis tracking is more significant in those cases. Based on the above results, in our implementation we set the number of hypotheses to 10; the latter facilitates operation with low error rates, and at the same time does not tax it computationally.

The threshold T (see Section 4.3.5) is used to provide flexibility on the characterization of pixels as inlier or occluded pixels. Intuitively, it represents the maximum allowable distance of a point to the hypothetical cylinder in order to be considered as inlier. To evaluate the impact of T, we conducted a series of experiments by altering its value and

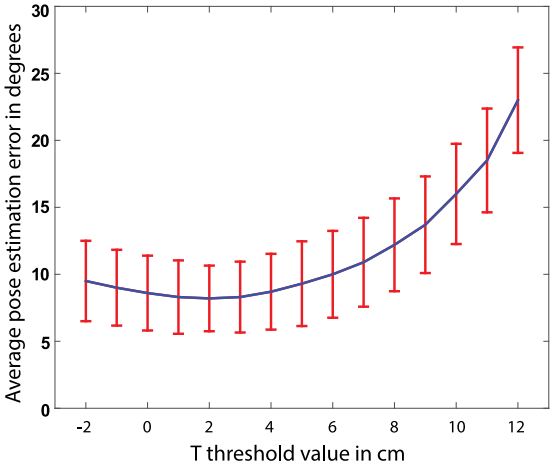

Fig. 11. Average error in body pose estimation as a function of the value of threshold T.

computing the corresponding average pose estimation error. The extracted results are presented in Fig. 11. When T assumed values within the range of -2 to 5 cm, the performance practically remained stable, with a slight increase close to the mid of the above range. For T values larger than 5, the performance gradually dropped. Evidently, the wide range $[-2, 5]$ of threshold values that allow for stable and high performance indicates that the value of T is not critical in our implementation. Currently, T is set at 2 cm, offering a safe operational value under all tested conditions.

Following threshold setting, we performed extensive experimentation to assess the performance of the proposed human pose estimation methodology and also contrast it to state-of-the-art methods. Detailed quantitative and qualitative experimental results are presented in the sequel. In all experiments the body pose is extracted for detected users with valid depths (face centroid) from the camera in the range from 1 to 4 m, due to decreasing depth accuracy for larger depths.

6.1 Quantitative Evaluation

The quantitative experiments were performed with (a) data from the Berkeley Multimodal Human Action Database (MHAD) [41] and (b) own collected data. MHAD constitutes a widely used dataset, and therefore was employed for quantitative assessment of our method as well as for comparative analysis against the NiTE [6] skeletonization tool of OpenNI [6]. Our own collected data were used for further evaluation of the proposed method. The relevant experiments involved up to three users, and prominent visual markers were used to facilitate extraction of ground truth data. This set of experiments allowed a thorough quantitative evaluation as well as its comparative assessment against state-of-the-art methods, such as the skeletonization module of Microsoft Kinect SDK [13] (based on the pose recovery methodology of [3], [4]) and NiTE. We note at this point that the Microsoft Kinect SDK skeletonization module was not assessed on the MHAD benchmark since, to the best of our knowledge, the available software is compatible only with *xed* files, which unfortunately is not technically possible for the MHAD sequences.

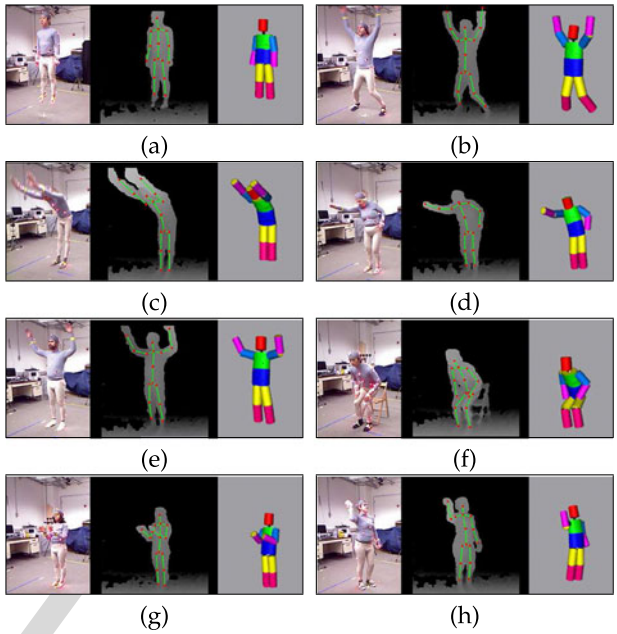

Fig. 12. Exemplar frames from the MHAD-based evaluation. 8 human subjects perform 8 different activities: (a) jumping in place, (b) jumping jacks, (c) bending, (d) boxing, (e) both hands waving, (f) sit down, (g) clapping, (h) throwing ball.

6.1.1 *Evaluation Based on Public Human Database MHAD*

The MHAD [41] database features 12 human subjects of both genders and various sizes, as depicted in leftmost images of Fig. 12. Each subject performs a set of 11 activities: jumping in place, jumping jacks, bending, boxing, both hands waving, right hand waving, clapping, throwing ball, sit down/stand up, sit down and stand up. A motion capture system provides ground truth information about the location of each joint of the users. Each user is observed by a set of cameras, namely conventional RGB cameras and RGB-D Kinect sensors. Throughout our experiments we used the RGB-D data provided by the "frontal" kinect (namely kin01). Fig. 12 depicts illustrative examples from the MHAD-based evaluation process, showing eight different users, of both genders and of varying ages, posture and sizes, performing various activities. The leftmost image of each column shows the RGB input of the "frontal" kinect, the middle image shows the estimated body skeleton superimposed on the corresponding depth input, while the rightmost image shows the 3D representation of the recovered pose.

For a rigorous quantitative analysis of our method's performance we measured established error statistics. In particular, we computed the mean angular error μE and the corresponding standard deviation σE (i.e., the estimation *precision*) by measuring the angle errors of all body joints. Table 2 provides the obtained statistics, namely the mean angular error (averaged differences between the actual and the estimated angles) μE and its standard deviation σE, expressed in degrees, together with the corresponding accuracy A, namely the percentage of estimations within 10 degrees from the ground truth.

We have conducted two sets of experiments using MHAD. The first set concerned the evaluation of performance across

TABLE 2
Comparative Analysis Using MHAD, with Respect to Different Users and Different Activities

	Different users Single activity (act. 04)						Different activities Single user (id 09)					
	TVR			NiTE			TVR			NiTE		
	μE	σE	A	μE	σE	A	μE	σE	A	μE	σE	A
Torso Joints	**7.75°**	**3.91°**	**95%**	10.32°	4.60°	87%	**7.58°**	**3.84°**	**95%**	10.83°	4.74°	85%
Arm Joints	**10.53°**	**4.92°**	**89%**	14.21°	5.68°	72%	**10.92°**	**5.10°**	**87%**	13.36°	5.39°	79%
Leg Joints	**9.25°**	**4.27°**	**91%**	12.48°	5.06°	82%	**9.1°**	**4.13°**	**92%**	11.95°	5.19°	84%

Performance statistics of our method are compared against NiTE™. μE is the mean angular error from the ground truth and σE the corresponding standard deviation, measured in degrees. Accuracy A depicts the percentage of estimations lying within 10 degrees from the ground truth.

different users. In that sense, 12 different subjects perform the same activity, i.e., activity 04 - boxing. The second set of experiments concerned the evaluation of performance with respect to different activities. For this purpose, we utilized sequences of the same person (subject id 09) performing all 11 activities. For both experimental sets, we have utilized the three first repetitions of each pair activity-user, provided by the dataset, thus summing up to approximately 20,000 frames. The extracted quantitative results are tabulated in Table 2. As can be observed, the performance of TVR compared favourably to the NiTE method for both sets of experiments. High accuracy rates and low overall errors were achieved, and at the same time our method outperformed NiTE in terms of estimation *precision* (σE).

6.1.2 Evaluation with Own Datasets

The proposed TVR-based pose estimation methodology enables to effectively cope with cluttered backgrounds and inter- and intra-person occlusions. Given that MHAD sequences feature uncluttered scenes with a single person performing various activities, to further assess the full potential of our methodology and rigorously evaluate its performance, we conducted additional comparative experiments, based on own collected sequences, annotated with ground truth information. The employed datasets, along with relevant annotation and processing results, are accessible at http://www.ics.forth.gr/cvrl/fbody/.

Six sequences, summing up to a total of approximately 3,500 frames, have been acquired as *xed* files via the Microsoft Kinect Studio and processed in order to provide performance statistics about our methodology, using marker-based ground truth data for the body joints. The examined sequences involve one, two or three users, acting in the scene, with apparent self-occlusions or occlusions among different users. In the sequences with multiple users, only one person is tracked to obtain quantitative assessment figures and the other present person(s) is(are) assumed to have a "dummy role", that is to partially occlude the subject that is tracked. The mentioned experiments are used to extract performance statistics of our methodology and compare it against state-of-the-art methods, namely the skeletonization module of Microsoft Kinect SDK([3], [4], [13]) and NiTE.

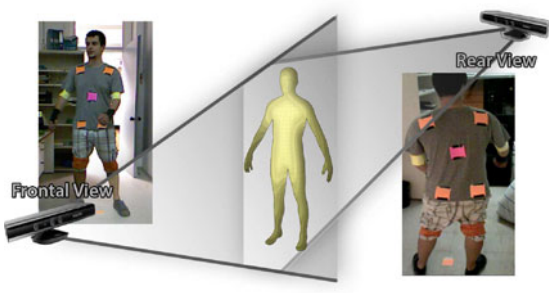

Fig. 13. Two-camera setup for ground truth extraction. Prominent color markers are attached on both sides of the user's body, denoting the underlying joint. Each marker is visible by at least one of the two cameras at all times.

Ground truth data: A visual marker-based acquisition technique along with a two-camera setup has been employed to obtain ground truth in our experiments, as illustrated in Fig. 13. Prominent color markers are attached on the both sides of the user's body, corresponding to the underlying joints. Semi-automatic annotation on the RGB-D sequences leads to the computation of the 3D position of each joint. The second (rear) camera is placed behind the user and slightly elevated in order to avoid interference between the two cameras.[1] The two-camera setup guarantees that all markers are visible by at least one of the cameras at all times and, therefore, facilitates the provision of ground truth information even for occluded joints. It is worth mentioning that the rear camera is only used for this purpose -i.e., ground truth computation of occluded joints- and is not utilized for the pose recovery process.

Comparative evaluation: For the actual quantitative assessment, each sequence has been processed by our methodology and the skeletonization module provided in the Microsoft Kinect SDK [13] and NiTE. The extracted results are summarised in Table 3. In order to better highlight the performance of the proposed methodology in the presence of occlusions, results in Table 3 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 a means to illustrate the effectiveness of *TVR*-based pose estimation in such cases. As can be observed, overall our methodology compares favourably to Microsoft Kinect SDK both in μE and σE, while it exhibits higher accuracy rates in all examined cases; at the same time, it outperforms NiTE with respect to all statistics. More importantly, however, in all cases of occluded interacting users, our method outperformed the state-of-the-art, demonstrating significantly more accurate performance.

Snapshots from the above quantitative experiments are shown in Fig. 14, where results are presented on four

1. For further information on Kinect calibration, the interested reader may also refer to [42].

TABLE 3
Comparative Assessment Results

	TVR						Microsoft SDK						NiTE					
	Overall			Under Occlusions			Overall			Under Occlusions			Overall			Under Occlusions		
	μE	σE	A	μE	σE	A	μE	σE	A	μE	σE	A	μE	σE	A	μE	σE	A
Torso Joints	**9.32°**	**3.92°**	**96%**	**13.52°**	**5.97°**	**86%**	12.14°	5.15°	93%	17.56°	7.5°	82%	13.05°	5.70°	85%	18.48°	7.75°	79%
Arm Joints	8.52°	5.47°	89%	**17.20°**	**6.45°**	**83%**	**8.37°**	**5.12°**	**90%**	21.07°	9.83°	74%	8.92°	5.63°	86%	24.13°	10.90°	69%
Leg Joints	7.32°	4.53°	**94%**	**13.09°**	**7.40°**	**83%**	**7.15°**	**4.49°**	91%	14.66°	9.57°	78%	7.93°	4.60°	78%	16.03°	10.06°	75%

μE = mean angular error throughout the sequences, σE = standard deviation of error and A = the accuracy rate. μE and σE are expressed in degrees.

exemplar frames of the sequences used for quantitative evaluation. For each frame, we provide the original RGB input (far-left image), the resulting skeleton of the recovered pose for the proposed methodology (image within the green frame), along with the corresponding estimation of the Microsoft Kinect SDK skeletonization module (image within the red frame) and NiTE (far-right image within the blue frame). Evidently, whenever the tracked user is occluded, the performance of both the Kinect SDK and NiTE methodologies deteriorate, whereas the performance of the proposed TVR-based method remains unaffected, as also shown in Table 3. For the interested reader, the examined sequences are publicly available at the aforementioned URL link.

6.2 Qualitative Results

Additionally to the experiments for quantitative and comparative evaluation, we have extensively tested our methodology in numerous scenarios of varying difficulty, with single and multiple users, involving intra- and inter-person occlusions. Illustrative instances from the named experiments are presented in Fig. 15. As can be observed, the proposed methodology succeeded in effectively tracking the pose of the full body in all cases, remaining unaffected by interactions and intra- or inter-person occlusions.

In addition, we have also investigated scenarios where the TVR-based methodology could possibly fail to provide an accurate pose estimation. Two representative such cases are illustrated in Figs. 16 and 17, where the leftmost image depicts the RGB input, the second left depicts the scene's 3D point cloud and the segmented human-body (in green), and the two rightmost images depict the 3D and skeletonized recovered pose. In the first case (Fig. 16), the left arm of the observed human is initially totally occluded (see frame 520 in Fig. 16). When a part of it re-appears (see frame 550),

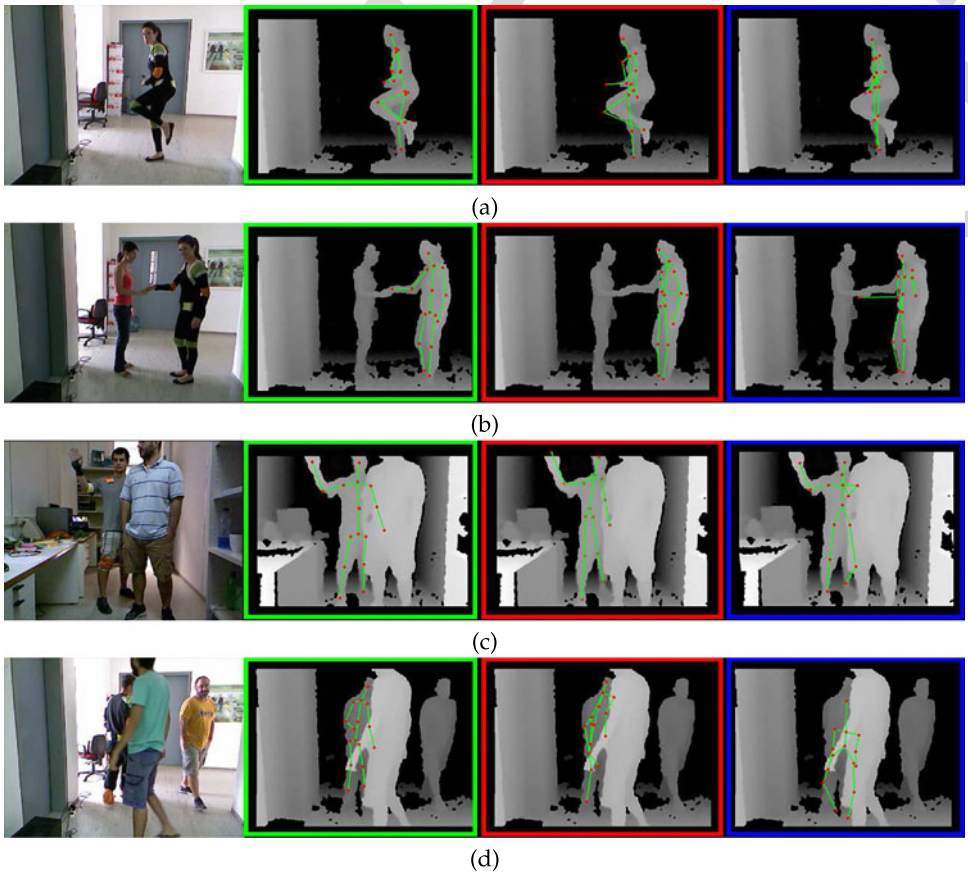

Fig. 14. Illustrative frames captured from the full body pose tracking sequences used for the comparative evaluation, where only the user wearing the markers is tracked and the other user(s) play the "dummy" role of the occluder. The results of the proposed methodology (image within the green frame) along with the corresponding estimation of the Microsoft Kinect skeletonization module (image within the red frame) and NiTE™ (far-right image within the blue frame) are superimposed on the depth image as the skeleton of the underlying pose.

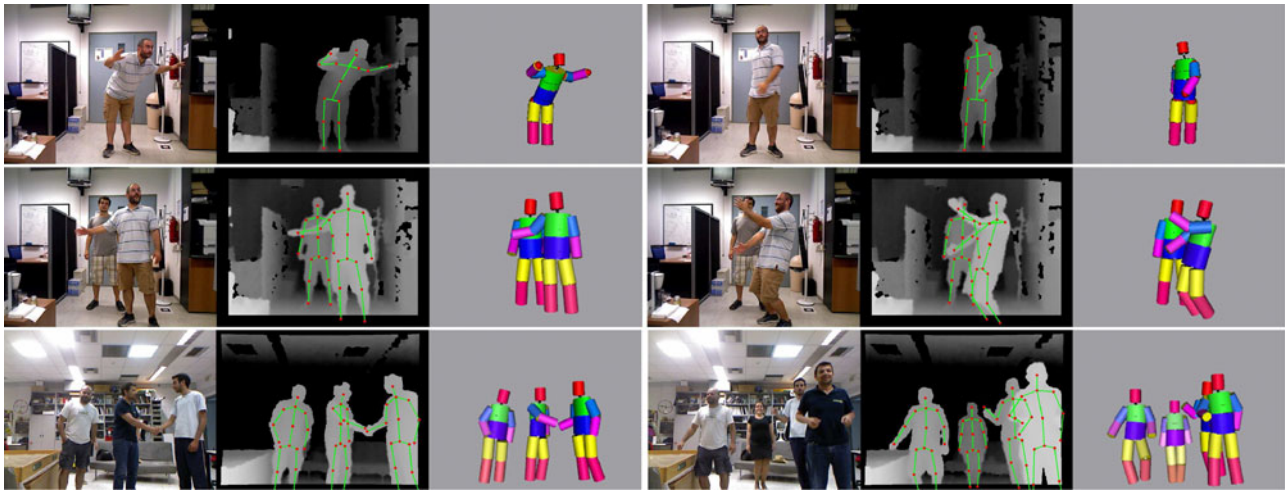

Fig. 15. Full body pose tracking results for single and multiple users scenarios. Users, of different sizes, move act and interact freely in the scene, causing multiple and severe inter- and intra-occlusions. In all instances, the far-left image is the RGB input while the middle and far-right images illustrate the skeletonized pose and the 3D pose, respectively.

the hypothesis sampling and tracking procedure fails to correctly estimate the arm pose, due to lack of hypotheses around the true pose. The second case (Fig. 17) shows the effect of severe errors in human-body segmentation. The fact that the user is adjacent to a very large object, namely the bookcase, negatively affects human-body segmentation by including large areas of the 3D point cloud which belong to the background. The latter results, in turn, in inaccurate pose estimation for the two left limbs.

7 Conclusions

In this paper we presented a robust, model-based approach for full body pose estimation and tracking in complex, real life scenarios in the presence of severe occlusions. The latter is a major advantage of the proposed approach 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 the *Top View Reprojection* concept, competent at treating the different body parts in a unified manner. Accordingly, the latter constitutes a main contribution of our work with proven capacity in recovering the full-body pose.

In all examined scenarios, the derived TVR-based formulation proved tolerant to human body diversities, caused by the users' gender, anthropometry and posture. This is due to the fact that no fixed model is employed, but model sizes are implicitly estimated with respect to the height of the user, based on established anthropometric measurements. This is also attested by the obtained results, where users of different gender, size and posture, are effectively tracked. Additionally, the proposed approach offers high tolerance to segmentation faults. As can be observed in the provided results, our methodology effectively coped with cases where accurate body segmentation is not feasible, such as in the case of the sitting action in the MHAD dataset or in the interaction scenarios presented in the illustrative analysis section (Section 6.2). However, the most important feature of the TVR-based approach regards its potential to cope with severe occlusions, either across body parts of the same user, or across different users. As demonstrated by the extensive experimental assessment, the proposed approach effectively treats occlusions via proper evaluation in the TVR scoring function.

Notwithstanding the effective and successful operation of our method in the aforementioned cases, failures may be experienced when the underlying assumptions are violated. Besides relevant results presented in Section 6.2 and in Figs. 16 and 17, further issues may affect the method's performance. Model initialization is accomplished assuming standing persons in upright positions as initial body configurations. Violations of this assumption, e.g., when a person is sitting, may lead to erroneous human-model instantiation and hence pose estimation. Additionally, the method is less tolerant to appearance changes caused by external factors, such as clothing variations. This is a common problem

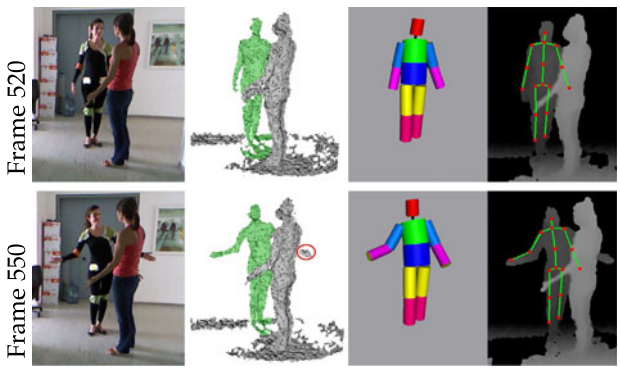

Fig. 16. Failure case involving tracking of occluded part. At frame 520 (top row) the left arm is totally occluded. After 30 frames (bottom row) a part of the arm re-appears, but since it remains partially occluded, tracking fails to recover the correct pose.

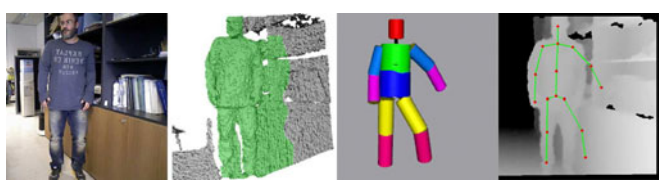

Fig. 17. Failure case involving human-body segmentation. Segmentation process fails to correctly segment the human from the scene, resulting to erroneous pose estimation.

throughout the relevant literature, leading to kinematic inconsistencies of the recovered pose and, thus, to inferior performance.

Our planned future work will be steered towards addressing the above issues. At the same time, it involves two immediate and interesting goals. The former regards the improvement of our method's performance using a more natural way of tracking, by introducing, apart from further kinematic constraints, physics rules to predict and detect more accurately the joints configuration. The latter addresses the study of more complex and involved interactions among users, including occlusions with longer durations, a case that challenges most contemporary approaches to pose-recovery. Moreover, the computational complexity constitutes a further challenge, since as already stated computational costs hinder real-time execution. The current single core CPU implementation provides processing rates of 3 to 4 frames per second. A first step towards speeding up the implemented system is to exploit the parallel execution capabilities of modern GPUs, which with a convenient global optimization, can possibly lead to real-time pose recovery and tracking.

Acknowledgments

This work has been partially supported by the EU Information and Communication Technologies Research Project JAMES (FP7-045388).

References

[1] Microsoft Corporation. (2012). Microsoft Kinect for Xbox 360. Redmond WA. [Online]. Available: http://www.xbox.com/en-US/xbox-360/accessories/kinect

[2] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 755–762.

[3] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1297–1304.

[4] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

[5] M. Sigalas, M. Pateraki, and P. Trahanias, "Robust articulated upper body pose tracking under severe occlusions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 4104–4111.

[6] Structure Sensor. (2015). OpenNI. [Online]. Available: http://structure.io/openni

[7] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[8] T. B. Moeslund, A. Hilton, V. Kruger, and L. Sigal, Eds., *Visual Analysis of Humans—Looking at People.* New York, NY, USA: Springer, 2011.

[9] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vis. Image Understanding*, vol. 108, pp. 4–18, 2007.

[10] S. Escalera, "Human behavior analysis from depth maps," in *Proc. 7th Int. Conf. Articulated Motion Deformable Objects*, 2012, vol. 7378, pp. 282–292.

[11] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recog. Lett.*, vol. 34, no. 15, pp. 1995–2006, 2013.

[12] J. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recog. Lett.*, vol. 48, pp. 70–80, 2014.

[13] S. Kean, J. C. Hall, and P. Perry, "Microsoft's Kinect SDK," in *Meet the Kinect.* New York, NY, USA: Springer, 2011, pp. 151–173.

[14] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 415–422.

[15] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 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. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 726–732.

[19] J. Charles and M. Everingham, "Learning shape models for monocular human pose estimation from the Microsoft Xbox Kinect," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1202–1208.

[20] 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*. New York, NY, USA: Springer, 2013, pp. 71–98.

[21] M. Kiefel and P. V. Gehler, "Human pose estimation with fields of parts," in *Proc. Comput. Vis.*, 2014, pp. 331–346.

[22] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2010, pp. 1–8.

[23] Y. Zhu and K. Fujimura, "Constrained optimization for human pose estimation from depth sequences," in *Proc. Asian Conf. Comput. Vis.*, pp. 408–418, 2007.

[24] Y. Zhu and K. Fujimura, "A Bayesian framework for human body pose tracking from depth image sequences," *Sensors*, vol. 10, pp. 5280–5293, 2010.

[25] E. Kalogerakis, A. Hertzmann, and K. Singh, "Learning 3d mesh segmentation and labeling," *ACM Trans. Graph.*, 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 *Proc. IEEE Int. Conf. Robot. Autom.*, 2010, pp. 3108–3113.

[27] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Estimating human 3d pose from time-of-flight images based on geodesic distances and optical flow," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshops.*, 2011, pp. 700–706.

[28] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli, "Decision tree fields," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1668–1675.

[29] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *Proc. Comput. Vis.*, 2014, pp. 33–47.

[30] D. Grest, J. Woetzel, and R. Koch, "Nonlinear body pose estimation from depth images," in *Proc. 27th DAGM Conf. Pattern Recog.*, 2005, vol. 3663, pp. 285–292.

[31] T. Koritnik, T. Bajd, and M. Munih, "A simple kinematic model of a human body for virtual environments," in *Advances in Robot Kinematics: Motion in Man and Machine*. New York, NY, USA: Springer, 2010, pp. 401–408.

[32] NASA. (1995). Man-systems integration standards - revision b. [Online]. Available: http://msis.jsc.nasa.gov/

[33] E. Churchill, J. McConville, L. Laubach, P. Erskine, K. Downing, and T. Churchill, *Anthropometric Source Book: Handbook of Anthropometric Data*, National Aeronautics and Space Administration (NASA), vol. 2, 1978.

[34] B. Stenger, A. Thayananthan, P. H. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical Bayesian filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1372–1384, Sep. 2006.

[35] H. Baltzakis, M. Pateraki, and P. Trahanias, "Visual tracking of hands, faces and facial features of multiple persons," *Mach. Vis. Appl.*, vol. 23, no. 6, pp. 1141–1157, 2012.

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

[37] H. Baltzakis and A. Argyros, "Propagation of pixel hypotheses for multiple objects tracking," in *Proc. Adv. Vis. Comput.*, 2009, vol. 5876, pp. 140–149.
[38] X. Zabulis, D. Grammenos, T. Sarmis, K. Tzevanidis, P. Padeleris, P. Koutlemanis, and A. A. Argyros, "Multicamera human detection and tracking supporting natural interaction with large-scale displays," *Mach. Vis. Appl.*, vol. 24, no. 2, pp. 319–336, 2013.
[39] L. Di Stefano and A. Bulgarelli, "A simple and efficient connected components labeling algorithm," in *Proc. Int. Conf. Image Anal. Process.*, 1999, pp. 322–327.
[40] D. B. Rubin, "Using the SIR algorithm to simulate posterior distributions," *Bayesian Statist.*, vol. 3, no. 1, pp. 395–402, 1988.
[41] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2013, pp. 53–60.
[42] C. Zhang and Z. Zhang, "Calibration between depth and color sensors for commodity depth cameras," in *Computer Vision and Machine Learning with RGB-D Sensors*. New York, NY, USA: Springer, 2014, pp. 47–64.

Markos Sigalas received the BSc degree in information & communication systems engineering from the University of the Aegean, Greece in 2005 and the MSc and PhD degrees in computer science from the University of Crete, Greece in 2008 and 2015, respectively. Currently, he is a postdoctoral researcher with the Computational Vision and Robotics Laboratory at the Institute of Computer Science of the Foundation for Research and Technology - Hellas (FORTH), Greece. His research is geared in the areas of visual human pose estimation and gesture recognition employed in human-computer interaction systems.

Maria Pateraki received the PhD degree in photogrammetry from the Swiss Federal Institute of Technology in Zurich (ETHZ), Switzerland in 2005 and has been a research associate at the University of Melbourne and the Cooperative Research Center for Spatial Information (CRC-SI) in Melbourne, Australia. Since 2008, she is a member of the Computational Vision and Robotics Laboratory at the Institute of Computer Science of the Foundation for Research and Technology - Hellas (FORTH), Greece. Her research interests include topics related to robotic vision, motion tracking, registration matching, and 3D reconstruction. She has published more than 45 papers in peer-reviewed journals and conference proceedings.

Panos Trahanias received the PhD degree in computer science from the National Technical University of Athens, Greece, in 1988. He is a professor with the Department of Computer Science, University of Crete and the Institute of Computer Science of the Foundation for Research and Technology - Hellas (FORTH), Greece. In the past he has been affiliated with the Department of Electrical & Computer Engineering, University of Toronto, Canada, and was a consultant to SPAR Aerospace Ltd., Toronto. Since 1993, he is with the University of Crete and FORTH. Currently, he is the head of the Computational Vision and Robotics Laboratory at FORTH where he is engaged in and supervises research and R & D projects in human-robot interaction, robot navigation, brain-based modeling, remote-access robotic systems, and augmented reality applications. He has been a cochair of the European Conference of Computer Vision 2010 (ECCV 2010) and Eurographics 2008. He has published more than 150 papers in technical journals and conference proceedings and has contributed in two books.