Tracking the articulated motion of the human body with two RGBD cameras

Damien Michel[a], Costas Panagiotakis[b], Antonis A. Argyros[a,c]

[a] *Institute of Computer Science, FORTH*
[b] *Department of Business Administration (Agios Nikolaos), TEI of Crete*
[c] *Computer Science Department, University of Crete*

Abstract

We present a model-based, top-down solution to the problem of tracking the 3D position, orientation and full articulation of the human body from markerless visual observations obtained by two synchronized RGBD cameras. Inspired by recent advances to the problem of model-based hand tracking [15], we treat human body tracking as an optimization problem that is solved using stochastic optimization techniques. We show that the proposed approach outperforms in accuracy state of the art methods that rely on a single RGBD camera. Thus, for applications that require increased accuracy and can afford the extra complexity introduced by the second sensor, the proposed approach constitutes a viable solution to the problem of markerless human motion tracking. Our findings are supported by an extensive quantitative evaluation of the method that has been performed on a publicly available data set that is annotated with ground truth.

Keywords: markerless human motion capture, 3D human tracking, 3D pose estimation, articulated object tracking, 3D reconstruction

---

Email address: `argyros@ics.forth.gr` (Antonis A. Argyros)
URL: `http://www.ics.forth.gr/~argyros` (Antonis A. Argyros)

1. Introduction

The estimation of the articulated motion of the human body is very important to a number of real world applications, ranging from surveillance to game design and human computer interaction. It is considered to be a difficult problem because of its high dimensionality and the variability of the tracked person regarding appearance, body dimensions, etc. A number of practical approaches simplify or even avoid these problems by using special hardware or by interfering with the subject and/or the environment by means of visual markers or full body suits [25]. However, unobtrusive, markerless tracking is definitely preferable since it does not interfere with the environment, the subject and its actions. The methods that use markerless visual data as their only input fall into two basic categories, the top-down and the bottom-up ones. Top-down approaches can provide accurate, physically plausible solutions at the cost of a high computational complexity. Bottom-up methods are typically faster, but rely on a discrete set of training poses whose selection determines the accuracy of the obtained results.

In this paper, we propose a model-based, top-down solution to the problem of tracking the 3D pose and articulation of a human body. This is formulated as a optimization problem that minimizes the discrepancy between the 3D occupancy of hypothesized instances of a human body model and the volume reconstructed from the observations. The input to the method comes from two wide baseline, extrinsically calibrated, off-the-shelf RGBD sensors [27] whose depth maps are fused to give rise to the required volumetric representation of the human body. The required volumetric representation can also be obtained by computing the visual hull of a human body figure through standard techniques [24] employing a network of multiple, conventional cameras. Nevertheless, the setup of two RGBD cameras is preferable due to its lower cost and complexity.

Optimization is performed based on an a stochastic method (Particle

Swarm Optimization - PSO) [9]. We demonstrate experimentally the accuracy achieved by the baseline PSO (*bPSO*) optimization method that borrows directly from recent advances on the problem of tracking the articulated motion of the human hand [15]. We also propose a new variant called perturbed PSO (*pPSO*) which systematically perturbs the solutions provided by *bPSO*. We demonstrate that *pPSO* outperforms *bPSO*. We also compare both *bPSO* and *pPSO* with a widely employed method [17] which we will refer to as *OpenNI* for estimating the human skeleton based on a single RGBD camera. Experimental results show that compared to *OpenNI*, the proposed *pPSO* method provides more accurate results. Thus, in applications where increased accuracy is worth the extra complexity introduced by the second sensor, *pPSO* is the preferred choice. On the other hand, both *bPSO* and *pPSO*, being model-based tracking approaches, require knowledge of the parameters of the human body and its 3D pose in the first frame of a sequence. To address these practical problems, we also propose and evaluate another variant, called *HYBRID*, which combines *pPSO* and *OpenNI*, aiming at combining the merits of both in a single method.

The rest of the paper is organized as follows. Section 2 reviews existing approaches to the problem of markerless human motion capture and tracking. Section 3 describes the proposed approaches, by detailing the human body model employed, the observation model, the objective function used to compare hypotheses and observations as well as the optimization methods used to minimize it. It also presents how *pPSO* and *OpenNI* are combined into the *HYBRID* method. Section 4 presents the experimental evaluation of the proposed method in a standard dataset that in annotated with ground truth. Finally, Section 5 summarizes the paper by drawing the main conclusions from this research.

2. Related work

Because of its high theoretical and practical interest, human motion capture based on vision has been the theme of numerous research efforts. The complete review of these works is beyond the scope of this paper. The interested reader is referred to [11, 19] where extended surveys are provided. More recently, Chen et al. [2] surveyed methods for human motion estimation based on depth cameras.

Most commercial solutions to the problem of human motion capture make use of special markers that are placed on carefully selected (e.g., joints) points of the subject's body. In this paper we are interested in markerless motion capture techniques because, being unobtrusive, present obvious practical advantages over the marker-based solutions.

Markerless human motion capture techniques may be classified into two broad classes, the *bottom-up* and the *top-down* ones. Bottom up methods [22, 1, 20, 18, 21] extract a set of features from the input images, and try to map them to the human pose space. This is achieved with a learning process that involves a typically large database of known poses that cover as much as possible the whole human poses search space. The type of descriptors employed, the mapping method and the actual poses database are the factors determining the accuracy and efficiency of these methods. Due to their nature, most of their computing time is spend on the offline processes of database creation and mapping, while the online performance is rather good.

Top-down approaches [4, 6, 5, 26, 3, 28] use a fully articulated model of the human body and try to estimate the joints angles that would make the appearance of this model fit best the visual input. The model is usually made of a base skeleton and an attached surface. In some methods, complex surface deformations are allowed [6]. Having defined a model of the human body, different pose hypotheses can be formed. A typical top-down method consists of generating hypotheses and comparing them to the input visual data. The comparison is performed based on an objective function that measures the discrepancy between a pose hypothesis and the actual observations. The minimization of this objective function determines the pose that best explains the available observations. Typically, this is formulated as an optimization problem that amounts to the exploration of a very high dimensional search space. Kinematic constrains based on physiological data are often applied to the model, excluding non realistic poses and reducing significantly that search space. Constraining not only the pose but also the motion itself can further help reducing the complexity, for example with Kalman filters [10]. However, this means a reduced generality and the necessity to build and learn human motion models.

The main advantage of top-down methods is their flexibility. The employed model can be changed easily, and the whole search space can be explored without any form of training. The price to pay for this flexibility is the computational cost of the online process. Due to their generative nature, most of the computational work needs to be performed online. Two more shortcomings is the requirement for knowing the body model parameters of each individual and the requirement of providing an initial pose to be tracked.

Instead of trying to estimate the full body model in a single step, a variety of methods first identify body parts. Then, they either report them as the solution or they further combine them into a full model [20, 21]. As in the case of hand tracking and according to the related categorization of Oikonomidis et al. [16], we can identify *disjoint evidence methods* and *joint evidence methods* [4, 6, 5, 26, 3, 28]. Joint evidence methods handle effortlessly collisions, self occlusions and all part interactions while disjoints evidence methods have to handle them explicitly.

This paper presents a model-based, top-down pose estimation method that employs a single hypothesis. Furthermore, it is a joint-evidence method. The 3D body pose recovery is treated as a minimization problem whose objective function quantifies the discrepancy between the 3D structure and appearance of hypothesized 3D body model instances, and visual observations of a human body. Observations come from two off-the-shelf Kinect sensors. Optimization is performed through a variant of PSO tailored to the needs of the specific problem. Zhang et al. [28] proposed a solution to articulated human motion tracking that is also based on two RGBD sensors. Their approach differs in the observation model that is being used and in the employed optimization technique. Other versions of PSO have been employed in the past for human body pose tracking [26, 3, 12], as well as for multicamera-based and RGBD camera-based hand pose estimation [14, 15]. For example in [26] body-pose hypotheses are used to render silhouettes that are compared with respective observations. They adopt a hierarchical approach to the problem and employ a PSO variant to solve it. Their approach differs from our methodology both in the observation model and in the optimization strategy. In particular, we propose and present a novel PSO variant (*pPSO*). We also present another variant, called *HYBRID*, which combines the *pPSO* with the *OpenNI* method and lifts the requirements for knowing the body model parameters and the initial body pose.

An extensive evaluation on a standard, publicly available data set annotated with ground truth shows that *pPSO* outperforms the baseline (*bPSO*) optimization method and is more accurate than an extensively used, bottom up *OpenNI* approach [17] to the same problem. Finally, the *HYBRID* approach performs slightly worse compared to *pPSO* with respect to accuracy. This is because *pPSO* operates with more accurate, manually derived human body models, while *HYBRID* estimated them automatically, but with some error. Nevertheless, *HYBRID* is far more practical as it avoids cumbersome initialization processes.

3. Tracking human body articulations

The input to the proposed method is a volumetric representation of the human body (Figure 1(e)). This can be obtained by two RGBD sensors that are configured in a wide baseline setup, or by computing the visual hull of the human body based on a number of conventional RGB cameras. The first option is preferable because it involves fewer cameras and because the resulting volumetric representation describes more accurately a 3D shape compared to its visual hull. The depth information also facilitates the segmentation of the human figure from the rest of the environment.

The adopted 3D human model comprises of a set of appropriately assembled geometric primitives. Each body pose is represented as a vector of 35 parameters. Body articulation tracking is formulated as the problem of estimating the 35 body model parameters that minimize the discrepancy between the body hypotheses and the actual observations. To quantify this discrepancy, a representation of the volume occupied by a given body model is produced and compared to the volumetric representation generated by the two RGBD cameras. An appropriate objective function is thus formulated and a variant of PSO is employed to search for the optimal body configuration. The result of this optimization process is the output of the method for the given frame. Temporal continuity is exploited to track the body articulation in a sequence of frames. The remainder of this section describes these algorithmic steps in more detail.

3.1. Observing a human

At a certain moment in time, the input to the method is a set of two 640×480 depth images of a human, as provided by two intrinsically and extrinsically calibrated RGBD sensors [27]. Figure 1(a), (b) and (c), (d), show the RGB and depth information acquired by two such sensors. Foreground is segmented through change detection that is performed on the depth information. More specifically, depth views of the environment without and with the human are available. Image points that exhibit pixelwise depth differences that exceed a certain threshold are detected and attributed to the scene foreground. The threshold used in this process is determined based on a study of the depth error estimation of the Kinect [23]. The resulting largest foreground blob in each depth image is kept for further consideration. A conservative estimation of the human spatial extend is performed by applying a closing morphological operator to these blobs with a circular mask of radius $r = 1$. Due to sensor limitations, the depth of some points that lie within the detected foreground is unknown. However, it is necessary to give at least an approximate depth value to these points in order to produce a correct 3D reconstruction that is needed for further processing. Thus, the depth at such points is set equal to the median of the non-null depths of points within a radius of 2 pixels. Averaging instead of median filtering was also tested, giving rise to negligible differences in accuracy. The depth values of the background pixels is set to infinity.

A 3D space of $150 \times 150 \times 150$ voxels is then considered. Each voxel is a cube with side equal to $s_v = 15mm$ resulting in a volumetric representation of a 3D space of $2.25 \times 2.25 \times 2.25$ meters. The center of this space is set equal to the mean position of the 3D points located onto the largest foreground blob of one of the two RGBD cameras. Each voxel of this space is set to 1 representing the initial assumption that the whole voxel space is fully occupied by the human figure. Then, the depth values of the two extrinsically calibrated RGBD cameras are used to carve out voxels that are not occupied. More specifically, for each 3D voxel v we compute its Euclidean distance d from an RGBD camera and compare this to the distance $\hat{d}$ that is estimated from the depth values provided by that camera. If $d < \hat{d}$ then this voxel should be carved out and takes the value of 0. This test is performed for both RGBD views. At the end of this process, voxels with a value of 1 provide the volumetric representation o_v of the human. An example of such a representation is shown in Figure 1(e).

We also compute the outer surface o_s of o_v. On o_s, we apply a 3D distance transform using a spherical kernel of a radius equal to 7 voxels. In the resulting map o_{sd}, voxels that also belong to o_s have a value of 1, voxels that are more than 7 voxels apart from any surface voxel have a value of 0 and the rest of the voxels have a value from 1 to 0 that is inversely proportional to their distance (0 to 7 voxels) to the closest voxel of o_s.

The observation model $o = \{o_v, o_{sd}\}$ that feeds the rest of the process consists of o_v and o_{sd}.

3.2. Modeling a human

The employed human model consists of a main body, two legs, two arms and the head (Figure 1(f)). The main body is modeled with two articulated elliptic cylinders and three ellipsoids for the caps and the junction. The head is made of one cylinder and a sphere. Each arm consists of three spheres and two truncated cones, while a leg has two such cones, two spheres for the knee and the ankle, respectively, and one ellipsoid for the foot. In Figure 1(f), the human model is depicted with color-coded geometric primitives (yellow for elliptic cylinders, red for ellipsoids, green for spheres and blue for truncated cones). For the *bPSO* and the *pPSO* variants, the parameters of these primitives (lengths, radii, etc) are manually set. For the *HYBRID* approach, these are automatically estimated based on the output of the *OpenNI* method.

The kinematics of each arm is modelled using six parameters encoding angles. Two parameters determine the shoulder position with respect to the torso, three parameters the upper arm with respect to the shoulder and one parameter the elbow with respect to the upper arm. Six parameters are also used for a leg, three for the root, one for the knee and two for the ankle. Two parameters are used for the head, and three parameters for the articulation between the torso and the hip. The global position of the body is represented using a fixed point on the hip. The global orientation is parametrized using Euler angles. The above parametrization encodes a 35 degrees of freedom (DOFs) human model with each DOF represented by a single parameter.

3.3. Evaluating a human hypothesis

Having defined the parametric 3D model of a human, the goal is to estimate the model parameters that are most compatible to the visual observations (Section 3.1). To do so, given a human pose hypothesis h, a volumetric representation h_v of the human model at pose h is generated through graphics rendering. The volume h_v is rendered in a voxel space with identical characteristics to those of o_v. A distance function $D_v(h_v, o_v)$ is defined as follows:

$$D_v(o_v, h_v) = 1 - \frac{2\sum(o_v \wedge h_v)}{\sum(o_v \wedge h_v) + \sum(o_v \vee h_v)}. \tag{1}$$

Intuitively, D_v quantifies the volumetric discrepancy between the observation volume o_v and the hypothesis volume h_v. In Eq.(1), symbols $\wedge$ and $\vee$ denote logical operations between the binary values of corresponding voxels and summations are over the set of all voxels. When the volumes h_v and o_v are disjoint, the quotient in Eq.(1) is equal to 0. If these volumes are identical and coincide, the quotient is equal to 1. Thus, D_v is equal to 0 if volumes coincide and 1 if they are totally disjoint.

Besides volumetric discrepancy, we also compute a surface alignment discrepancy. To define this, we first compute the outer surfaces h_s of the volumetric representation h_v of the hypothesis h. Then, the surface alignment discrepancy $D_s(o_{sd}, h_s)$ is defined as:

$$D_s(o_{sd}, h_s) = 1 - \frac{1}{n_p}\sum(o_{sd} \cdot h_s). \tag{2}$$

In Eq.(2), the sum is over all voxels and $\cdot$ denotes standard multiplication of the values of the corresponding maps. o_{sd} is defined as in Section 3.1. Thus,

D_s takes a value of 0 if the surface of the hypothesis coincides perfectly with that of the observation.

Given the distance functions D_v and D_s, it is now possible to define the function $E(o, h)$ that measures the discrepancy between the observation o and a given body pose hypothesis h:

$$E(o, h) = D_v(o_v, h_v) + D_s(o_{sd}, h_s). \tag{3}$$

The minimization of $E(o, h)$ with respect to h yields the body pose that best (as quantified by the objective function) explains the observations. The next section details how this minimization is actually achieved.

It should be noted that the reconstructed volume of the subject is a superset of his/her actual volume. This is a direct effect of the fact that 3D reconstruction is performed with a space-carving-like method, which cannot handle occlusions. As an example, assume that a human is observed from sideways and the volume between his arm and torso cannot be reconstructed, as the arm occludes this space. Having two RGBD cameras in a wide baseline configuration minimizes this type of effects but does not eliminate them. This is exactly why the objective function of the optimization process employs the term $D_s(o_{sd}, h_s)$ that is related to the coverage of the surface of the reconstruction. As the 3D reconstruction is a superset of the actual volume, solutions that "float" in the 3D reconstructed space are equally good in terms of volume coverage. For the case of the previous example, these are all arm configurations that occupy some of the reconstructed volume. From those, the additional surface coverage term selects the one that best matches also the visible surface of the 3D reconstruction. If the 3D reconstruction did not suffer from these extra, non-carved voxels, then the volume term would be enough to guide optimization and the surface term would not introduce any further useful constraints.

3.4. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a popular optimization algorithm that was introduced by Kennedy and Eberhart in [8, 9]. PSO looks for the optimum of an objective function employing a population of entities that evolve according to rules that emulate social interaction.

Cenrtal to PSO are the notions of *particles* and *generations*. A particle holds a position/candidate solution in the parametric space where the search is performed. Each particle can estimate the fitness of its position by evaluating the objective function at that point. Each particle is aware of the position at which it has achieved its own best objective function value. It also knows the global best position that has ever been achieved by any of the rest of the particles. Two forces are defined that attract a particle to these two positions. The particles evolve themselves by moving in the search space under the previously described forces in iterations called generations. The details of this process are provided in [15].

It has been observed that given enough particles and generations, the swarm reaches the global minimum of the objective function. The required number of particles and generations is problem-dependent and, thus, experimentally identified. A number of studies have shown that PSO is very competent in optimizing complex, multidimensional, multimodal, non-differentiatable objective functions. The product of the number of particles to that of generations determines the computational requirements of the optimization process. This is because this product represents the number of objective function evaluations that constitutes the most computationally demanding part of the algorithm.

Typically, the particles are initialized at random positions and zero velocities. Each dimension of the multidimensional parameter space is bounded in some range. As in [15], if during the position update a particle has a velocity that forces it to move to a point p_o outside the bounds of the parameter space, that particle effectively moves to the point p_b inside the bounds that minimizes the distance $|p_o - p_b|$.

3.5. Baseline PSO (bPSO)

In this work, PSO operates on the 35-dimensional 3D body pose parameter space. This also implies that the intrinsic human model parameters (lengths and radii of the primitives of the human model) need to be known in advance. The objective function to be optimized (i.e., minimized) is $E(O, h)$ (Eq. 3) and the population is a set of candidate 3D body poses hypothesized for a single frame. Thus, the process of tracking a human requires the solution of a sequence of optimization problems, one for each acquired frame. By exploiting temporal continuity, the solution over frame F_t is used to generate the initial population for the optimization problem for frame F_{t+1}. More specifically, the first member of the population h_{ref} for frame F_{t+1} is the solution for frame F_t. This implies that for the first frame F_0, a human body configuration close to the actual one needs to be provided. The rest of the population consists of perturbations of h_{ref}. The variance of these perturbations is experimentally determined and depends on the characteristics of the observed motion and the image acquisition frame rate. The optimization for frame F_{t+1} is executed for a fixed amount of generations. After all generations have evolved, the best hypothesis h_{best} constitutes the solution for time step $t + 1$.

3.6. Perturbed PSO (pPSO)

It has been verified experimentally that *bPSO* is competent in estimating the $6D$ global pose of the human body. However, the estimation of the 29 remaining parameters that are related to limb angles is not equally satisfactory. The swarm often gets stuck to local minima. To overcome this problem and to increase accuracy, we propose a PSO variant which we call *pPSO* that performs systematic perturbations/randomization on the articulation parameters. More specifically, the human body model is decomposed into seven branches, as shown in Figure 2. Each branch consists of a set b_p of primitives and has a set b_d of internal articulation parameters. *pPSO* operates exactly as *bPSO* for a percentage of its generations. This percentage has been identified experimentally to be 40%. After those generations, each particle is perturbed in a very specific way. First, one branch is randomly selected. Then, only the parameters of this branch are perturbed and replicated in the global particle representation. Additionally, the local (particle-dependent) best position for this particle is reset to the new particle position. After each and every particle is perturbed in this way, all particles are left to interact as in the *bPSO* scheme for g_p generations. In all reported experiments, the value of g_p was set to 6 generations. The process is repeated until the rest 60% of the PSO generations are lapsed.

Two different perturbation strategies have been identified and tested. In the first one, samples are drawn from a uniform distribution in the range of minimum/maximum allowed values of the corresponding parameter. In the second case, samples are drawn from a Gaussian distribution centered at the particle's previous position and with a standard deviation equal to one sixth of the range of the corresponding uniform distribution. It has been verified experimentally that both result in the same tracking error but the error variance is slightly higher for the case of Gaussian perturbation. In fact, Gaussian perturbation performs better on slow actions, but worse on fast actions. This can be explained by the fact that Gaussian sampling performs a more local search in the parameter space compared to the uniform sampling, making it more difficult to recover from track loses.

Special care should be taken when a particle is perturbed with respect to its torso or hip branches. As shown in Figure 2, these two branches are not leafs in the kinematics hierarchy. Thus, the perturbation of these branches affects the parameters of the rest of the branches, too. For this reason, as soon as these branches are perturbed, the global human kinematics model consistency needs to be enforced. This is achieved by employing inverse kinematics. Consequently, a perturbation on the torso or the hip will in fact influence most, if not all the 35 parameters.

The particular scheme for perturbing particles/candidate solutions is justified by the study of the morphology of the human body and the objective function of the optimization problem. The human torso accounts for most of the body's volume and, therefore, for the largest part of the objective function. Fine tuning a solution requires checking alternative configurations of the human limbs that are much smaller in size and less influential to the objective function. Thus, a targeted particle perturbation that affects only a branch at a time gives more chances to the algorithm to explore the true minimum of the objective function.

Another reason why perturbation proves valuable stems from the coarseness of the employed model. Consider, for example, the case in which the arms of a subject are stretched straight and turned around at the shoulder joint for 90 degrees. In this particular case, the method will probably loose tracking of the roll around the shoulder joint because this motion does not produce some significant, observable difference in the volume occupied and the surface covered by the subject. Due to the perturbation step of *pPSO*, several solutions relatively far from the computed one will be tested. This prevents *pPSO* from getting trapped in local minima and enables the effective tracking of subsequent, unambiguous motions.

3.7. Hybrid human body pose tracking (HYBRID)

As stated is sections 3.5 and 3.6, both *bPSO* and *pPSO* require

- Knowledge of the human body shape parameters (i.e., lengths, radii of the geometric primitives comprising the human body model).

- A coarse estimation of the human body pose for the first frame of a sequence.

These requirements hinder the practical exploitation of these algorithms because their fulfilment is associated with considerable effort. In order to alleviate this problem, we capitalize on the *OpenNI* appearance-based method for human skeleton estimation [17] to come up with a new variant, which we call *HYBRID* and which operates as follows.

At a first stage, given an input RGBD sequence of an articulating human, the *OpenNI* method is employed to estimate the articulation. The result of this process is twofold:

1. The lengths of the parts of a skeletal model of the human body for each frame of the sequence.
2. An estimation of the human body pose for each frame of the sequence.

Based on (1), we compute a human body model of constant shape parameters that consists of primitives that are compatible to those estimated by the *OpenNI* method. More specifically, on top of the 35 mobilities of our model, 9 parameters (constant for each sequence) are added to be able to fit the human body model to a particular subject. These are the upper body length (UBL), the lower body length (LBL), the shoulders-neck distance (SND), the head neck distance (HND), the legs-hip distance (LHD), the back-arm length (BAL), the forearm length (FAL), the back-leg length (BLL) and the front-leg length (FLL). Table 2 presents ground truth values as well as estimations of these parameters for a number of subjects. The parenthesis next to the name of each parameter refers to the corresponding body segment(s) in Figure 2.

For a certain sequence, the human skeletons estimated by *OpenNI* provide the 3D positions of human body joints and extremities. For each valid frame, the aforementioned distances are calculated (9 parameters, 15 distances, since some appear twice for the two arms and legs). The median value for each parameter across all the sequence is selected as the representative one. Other dimensions (i.e, radii of primitives) are set accordingly, based on anatomical studies.

These parameters are then used to define the human body model and inverse kinematics fits the model to the *OpenNI* solution for the first frame. Then, *pPSO* is employed to track the derived human model. Moreover, the solution suggested for each frame by *OpenNI* identifies a particle for *pPSO*. This has the additional advantage that in case of a tracking loss from *pPSO*, tracking can be recovered by considering the fairly accurate *OpenNI* recommendation as a candidate solution.

4. Experimental evaluation

The experimental evaluation of the proposed method was based on the Berkeley Multimodal Human Action Database (MHAD) [13]. This dataset features 12 human subjects.

Figure 3 shows one frame of each subject. From this figure it can be verified that the employed data set includes subjects of considerable variability with respect to age, size and body types. This is also shown quantitatively in Table 2.

The subjects of the dataset perform 11 different activities (01-jumping, 02-jumping jacks, 03-bending, 04-punching, 05-waving two hands, 06-waving one hand, 07-clapping, 08-throwing, 09-sit down/stand up, 10-sit down and 11-stand up). In each sequence, each activity is repeated several times. A motion capture system has been used to provide ground truth information regarding the position of all joints in all sequences. Additionally, the activities are recorded with a multicamera setup consisting of several conventional cameras as well as by two extrinsically calibrated Kinect sensors. In all experiments reported in this paper, the RGBD data provided by the two

Kinect sensors feed the proposed methods. The resulting tracking results are compared against the ground truth resulting from the motion capture data.

To quantify the accuracy in body pose estimation, we adopt the metric used in [7]. More specifically, the distance between a set of corresponding 3D points in the ground truth and in the estimated body model is measured. Each such point (four per leg, three per arm and one for the head) is marked in Figure 2 with a red “x”. The average of all these distances over all the frames of the sequence constitutes the resulting error estimate Δ. Another metric reports the percentage of these distances that are within some predefined threshold A_t. We will refer to this metric as the accuracy in human body pose estimation. A_t was set to $10cm$ for all experiments. For example, an accuracy of $A_t = 70\%$ for a sequence means that 70% of the joints were estimated at positions that are within less than $10cm$ from the ground truth, in all frames.

Several experiments were carried out to assess quantitatively and qualitatively the accuracy and the performance of the proposed human articulation tracking method. The goal of the first experiment was to assess the error in joints position estimation as a function of the computational budget devoted to PSO. To do so, we choose one of the sequences of the MHAD dataset that consists of 80 consecutive human poses showing a human performing activity 02 (jumping jacks). The rationale for selecting this particular activity and sequence is that (a) it is executed in high speed and (b) it involves the whole body, so all body model parameters change values as a function of time. Thus, it is expected that this sequence constitutes a worst case scenario, at least among activities represented in the specific dataset.

Figure 4 illustrates the error Δ in joints position estimation as a function of the *pPSO* parameters (number of generations and particles per generation). As explained in Section 3, the product of these parameters determines the computational budget of the proposed methodology, as it accounts for the number of objective function evaluations. The horizontal axis of the plot denotes the number of PSO generations. Each plot of the graph corresponds to a different number of particles per generation. Each point in each plot is the average of the error Δ for 5 runs of an experiment with the specific parameters. A first observation is that Δ decreases monotonically as the number of generations increase. Additionally, as the particles per generation increase, the resulting error decreases. Nevertheless, employing more that 65 generations and more than 200 particles results in a reduction of the error Δ that is disproportionally low compared to the increase in the required computational budget. For this reason, 200 particles evolving in 65 generations was retained in all further experiments.

The second experiment aimed at evaluating the performance of the methods across different human subjects. All twelve sequences showing the twelve different subjects performing the same activity (activity 04, boxing) were considered. *bPSO* and *pPSO* require knowledge of the parameters of the human body models as well as body configuration parameters for the first frame of a sequence. In our experiments, these subject-specific model parameters and initial model configurations were estimated manually for *bPSO* and *pPSO*, and automatically for *HYBRID*, based on the results obtained by the *OpenNI* method (see Section 3.7). Additionally, the *pPSO*, *bPSO* and *HYBRID* methods were assigned exactly the same computational budget.

Figure 5 illustrates the error Δ and the accuracy of the *pPSO* and the *HYBRID* methods. For the purposes of comparative evaluation, errors and accuracies are also provided for *bPSO* and for the *OpenNI* skeleton estimation method [17]. Table 1 summarizes the individual errors and accuracies shown in Figure 5. It can be verified that the *pPSO* method outperforms all other methods in all aspects (average error, standard deviation of error and accuracy). *HYBRID* outperforms the *bPSO* and *OpenNI*, showing that model-based optimization improves the guess made by *OpenNI*. Still,

Method	Mean Δ	Std. Δ	Accuracy (%)
OpenNI	52.9	49.5	87.3
bPSO	62.2	69.5	82.3
pPSO	**41.8**	**33.1**	**94.4**
HYBRID	45.5	44.1	92.5

Table 1: Comparison of *pPSO* and *HYBRID* with the baseline PSO method (*bPSO*) and the *OpenNI* method for the case of different humans performing the same action (boxing).

HYBRID does not outperform *pPSO*. The reason for this is that *HYBRID* operates on automatically estimated human body models that are less accurate compared to the ones on which *pPSO* operates (identified manually). Table 2 shows characteristic distances and metrics regarding the 12 human subjects as measured manually (columns G) and as estimated in the *HYBRID* algorithm (columns H). It can be verified that the body dimensions estimated by *HYBRID* deviate considerably from the actual, ground truth measurements.

One interesting question that arises is how *HYBRID* would perform if it was provided with the manually estimated human body models on which *pPSO* operates. It turns out that in this case, mean Δ is 40.9, the standard deviation of Δ is 32.6 and the accuracy is 94.7%. Thus, *HYBRID* outperforms marginally *pPSO* in that case. The slight difference in performance between *HYBRID* and *pPSO* is explained by the fact that no track loss occurred during *pPSO* tracking from which *HYBRID* could recover.

In a third experiment, the goal was to assess the proposed method with respect to different activities. For that purpose, the evaluation was performed on image sequences showing a single subject performing the eleven different

Subject	S01		S02		S03		S04		S05		S06	
Metric	G	H	G	H	G	H	G	H	G	H	G	H
UBL (HI)	26	19	30	21	33	21	29	20	32	22	28	20
LBL (IJ)	15	19	17	21	18	21	17	20	19	22	17	20
SND (CH, C'H)	19	15	19	15	17	14	15	15	17	16	19	14
HND (GH)	20	25	20	25	20	25	20	21	20	26	20	20
LHD (FJ, F'J)	10	9	11	9	10	8	9	9	9	9	9	8
BAL (BC, B'C')	24	25	28	27	31	28	24	23	26	28	26	26
FAL (AB, A'B')	23	26	25	31	26	32	24	25	25	31	24	27
BLL (EF, E'F')	36	41	43	47	44	47	37	39	42	45	42	44
FLL (DE, D'E')	42	37	48	42	47	43	41	35	45	42	45	41

(a)

Subject	S07		S08		S09		S10		S11		S12	
Metric	G	H	G	H	G	H	G	H	G	H	G	H
UBL (HI)	25	20	30	20	27	20	27	21	28	20	24	20
LBL (IJ)	15	20	18	20	15	20	16	21	16	20	21	20
SND (CH, C'H)	17	17	18	15	15	13	17	14	17	15	18	14
HND (GH)	20	20	20	21	20	17	20	24	20	25	20	24
LHD (FJ, F'J)	8	7	9	9	8	7	8	8	8	9	9	8
BAL (BC, B'C')	22	25	24	26	23	22	27	25	26	27	25	25
FAL (AB, A'B')	22	24	24	26	23	27	24	29	24	28	22	27
BLL (EF, E'F')	35	39	39	40	35	42	41	43	41	44	38	43
FLL (DE, D'E')	41	35	43	39	41	37	43	40	44	41	41	39

(b)

Table 2: Characteristic measures of the shape of the human subjects of the MHAD dataset, (a) subjects 01-06, (b) subjects 07-12. Columns (G) are the manually measured, ground truth values and columns (H) the one estimated by the *HYBRID* method. The parenthesis next to the name of each measure refers to the corresponding body segment(s) in Figure 2.

Method	Mean Δ	Std. Δ	Accuracy (%)
OpenNI	54.5	46.2	86.3
bPSO	50.6	48.8	89.5
pPSO	**39.3**	27.3	**96.3**
HYBRID	42.8	**25.2**	**96.3**

Table 3: Comparison of *pPSO* and *HYBRID* with the baseline PSO method (*bPSO*) and the *OpenNI* method for the case of all actions performed by the same subject (subject 09).

activities. Figure 6 illustrates the obtained results in a way analogous to that of Figure 5. Again, *pPSO* outperforms the rest of the methods with respect to the mean error Δ. It should also be noted that for actions like bending (action 03) and sit-down/stand-up (action 09) that exhibit considerable self- and body-object occlusions, the proposed method performs considerably better. In this experiment, the *HYBRID* method has the smallest error variance and equal accuracy to that of *pPSO*.

We again tested the performance of *HYBRID* in the case that it is fed with the manually estimated human body model for that subject. It turns out that mean Δ is 37.9, the standard deviation of Δ is 22.9 and the accuracy is 98.2%. Thus, *HYBRID* outperforms *pPSO* in that case, showing that, given accurate models, the combination of the bottom-up *OpenNI* method with the top-down *pPSO* method improves the tracking performance.

Finally, Table 4 summarizes all performed experiments. It can be verified that *pPSO* achieves a significant reduction in mean error and error variance compared to the rest of the methods as well as a significant increase in accuracy. If *HYBRID* is employed on the manually estimated human body models, then mean Δ becomes 39.7, the standard deviation of Δ becomes 28.4

Method	Mean Δ	Std. Δ	Accuracy (%)
OpenNI	54.5	46.2	86.3
bPSO	50.6	48.8	89.5
pPSO	**39.3**	**27.3**	**96.3**
HYBRID	44.5	35.4	94.2

Table 4: Aggregate comparison of *pPSO* and *HYBRID* with the baseline PSO method (*bPSO*) and the *OpenNI* method for all tested sequences.

and and the accuracy becomes 96.3%. So, its performance is considerably improved, but it does not outperform that of *pPSO*.

Figure 7 shows characteristics snapshots of the MHAD dataset and the skeletons that have been extracted by the *pPSO*, *OpenNI* and *bPSO* methods superimposed on the RGB frame of one of the two employed RGBD sensors. Finally, Figure 8 provides additional characteristic examples of the solutions provided by the *pPSO* method. A much more complete qualitative assessment of the performance of the proposed method can be performed based on the supplementary material accompanying this paper which is available at `http://youtu.be/n5irgHVuFwc`. It should be noted that no temporal smoothing has been performed between successive frames.

The proposed method runs on a computer equipped with a 8-core Intel i7 950 CPU, 4 GBs RAM. On this system, the average computing time for our non-optimized CPU-only implementation is 20 sec/frame. However, all involved computations are inherently data parallel and tailored for a GPU implementation. This is also evidenced by the real-time performance (20 fps) that is achievable by GPU implementations of similar approaches for the case of 3D hand tracking [15].

5. Discussion

We proposed a model-based method for tracking the articulated motion of the human body using a volumetric 3D representation that is built by fusing the depth measurements provided by two calibrated RGBD sensors. The proposed method follows a hypothesize-and-test approach that casts the articulated motion tracking problem into a search problem in a high-dimensional space. Searching is performed with a stochastic optimization technique, called PSO, resulting in a baseline implementation called *bPSO*. We also proposed a perturbation scheme that is applied on top of the *bPSO* solutions that results in the *pPSO* method. Finally, in order to raise the practical difficulties the limitations of *pPSO* with respect to its need for tailored human models and initialization in the first frame, we proposed the *HYBRID* method that combines *pPSO* with *OpenNI*. A series of experiments performed on a ground-truth-annotated data set demonstrated quantitatively and qualitatively that *pPSO* outperforms in error and accuracy the rest of the methods. This is even more striking in the challenging cases where the body configuration exhibits significant self occlusions. Thus, in situations where small error and high accuracy is more important that the burden and the overhead of using a second RGBD sensor, the proposed *pPSO* marker-less human articulations tracking method constitutes an attractive approach. *HYBRID* performs worse than *pPSO* because of the less accurate (yet automatic) estimation of the human body models. Still, the fact that *HYBRID* is fully automatic, is a significant advantage that, depending on application, might be more important than its lacking accuracy. In fact, as demonstrated experimentally, if *HYBRID* is given the chance to operate on accurate (non-automatically extracted) human body models, it performs comparably and, in some cases, better compared to *pPSO*.

Acknowledgements

This work was partially funded by the European Commission under contract FP7-IST-288146 HOBBIT and by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Project: THALIS-UOA- ERASITECHNIS MIS 375435.

[1] Bisacco, A., Ming-Hsuan, Y., Soatto, S., 2007. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In: IEEE Computer Vision and Pattern Recognition.

[2] Chen, L., Wei, H., Ferryman, J., 2013. A survey of human motion analysis using depth imagery. Pattern Recognition Letters 34 (15), 1995 – 2006.

[3] Corazza, S., Mundermann, L., Gambaretto, E., Ferrigno, G., Andriacchi, T., 2010. Markerless motion capture through visual hull, articulated icp and subject specific model generation. International Journal of Computer Vision 87 (1-2), 156–169.

[4] Deutscher, J., Reid, I., 2005. Articulated body motion capture by stochastic search. International Journal of Computer Vision 61 (2), 185–205.

[5] Gall, J., Rosenhahn, B., Brox, T., Seidel, H.-P., 2010. Optimization and filtering for human motion capture. International Journal of Computer Vision 87 (1-2), 75–92.

[6] Gall, J., Stoll, C., de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H. P., 2009. Motion capture using joint skeleton tracking and surface

estimation. In: IEEE Computer Vision and Pattern Recognition. pp. 1746–1753.

[7] Hamer, H., Schindler, K., Koller-Meier, E., Van Gool, L., 2009. Tracking a Hand Manipulating an Object. In: IEEE International Conference on Computer Vision.

[8] Kennedy, J., Eberhart, R., Jan. 1995. Particle Swarm Optimization. In: International Conference on Neural Networks. Vol. 4. IEEE, pp. 1942–1948.

[9] Kennedy, J., Eberhart, R., Yuhui, S., 2001. Swarm intelligence. Morgan Kaufmann.

[10] Mikic, I., Trivedi, M., Hunter, E., Cosman, P., 2003. Human body model acquisition and tracking using voxel data. International Journal of Computer Vision 53 (3), 199–223.

[11] Moeslund, T. B., Hilton, A., Kru, V., 2006. A Survey of Advances in Vision-based Human Motion Capture and Analysis. Computer Vision and Image Understanding 104, 90–126.

[12] Mussi, L., Ivekovic, S., Cagnoni, S., 2010. Markerless articulated human body tracking from multi-view video with gpu-pso. In: Tempesti, G., Tyrrell, A., Miller, J. (Eds.), Evolvable Systems: From Biology to Hardware. Vol. 6274 of Lecture Notes in Computer Science. pp. 97–108.

[13] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R., 2013. Berkeley MHAD: A Comprehensive Multimodal Human Action Database. In: IEEE Workshop on Applications on Computer Vision (WACV).

[14] Oikonomidis, I., Kyriazis, N., Argyros, A. A., 2010. Markerless and Efficient 26-DOF Hand Pose Recovery. Asian Conference on Computer Vision, 744–757.

[15] Oikonomidis, I., Kyriazis, N., Argyros, A. A., 2011. Efficient Model-based 3D Tracking of Hand Articulations using Kinect. In: British Machine Vision Conference. Dundee, UK.

[16] Oikonomidis, I., Kyriazis, N., Argyros, A. A., 2011. Full DOF Tracking of a Hand Interacting with an Object by Modeling Occlusions and Physical Constraints. In: IEEE International Conference on Computer Vision.

[17] OpenNI, November 2010. OpenNI User Guide. OpenNI organization, last viewed 19-01-2011 11:32.
URL http://www.openni.org/documentation

[18] Pons-Moll, G., Leal-Taixe, L., Truong, T., Rosenhahn, B., 2011. Efficient and robust shape matching for model based human motion capture. In: Mester, R., Felsberg, M. (Eds.), Pattern Recognition. Vol. 6835 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 416–425.

[19] Poppe, R., 2007. Vision-based human motion analysis: An overview. Computer VIsion and Image Understanding 108 (1-2), 4 – 18, special Issue on Vision for Human-Computer Interaction.

[20] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-Time Human Pose Recognition in Parts from Single Depth Images.

[21] Sigal, L., Isard, M., Haussecker, H., Black, M., 2012. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. International Journal of Computer Vision 98 (1), 15–48.

[22] Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D., 2005. Discriminative density propagation for 3d human motion estimation. In: IEEE Computer Vision and Pattern Recognition. Vol. 1. pp. 390–397 vol. 1.

[23] Smisek, J., Jancosek, M., Pajdla, T., 2011. 3d with kinect. In: IEEE ICCV Workshops. pp. 1154–1160.

[24] Tzevanidis, K., Zabulis, X., Sarmis, T., Koutlemanis, P., Kyriazis, N., Argyros, A., 2010. From multiple views to textured 3d meshes: a gpu-powered approach. In: ECCV Workshops 2010. pp. 5–11.

[25] Vicon, 2013. Vicon: Motion capture systems.
URL `http://www.vicon.com`

[26] Vijay, J., Trucco, E., Ivekovic, S., 2010. Markerless human articulated tracking using hierarchical particle swarm optimisation. Image and Vision Computing 28 (11), 1530–1547.

[27] Wilson, J. L., 2010. Microsoft kinect for xbox 360. PC Magazine Communications.

[28] Zhang, L., Sturm, J., Cremers, D., Lee, D., Oct. 2012. Real-time human motion tracking using multiple depth cameras. In: Proceedings of the International Conference on Intelligent Robot Systems (IROS).

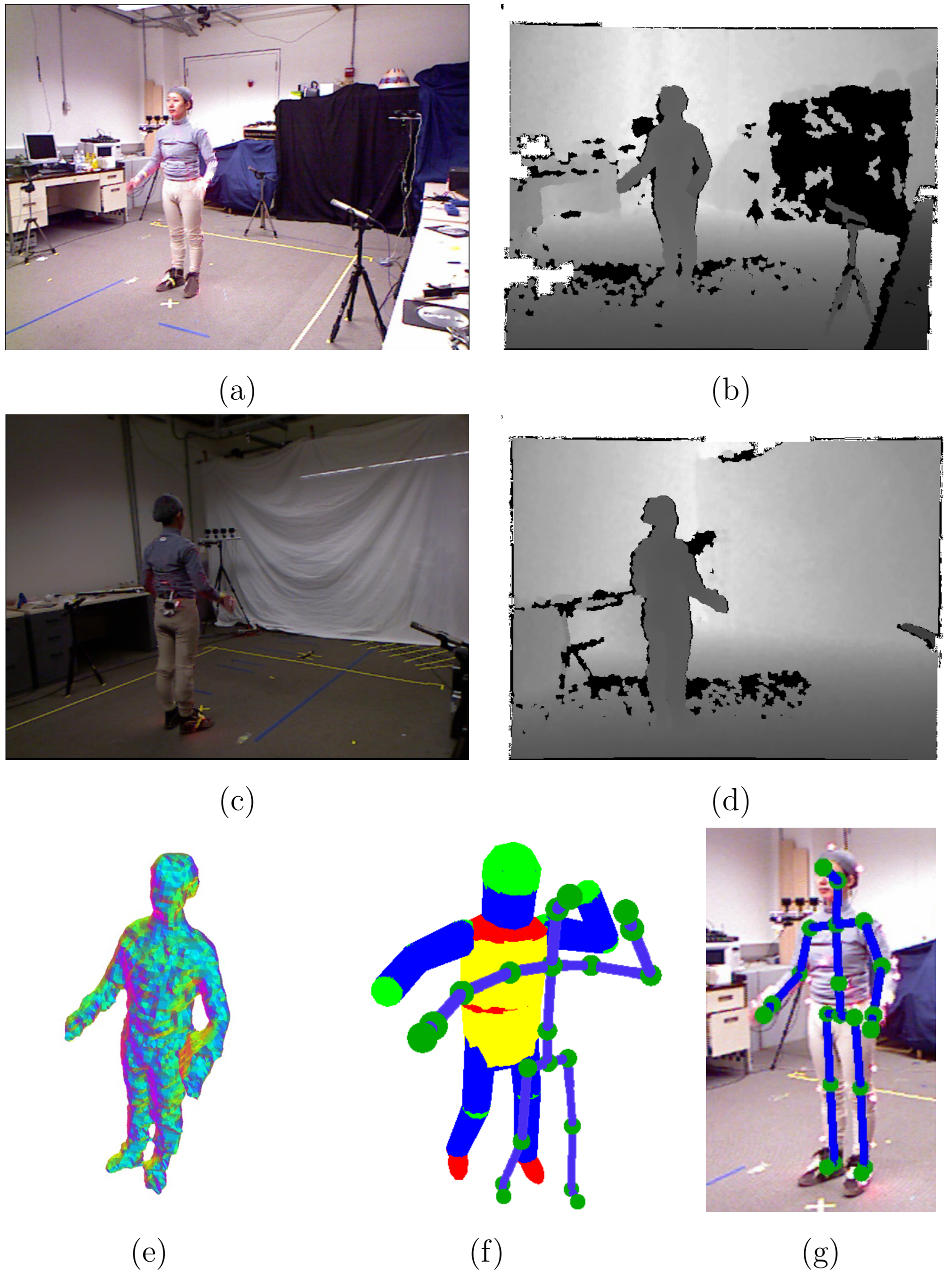


Figure 1: Graphical illustration of the proposed method. Two RGB frames ((a), (c)) and the corresponding depth maps ((b), (d)). The volume (e) occupied by the person is reconstructed using the depth maps. The proposed method fits the employed human body model (f) to this volume, recovering the body articulation (g).

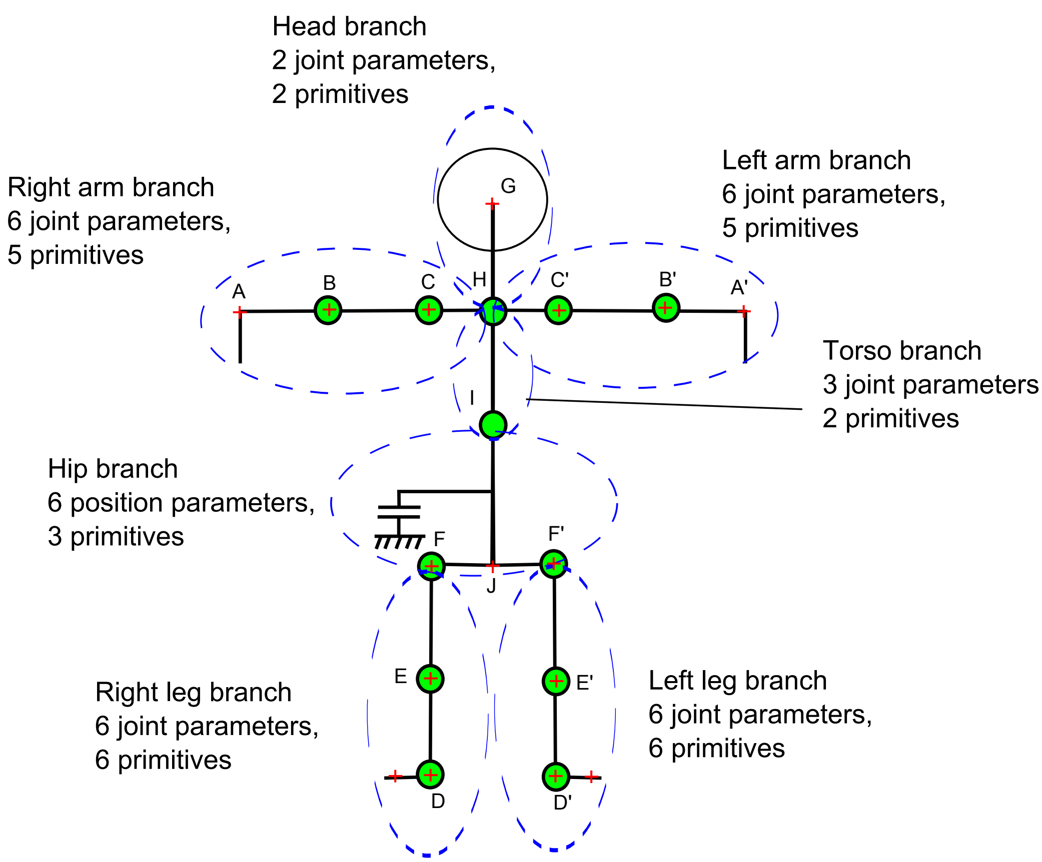


Figure 2: Definition of human body branches. The perturbation of the torso and hip branches results in changes in the parameters of their child branches. Model points with a red "x" denote joints whose 3D position is taken into account in defining the tracking error in the quantitative experimental evaluation of the method.

Figure 3: The twelve subjects of the MHAD dataset.

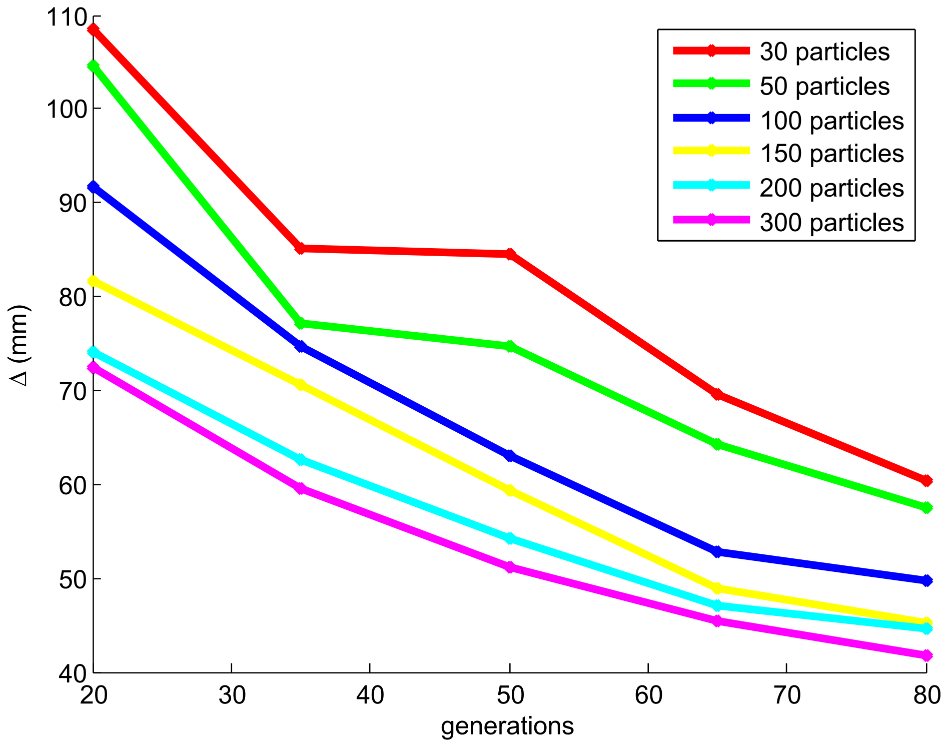
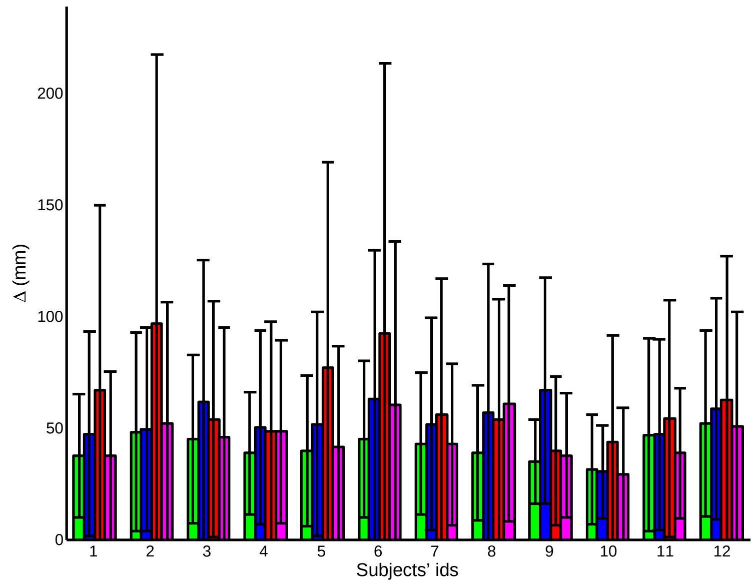
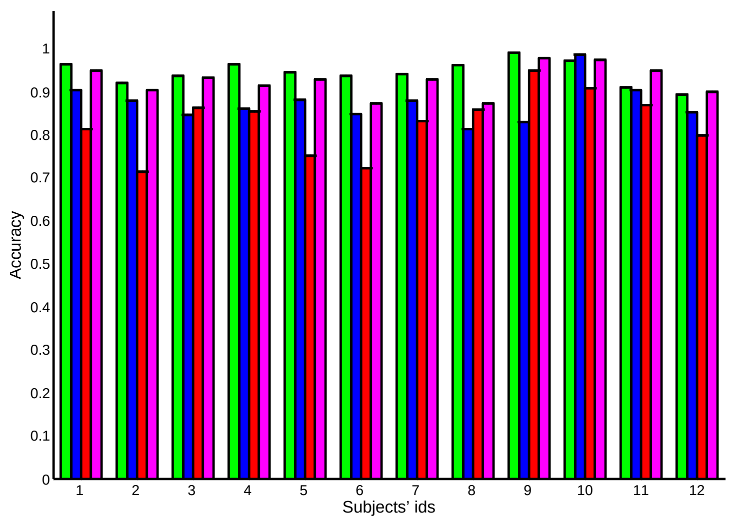


Figure 4: Quantitative evaluation of the performance of the method with respect to the *pPSO* parameters.

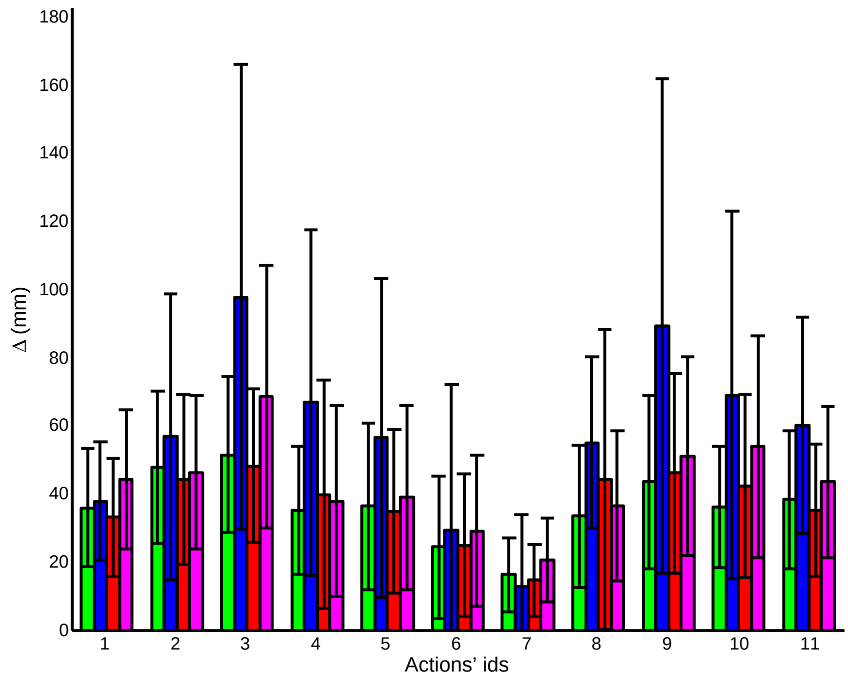


(a)

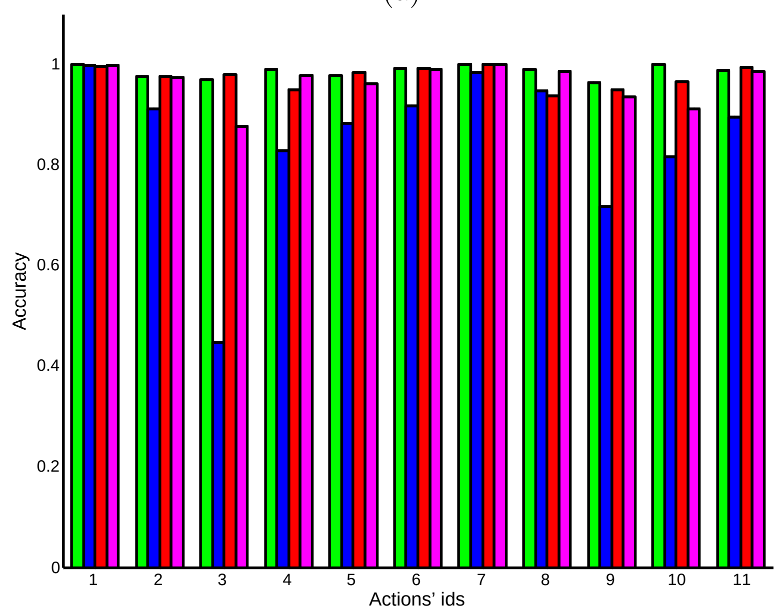


(b)

Figure 5: Quantitative evaluation of the method applied to 12 subjects performing the same action (boxing). (a) Error Δ and variances, (b) accuracy for the proposed method (*pPSO*, green bars), baseline method (*bPSO*, red bars), *OpenNI* human skeleton estimation (blue bars) and *HYBRID* (purple bars).



(a)



(b)

Figure 6: Quantitative evaluation of the method applied to 11 actions performed by the same person. (a) Error Δ and variances, (b) accuracy for the proposed method (green bars) *OpenNI* human skeleton estimation (blue bars) and *HYBRID* method (purple bars).

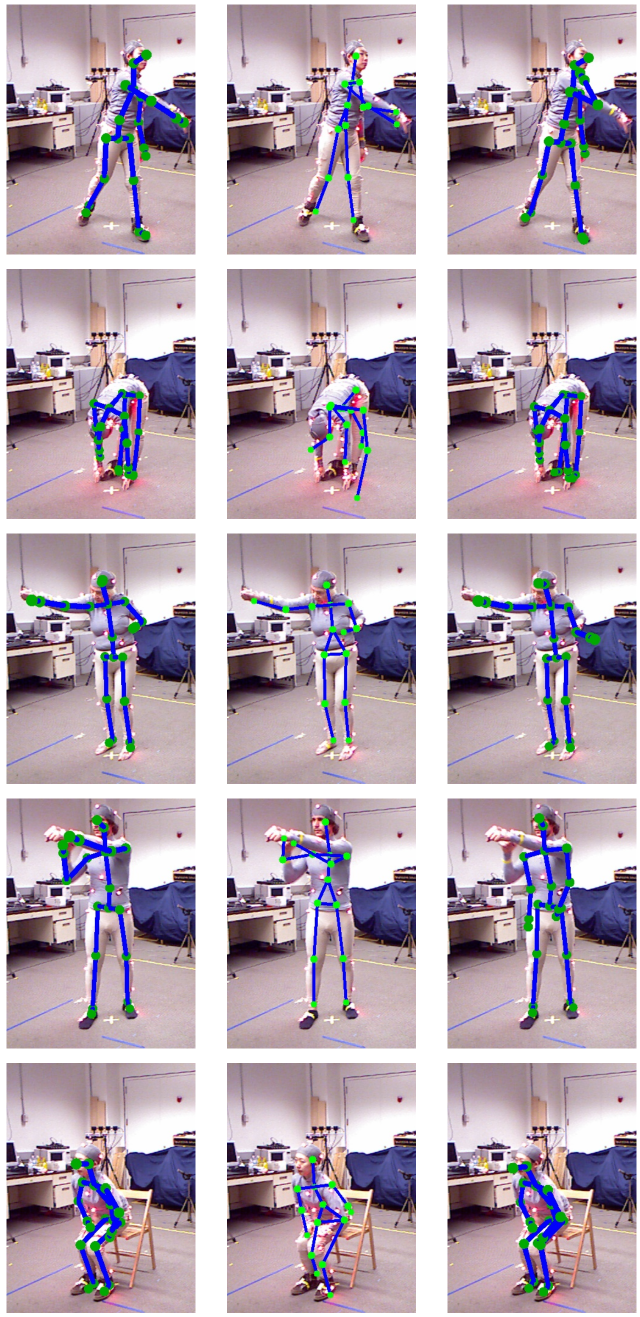


Figure 7: Qualitative comparison of the *pPSO* (left), OpenNI (middle) and *bPSO* (right) methods based on characteristic frames of the MHAD dataset.

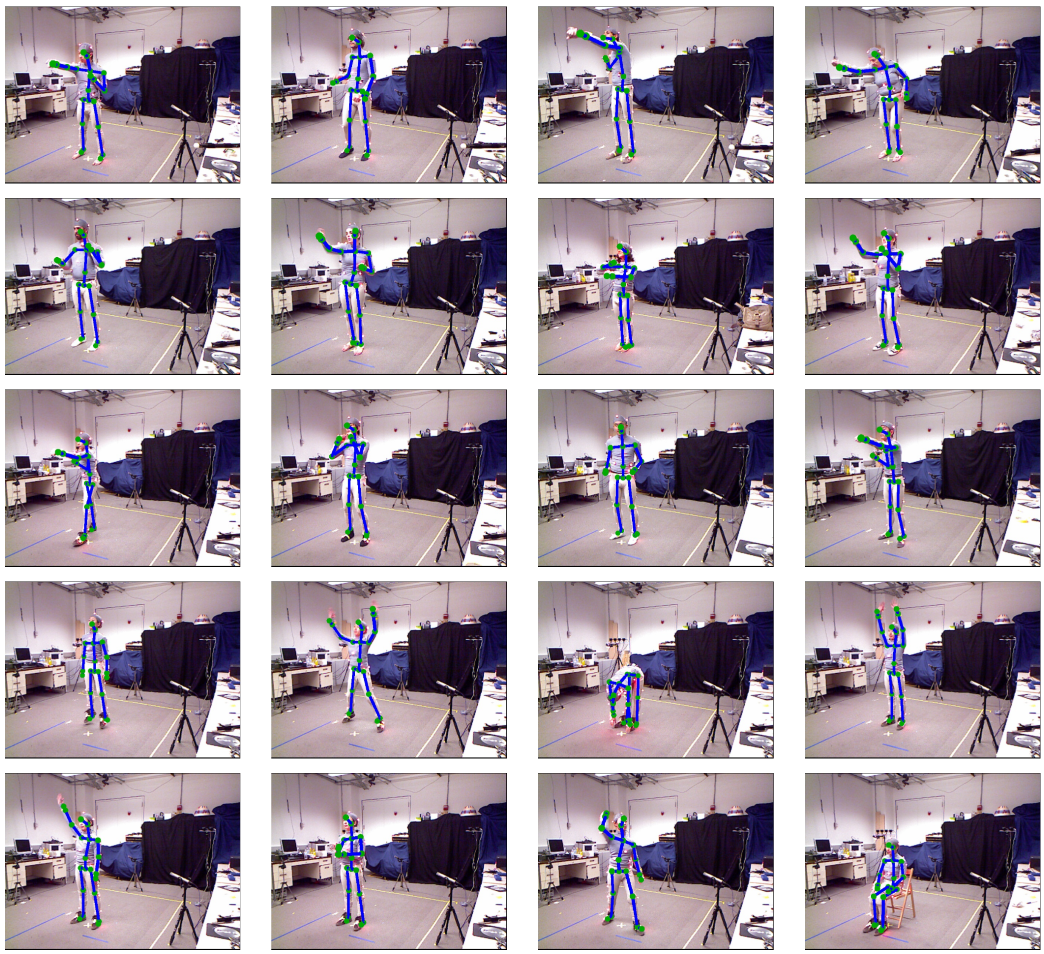


Figure 8: Various configurations on different subjects evaluated by the method.