Markerless 3D Human Pose Estimation and Tracking based on RGBD Cameras: an Experimental Evaluation

Damien Michel
ICS-FORTH
N. Plastira 100, Vas. Vouton
Heraklion, Crete, Greece
michel@ics.forth.gr

Ammar Quammaz
CSD/UoC and ICS-FORTH
Vassilika Vouton
Heraklion, Crete, Greece
ammarkov@ics.forth.gr

Antonis Argyros
CSD/UoC and ICS-FORTH
Vassilika Vouton
Heraklion, Crete, Greece
argyros@ics.forth.gr

ABSTRACT

We present a comparative experimental evaluation of three methods that estimate the 3D position, orientation and articulation of the human body from markerless visual observations obtained by RGBD cameras. The evaluated methods are representatives of three broad 3D human pose estimation/tracking methods. Specifically, the first is the discriminative approach adopted by OpenNI. The second is a hybrid approach that depends on the input of two synchronized and extrinsically calibrated RGBD cameras. Finally, the third one is a recently developed generative method that depends on input provided by a single RGBD camera. The experimental evaluation of these methods has been based on a publicly available data set that is annotated with ground truth. The obtained results expose the characteristics of the three methods and provide evidence that can guide the selection of the most appropriate one depending on the requirements of a certain application domain.

Keywords

Human body tracking; articulated motion tracking; human skeleton tracking; 3D human pose estimation

1. INTRODUCTION

The estimation of the articulated motion of the human body is very important to a number of real world applications, including but not limited to surveillance, gaming, medical rehabilitation, human-robot interaction, smart environments and many others. It is considered to be a challenging problem because of its high dimensionality, the variability of the tracked persons regarding their appearance and sizes, the spatially and temporally extended (self)-occlusions, etc. A number of practical approaches simplify or even avoid these problems by using special hardware that is placed on the environment and/or markers/full body suits worn by the persons to be tracked. However, these are invasive solutions. Uunobtrusive, markerless tracking is definitely preferable since it does not interfere with the environment, the subject and the performed actions.

The methods that use markerless visual data as their only input fall into three basic categories, the generative, the discriminative and the hybrid ones. Each class of methods has its own characteristics, advantages and disadvantages. Discriminative methods are fast, but rely on a discrete set of training poses whose selection determines the accuracy of the obtained results. Typically, they operate as single frame pose estimation methods, so they do not need to be initialized and they do not drift. The generative approaches provide accurate, physically plausible solutions, typically at a high computational cost. They also require initialization for the first frame, and may suffer from drift and tracking failures that are often irrecoverable. Hybrid methods integrate generative and discriminative components towards combining the advantages of both worlds.

In this work, we evaluate three approaches for 3D human pose estimation, one representative of each class. The evaluation has been performed on a dataset annotated with ground truth. The obtained quantitative experimental results help in assessing the relative performance of these methods and in deciding which is preferable in which situation. Qualitative results are also presented, based on a scenario that involves 3D human motion tracking to support the teleoperation of a NAO humanoid robot.

The rest of the paper is organized as follows. Section 2 reviews existing approaches to the problem of markerless 3D human pose estimation and tracking. Section 3 provides the articulated model of the human that was employed in the experiments and was used to evaluate the performance of the 3D human pose estimation methods. Section 4 presents the dataset on which this evaluation has been performed as well as the evaluation metrics. Section 5 presents the performed experiments and discusses the obtained results. Finally, Section 6 summarizes the main conclusions of this study and outlines possible extensions.

2. RELATED WORK

Because of its high theoretical and practical interest, vision-based human motion capture has been the theme of several research efforts. The complete review of the relevant works is beyond the scope of this paper. The interested reader is referred to [Moeslund et al. 2006, Poppe 2007] where extended surveys are provided. More recently, Chen et al. [Chen et al. 2013] surveyed methods for human motion estimation based on depth cameras.

Most commercial solutions to the problem of human mo-

PETRA '17, June 21 - 23, 2017, Island of Rhodes, Greece

ISBN 978-1-4503-5227-7/17/06…$15.00

DOI: http://dx.doi.org/10.1145/3056540.3056543

tion capture make use of special markers that are placed on carefully selected points of the subject's body (e.g., joints). Nevertheless, markerless motion capture techniques are, by far, more interesting and preferable. Being unobtrusive, they present important practical advantages over the marker-based solutions, such as lower setup cost and complexity, no interference with the performed actions, etc.

Markerless human motion capture techniques may be classified into three classes, the *discriminative* (named bottom-up in [Michel et al. 2015]), the *generative* (named top-down, in [Michel et al. 2015]) and the hybrid ones. Discriminative methods [Sminchisescu et al. 2005, Bisacco et al. 2007, Shotton et al. 2011, Pons-Moll et al. 2011, Sigal et al. 2012] 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 space of possible human poses, or the part of it that is relevant to the application of interest. The type of descriptors employed, the mapping method and the actual poses database are the factors determining the accuracy and efficiency of these methods. Recent approaches based on CNNs have produced very promising results [Rogez and Schmid 2016, Yasin et al. 2016, Li et al. 2015]. Due to their nature, most of their computing time is spend on the offline processes of database creation and mapping, while the online computational performance is rather good. An advantage of the discriminative approaches is that they perform single frame pose estimation and they do not rely on temporal continuity. Thus, they do not require initialization and they do not suffer from drift.

Generative approaches [Deutscher and Reid 2005, Gall et al. 2009, Gall et al. 2010, Vijay et al. 2010, Corazza et al. 2010, Zhang et al. 2012, Michel et al. 2015] use an 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 [Gall et al. 2009]. Having defined a model of the human body, different pose hypotheses can be formed. A typical generative method consists of generating hypotheses and comparing them to the input visual data. The comparison is performed based on an objective function that quantifies 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 the high dimensional space of human poses. Kinematic constrains based on physiological data are often applied to the model, excluding non realistic poses and reducing that search space. Constraining not only the pose but also the motion itself can further help reducing the complexity, for example with Kalman filters [Mikic et al. 2003]. However, this means a reduced generality and the necessity to build and learn human motion models. 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 final solution or they further combine them into a full model [Shotton et al. 2011, Sigal et al. 2012]. As in the case of hand tracking and according to the related categorization of Oikonomidis et al. [Oikonomidis et al. 2011], we can identify *disjoint evidence methods* and *joint evidence methods* [Deutscher and Reid 2005, Gall et al. 2009, Gall et al. 2010, Vijay et al. 2010, Corazza et al. 2010, Zhang et al. 2012]. Joint evidence methods handle effortlessly collisions, self occlusions and all part interactions while disjoints evidence methods have to handle them explicitly. The main advantage of generative 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 of generative methods is that typically, they require knowledge of the body model parameters of the individual to be tracked and they must be initialized for the first frame of a sequence.

Characteristics	**OpenNI**	**HYBRID**	**FHBT**
Method type	Discriminative	Hybrid	Generative
Number of cameras	1	2	1
Auto-initialization	Yes, special pose	Yes, special pose	Yes, any pose
Initialization speed	Slow (>3 sec)	Slow (3 sec)	Instant (0.03 sec)
Auto recovery from failures	Yes	Yes	Yes
Handles various body types	Yes	Yes	Yes
Handles occlusions	No	No	Yes
Moving camera(s)	No	No	Yes
Ensure physical plausibility	No	Yes	No
Mode of operation	Online	Offline	Online
Real time performance	Yes	No	Yes

Table 1: Overview of the evaluated methods (*OpenNI* [OpenNI 2010], *HYBRID* [Michel et al. 2015], *FHBT* [Michel and Argyros 2016]) with respect to a number of key characteristics and properties.

Finally, hybrid methods have been proposed [Michel et al. 2015] that combine the benefits of the discriminative and generative approaches. The basic idea is to use a discriminative part that provides a rough 3D human pose for every frame, which is then refined based on a generative component. This way, hybrid methods manage to achieve the accuracy of the generative ones without need for initialization and with robustness to tracking failures.

2.1 The evaluated methods

This paper presents a comparison between three methods that fall within different categories of the classification scheme presented above.

The *OpenNI* method [OpenNI 2010]: This is a widely employed, purely discriminative method. It is applied on the input of a single RGBD camera.

The *HYBRID* method [Michel et al. 2015]: As a hybrid method[1], it consists of a discriminative and a generative component. The generative, joint evidence component of the

[1] See also https://youtu.be/n5irgHVuFwc

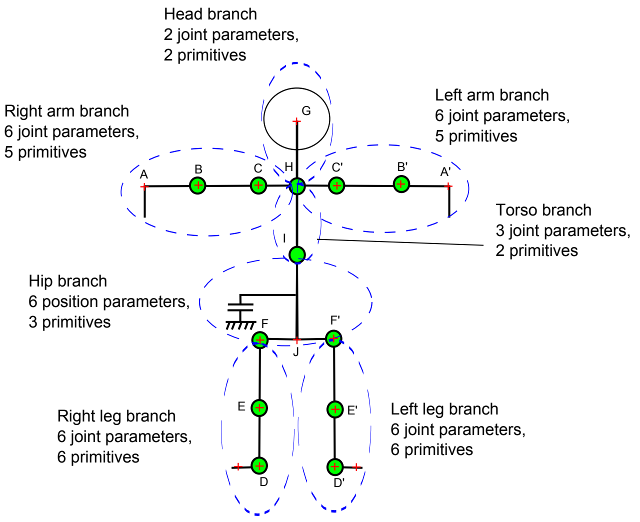


Figure 1: The employed human body model. Model points with a red "+" denote joints whose 3D position is taken into account in defining the tracking error.

method requires input from two extrinsically synchronized RGB-D cameras that is used to reconstruct the 3D volume occupied by the human body. Then, human pose estimation is formulated as an 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. To track the human pose, solutions for a certain frame are initialized at the vicinity of solutions estimated at the previous frame. However, the solution suggested by the discriminative component (*OpenNI*) of the method is also considered as a human pose hypothesis to (a) adjust the human model parameters to the tracked individual and (b) safeguard from abrupt human motions as well as from tracking failures.

The *FHBT* method [Michel and Argyros 2016]**:** This is a generative, disjoint evidence method that performs tracking by-detection. Body parts are identified independently and then assembled together in a complete model[2]

The main characteristics of the evaluated methods are summarized in Table 1. For more details, the reader is referred to the corresponding references. In general, *OpenNI* is a flexible and fast method. As suggested by the extensive evaluation performed in this paper, its accuracy is moderate. One of its main drawbacks is its long initialization time. The *HYBRID* method inherits the long initialization time from *OpenNI*. Moreover, it relies on a more complex setup of two extrinsically calibrated RGBD sensors and its computational requirements are quite high, i.e., near-real-time performance can only be achieved with an elaborate GPU-based implementation on a high-end computer featuring a state of the art graphics card. The experimental results demonstrate that in return, the *HYBRID* method outperforms the two others in terms of accuracy. The *FHBT* method is less accurate than *HYBRID* but on par with *OpenNI*. This fact, together with other advantages of the method (see Table 1), make it an attractive solution to a number of applications that require knowledge of the human body pose.

Figure 2: The twelve subjects of the MHAD dataset.

3. HUMAN BODY MODEL

The employed articulated model of the human consists of a main body, two legs, two arms and the head (Figure 1). 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.

On top of the 35 mobilities of this model, 9 parameters control the lengths of certain human body parts. 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 for these parameters for the subjects of the employed dataset. The parenthesis next to a parameter name refers to the corresponding body segment(s) in Figure 1.

It has to be noted that individual methods employ their own, internal models for 3D human pose estimation. The model described above and illustrated in Figure 1 is used for the evaluation of the performance of the benchmarked methods, only. Thus, the relation of the above model to the ones used internally by each method has been established and used to bring all results to the same reference frame so as to enable their direct comparison.

4. DATASET & EVALUATION METRICS

We provide information on the dataset that was employed for the evaluation of the 3D human pose estimation methods

[2] See also https://youtu.be/ZKlC9PA1IDg

Subject	S01			S02			S03			S04			S05			S06		
Metric	G	O	F	G	O	F	G	O	F	G	O	F	G	O	F	G	O	F
UBL (HI)	26	19	22	30	21	24	33	21	23	29	20	22	32	22	23	28	20	24
LBL (IJ)	15	19	22	17	21	24	18	21	23	17	20	22	19	22	23	17	20	24
SND (CH, C'H)	19	15	17	19	15	17	17	14	17	15	15	17	17	16	17	19	14	17
HND (GH)	20	25	23	20	25	23	20	25	23	20	21	21	20	26	26	20	20	22
LHD (FJ, F'J)	10	9	10	11	9	10	10	8	10	9	9	10	9	9	10	9	8	10
BAL (BC, B'C')	24	25	28	28	27	31	31	28	33	24	23	25	26	28	33	26	26	30
FAL (AB, A'B')	23	26	21	25	31	29	26	32	29	24	25	23	25	31	28	24	27	25
BLL (EF, E'F')	36	41	39	43	47	41	44	47	41	37	39	39	42	45	40	42	44	42
FLL (DE, D'E')	42	37	39	48	42	41	47	43	41	41	35	40	45	42	40	45	41	41

(a)

Subject	S07			S08			S09			S10			S11			S12		
Metric	G	O	F	G	O	F	G	O	F	G	O	F	G	O	F	G	O	F
UBL (HI)	25	20	21	30	20	24	27	20	21	27	21	23	28	20	23	24	20	23
LBL (IJ)	15	20	21	18	20	23	15	20	21	16	21	23	16	20	23	21	20	23
SND (CH, C'H)	17	17	17	18	15	16	15	13	17	17	14	17	17	15	17	18	14	17
HND (GH)	20	20	21	20	21	20	20	17	23	20	24	21	20	25	22	20	24	22
LHD (FJ, F'J)	8	7	10	9	9	10	8	7	10	8	8	10	8	9	10	9	8	10
BAL (BC, B'C')	22	25	26	24	26	28	23	22	24	27	25	30	26	27	31	25	25	27
FAL (AB, A'B')	22	24	23	24	26	24	23	27	22	24	29	24	24	28	24	22	27	24
BLL (EF, E'F')	35	39	38	39	40	41	35	42	38	41	43	40	41	44	41	38	43	41
FLL (DE, D'E')	41	35	37	43	39	41	41	37	40	43	40	40	44	41	41	41	39	41

(b)

Table 2: Body part lengths (in cm) for the human subjects of the MHAD dataset, (a) subjects 01-06, (b) subjects 07-12. Columns (G) are the manually measured, ground truth values, columns (O) the one estimated by the *OpenNI* method, and columns (F) are the ones estimated by the *FHBT* method. The parenthesis next to the name of each body part to the corresponding body segment(s) in Figure 1 (see also Section 3).

as well as on the criteria used to quantify their performance.

4.1 The MHAD data set

The comparative evaluation and comparison of the three human pose estimation/tracking methods was based on the Berkeley Multimodal Human Action Database (MHAD) [Ofli et al. 2013]. This dataset features 12 human subjects (see Figure 2). From this figure it can be verified that the MHAD data set involves subjects of considerable variability with respect to age, size and body types. This is also shown quantitatively in Table 2, columns (G), which provide the lengths of body parts for all the subjects.

The subjects 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. 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 methods employ the RGBD feeds (both of them for the *HYBRID* method and the same, single feed for the *FHBT* and *OpenNI* methods). The resulting tracking results are compared against the ground truth resulting from the motion capture data.

4.2 Evaluation metrics

To quantify the accuracy in body pose estimation, we adopt the metric used in [Hamer et al. 2009]. More specifically, the Euclidean distance between a set of corresponding 3D points (skeleton joints) 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 with a red cross in Figure 1. The average of all these distances over all the frames of the sequence constitutes the resulting error estimate Δ.

Another metric reports the percentage $A(t)$ of these distances that are within some predefined threshold t for a certain sequence. We will refer to this metric as the accuracy in human body pose estimation. For example, an accuracy of $A(10) = 70\%$ for a sequence means that in the frames of that sequence, 70% of the joints have been estimated within $10cm$ from the ground truth.

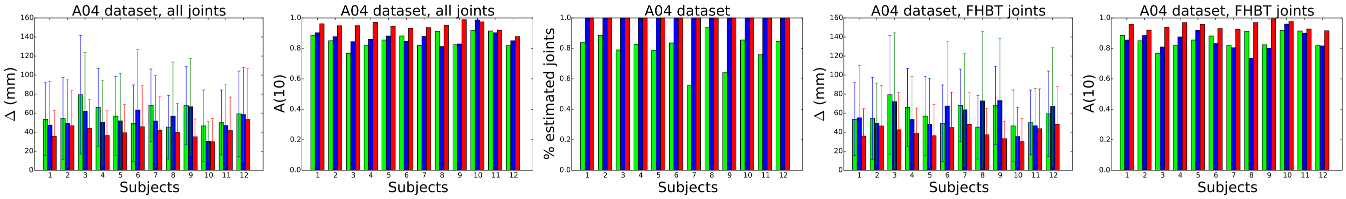
5. COMPARATIVE EVALUATION

Several experiments were carried out to assess quantitatively and qualitatively the accuracy and the performance of the evaluated human articulation tracking methods.

5.1 All-subjects-one-action experiment

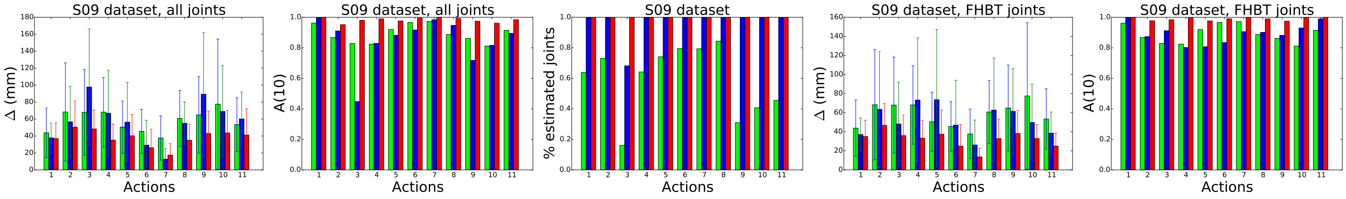
A first experiment aimed at evaluating the performance of the methods across different human subjects. All twelve sequences showing the twelve different subjects performing the same activity (activity 04, boxing) were considered.

Figure 3(a), (b) illustrate the error Δ and the accuracy $A(10cm)$, respectively, of the *FHBT*, *HYBRID* and *OpenNI* methods. These measures have been estimated over all the joints of the human model that a particular method estimates. It can be verified that the *HYBRID* method outperforms the *OpenNI* and *FHBT* methods which perform comparably. This is expected, given the fact that *HYBRID* uses



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

Figure 3: Quantitative evaluation of the method applied to 12 subjects performing the same action (boxing). (a) Error Δ and variances over all frames and joints. (b) Accuracy *A*(10*cm*) over all frames and joints. (c) Percentage of joints for which a method provided an estimation. (d), (e): Error Δ and accuracy *A*(10*cm*) over the joints for which *FHBT* provided an estimation. *FHBT*: green bars, *HYBRID*: red bars, *OpenNI*: blue bars.



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

Figure 4: Quantitative evaluation of the method applied to 11 actions performed by the same subject (s09). (a) Error Δ and variances over all frames and joints. (b) Accuracy *A*(10*cm*) over all frames and joints. (c) Percentage of joints for which a method provided an estimation. (d), (e): Error Δ and accuracy *A*(10*cm*) over the joints for which *FHBT* provided an estimation. *FHBT*: green bars, *HYBRID*: red bars, *OpenNI*: blue bars.

more information than the two other methods (two RGBD feeds as opposed to only one).

It has to be noted that the *FHBT* method estimates only a subset of the joints, depending on whether the method's confidence on them exceeds an internally set threshold. Figure 3(c) shows the percentage of joints that were estimated by each method. For the *OpenNI* and *HYBRID* methods this is always 100% while for the *FHBT* method this is 75% on average, across different subjects. In a subsequent measurement, we evaluated the error Δ and the accuracy $A(10cm)$ (Figures 3(d), (e), respectively) for all methods, but only over the joints and the frames for which the *FHBT* method provided some estimation. It can be verified that when error and accuracy is measured over these joints, the performance of the *OpenNI* and the *HYBRID* method increases. This indicates that the confidence that *FHBT* estimates for the joints is trustworthy.

5.2 All-actions-one-subject experiment

In a second 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 activities. Figure 4 illustrates the obtained results in a way analogous to that of Figure 3. Again, *HYBRID* outperforms the rest of the methods with respect to the mean error Δ and accuracy, while the rest two methods perform comparably. 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 *FHBT* method estimates the least number of joints (see Figure 4(c)).

5.3 Aggregated results

Table 3 summarizes the performed experiments by providing Δ, $A(10)$ numerical values for the cases of all-subjects-one-action and all-actions-one-subject experiments, as well as for the union of the corresponding datasets. A number of interesting conclusions can be drawn: (a) Overall, the *HYBRID* method is the one that results in the lowest errors and error variances and the highest accuracy, (b) *FHBT* and *OpenNI* perform comparably, (c) *FHBT* exhibits minimum performance variability between the two experiments with respect to Δ and its standard deviation, while *OpenNI* exhibits minimum performance variability with respect to $A(10)$ and, (d) *HYBRID* has the maximum variability in all metrics.

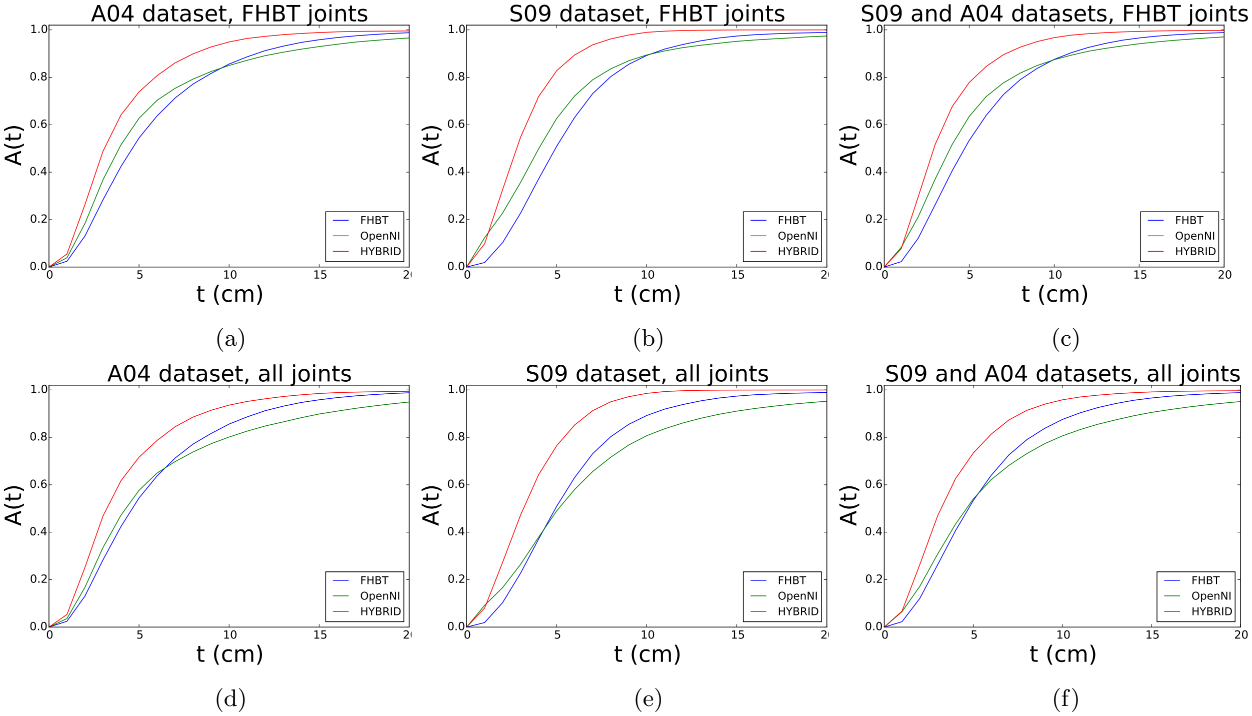
In the results of Figures 3 and 4, the accuracy $A(t)$ has been computed for $t = 10cm$. While this choice of t is compatible with the requirements of many applications, it is interesting to know how the accuracy of a certain method varies as a function of t. Figure 5 presents this information. For the three evaluated methods, we measure their accuracy for various values t in the range [0..20] cm. The top row of plots shows these results over the joints that the *FHBT* method estimated, while the bottom row shows the same results over all the joints. It can be verified that the *HYBRID* method is consistently more accurate compared to the other two, regardless of t. *FHBT* and *OpenNI* perform comparably. Moreover, the plots show for which error tolerances each method becomes preferable.

5.4 Estimation of body sizes

The *HYBRID* method relies on its discriminative part

Dataset	S09			A04			Aggregate		
Method	Mean Δ	Std. Δ	$A(10)$ (%)	Mean Δ	Std. Δ	$A(10)$ (%)	Mean Δ	Std. Δ	$A(10)$ (%)
FHBT	58.0/58.0	40.7/40.7	89.2/89.2	58.1/58.1	41.3/41.3	85.5/85.5	57.6/57.6	41.0/41.0	87.5/87.5
OpenNI	69.4/52.8	63.1/50.1	80.7/89.4	67.6/58.7	69.5/58.9	80.1/84.9	67.9/55.1	66.3/54.2	80.6/87.4
HYBRID	36.1/32.3	20.3/19.0	98.5/98.9	42.6/40.5	34.7/32.7	93.5/94.9	39.7/36.7	28.3/26.5	95.8/96.7

Table 3: Comparison of *FHBT*, *HYBRID* and *OpenNI* methods in all datasets. Mean Δ and std. of Δ are measured in mm. The two numbers in each slot of the matrix refer to the quantity measured over all joints/the quantity measured over the joints computed by *FHBT*.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5: The accuracy $A(t)$ of the evaluated methods as a function of t in the range $[0..20]$cm, for all experiments. Left column: all-subjects-one-action, middle column: all-actions-one-subject, right column: the union of the two datasets. Top row: results for the joints estimated by *FHBT*. Bottom row: results for the joints that each individual method estimates.

(which is the *OpenNI* method) in order to initialize the tracking process and to set the proper human body model parameters. The *FHBT* method has its own mechanism to provide an estimation of these parameters. Table 2 shows, for each subject, the ground truth information (columns (G)) as well as the ones estimated by the *OpenNI* (columns (O)) and *FHBT* (columns (F)) methods. It can be verified that the *FHBT* method is slightly more accurate in estimating body shape parameters than *OpenNI*. In particular, for each method, we computed the mean error (among subjects) in the estimation of each body part length. Then, we estimated the mean and the standard deviation of these errors for all body parts. The results show that the mean error in the estimation of the body parts for the *OpenNI* method is $3.44cm$ with a standard deviation of 0.68cm, while for the *FHBT* method we obtain a mean error of 2.86cm with a standard deviation of 0.71cm. The analysis also shows that the most inaccurate measurements are obtained for the human torso-related parts, while the lengths of the limb parts (arms, legs) are estimated more accurately.

5.5 Qualitative results

Figure 6 shows characteristic snapshots of the MHAD dataset and the skeletons that have been extracted by the *HYBRID*, *OpenNI* and *FHBT* methods superimposed on the RGB frame of one of the two employed RGBD sensors. It can be verified that the estimation performed by *HYBRID* (top row) is the most accurate one, followed by *OpenNI* (middle) and then by *FHBT* (bottom). Interestingly, the pose in the 2nd column cannot be estimated at all by *FHBT*. For the 5th pose, *FHBT* estimates only partial information.

Finally, Figure 7 provides representative snapshots from an experiment where *FHBT* was used to teleoperate a NAO humanoid robot through the tracking of the human body motion. A more complete view of the results are available at http://cvrlcode.ics.forth.gr/projects/fhbt/.

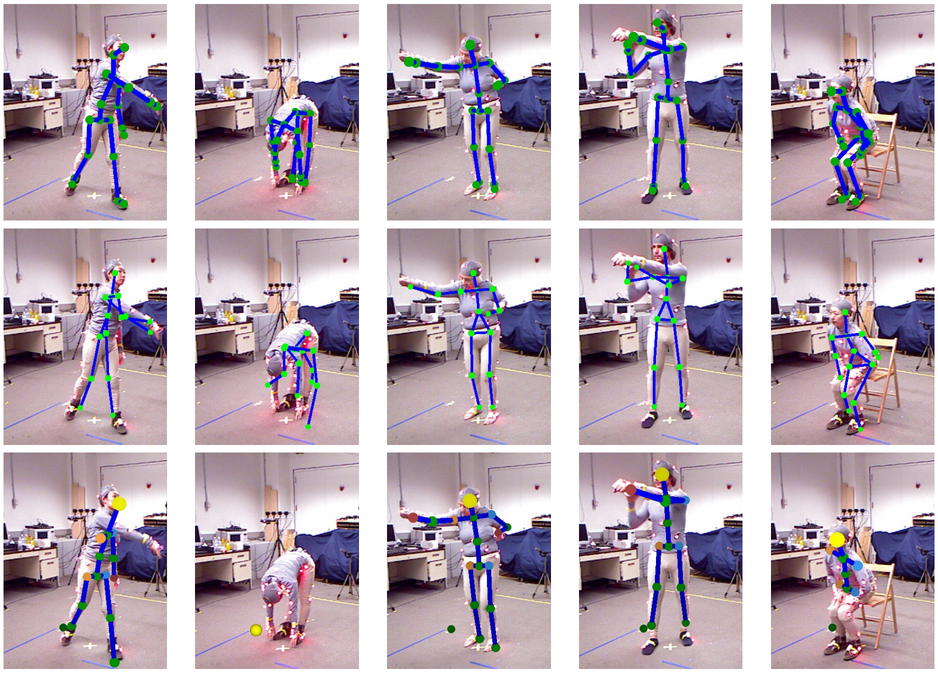


Figure 6: Qualitative comparison of the *HYBRID* (top), *OpenNI* (middle) and *FHBT* (bottom) methods based on frames of the MHAD dataset.

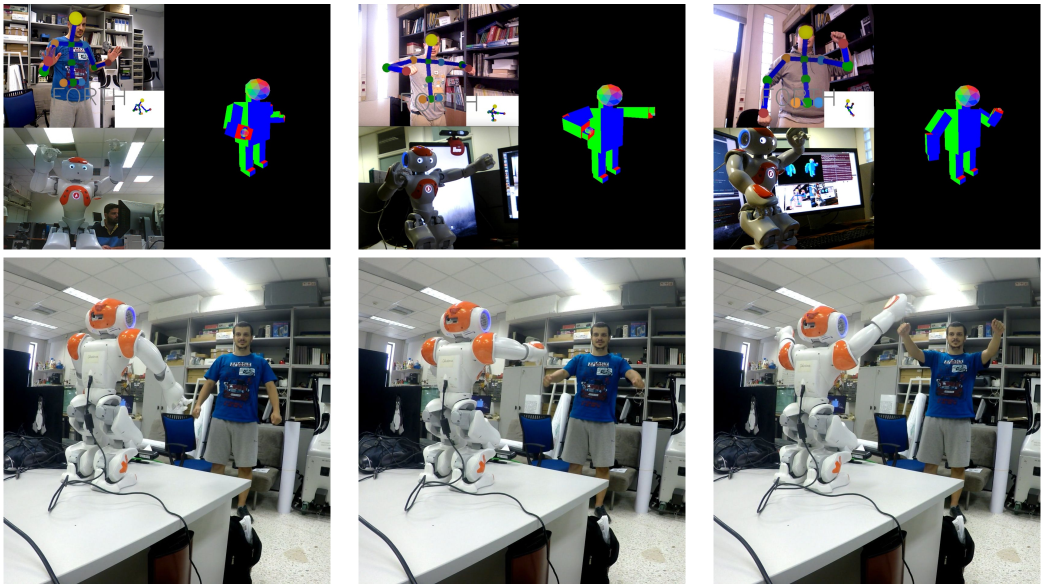


Figure 7: Snapshots from a *FHBT*-guided humanoid robot (NAO) teleoperation experiment. Top row: In each frame, we show the fitted human skeleton (top-left), the inferred NAO pose (right) and its actual robotic realization (bottom left). Notice that *FHBT* does not estimate the joints of the legs as they are not visible from the specific camera viewpoint. Bottom row: External views from the actual experiment.

6. DISCUSSION

We performed a comparative evaluation of three methods for the estimation of the articulated motion of the human body. A series of experiments performed on a ground-truth-annotated data set demonstrated quantitatively and qualitatively the performance of the evaluated methods. The results show that in situations where small error and high accuracy is more important that the burden and the overhead of using a second RGBD sensor, the *HYBRID* method is the preferred one. Interestingly, the *HYBRID* method is slightly less accurate than other purely generative methods like *pPSO* [Michel et al. 2015] that are aware of an accurate human body model. Still, the fact that *HYBRID* is fully automatic, is a significant advantage that, depending on application, might be more important than its lacking accuracy. Another result is that *FHBT* and *OpenNI* perform comparably. *FHBT* has some additional practical advantages that make it an attractive alternative for estimating human 3D pose. For example, it initializes instantly (in a single frame), can cope with partially visible human bodies and operates with a moving camera, even in jerky motion. It should be stressed that the employed MHAD dataset does not showcase such difficult situations which are, nevertheless, abundant in several real-life scenarios[3]. Future experimental work will address the quantification of the performance of *FHBT* in such scenarios that require 3D human pose estimation in the wild.

7. ACKNOWLEDGEMENTS

This work was partially supported by the EU FP7-ICT-2011-9 project WEARHAP and the H2020 project ACANTO.

8. REFERENCES

[Bisacco et al. 2007] A. Bisacco, Y. Ming-Hsuan, and S. Soatto. 2007. Fast Human Pose Estimation Using Appearance and Motion via Multi-dimensional Boosting Regression. In *IEEE CVPR*.

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

[Corazza et al. 2010] S. Corazza, L. Mundermann, E. Gambaretto, G. Ferrigno, and T. Andriacchi. 2010. Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation. *IJCV* 87, 1-2 (2010), 156–169.

[Deutscher and Reid 2005] J. Deutscher and I. Reid. 2005. Articulated Body Motion Capture by Stochastic Search. *IJCV* 61, 2 (2005), 185–205.

[Gall et al. 2010] J. Gall, B. Rosenhahn, T. Brox, and H-P. Seidel. 2010. Optimization and Filtering for Human Motion Capture. *IJCV* 87, 1-2 (2010), 75–92.

[Gall et al. 2009] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H. P Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *IEEE CVPR*. 1746–1753.

[Hamer et al. 2009] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. 2009. Tracking a Hand Manipulating an Object. In *IEEE ICCV*.

[Li et al. 2015] S. Li, W. Zhang, and A. Chan. 2015. Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation. In *IEEE ICCV*. 2848–2856.

[Michel and Argyros 2016] D. Michel and A. Argyros. 2016. Apparatuses, methods and systems for recovering a 3-dimensional skeletal model of the human body. (24 March 2016).

[Michel et al. 2015] D. Michel, C. Panagiotakis, and A. Argyros. 2015. Tracking the articulated motion of the human body with two RGBD cameras. *Machine Vision Applications* 26, 1 (2015), 41–54.

[Mikic et al. 2003] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. 2003. Human Body Model Acquisition and Tracking Using Voxel Data. *IJCV* 53, 3 (2003), 199–223.

[Moeslund et al. 2006] T. Moeslund, A. Hilton, and V. Kruger. 2006. A Survey of Advances in Vision-based Human Motion Capture and Analysis. *CVIU* 104 (2006), 90–126.

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

[Oikonomidis et al. 2011] I. Oikonomidis, N. Kyriazis, and A. Argyros. 2011. Full DOF Tracking of a Hand Interacting with an Object by Modeling Occlusions and Physical Constraints. In *IEEE ICCV*.

[OpenNI 2010] OpenNI. 2010. *OpenNI User Guide*. OpenNI organization.

[Pons-Moll et al. 2011] G. Pons-Moll, L. Leal-Taixe, T. Truong, and B. Rosenhahn. 2011. Efficient and Robust Shape Matching for Model Based Human Motion Capture. In *Pattern Recognition*, Rudolf Mester and Michael Felsberg (Eds.). LNCS, Vol. 6835. Springer, 416–425.

[Poppe 2007] R. Poppe. 2007. Vision-based human motion analysis: An overview. *CVIU* 108, 1-2 (2007), 4 – 18. Special Issue on Vision for Human-Computer Interaction.

[Rogez and Schmid 2016] G. Rogez and C. Schmid. 2016. MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. In *Advances in Neural Information Processing Systems*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 3108–3116.

[Shotton et al. 2011] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-Time Human Pose Recognition in Parts from Single Depth Images. In *IEEE CVPR*.

[Sigal et al. 2012] L. Sigal, M. Isard, H. Haussecker, and M. Black. 2012. Loose-limbed People: Estimating 3D Human Pose and Motion Using Non-parametric Belief Propagation. *IJCV* 98, 1 (2012), 15–48.

[Sminchisescu et al. 2005] C. Sminchisescu, A. Kanaujia, Zhiguo Li, and D. Metaxas. 2005. Discriminative density propagation for 3D human motion estimation. In *IEEE CVPR*, Vol. 1. 390–397.

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

[Yasin et al. 2016] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. 2016. A Dual-Source Approach for 3D Pose Estimation from a Single Image. In *IEEE CVPR*.

[Zhang et al. 2012] L. Zhang, J. Sturm, D. Cremers, and D. Lee. 2012. Real-Time Human Motion Tracking using Multiple Depth Cameras. In *Proceedings of the International Conference on Intelligent Robot Systems (IROS)*.

[3] See also https://youtu.be/ZKlC9PA1IDg