Principal Direction Analysis-based Real-time 3D Human Pose Reconstruction from a Single Depth Image

Dong-Luong Dinh
Dept. of Computer Engineering
Kyung Hee University
South Korea
luongdd@oslab.khu.ac.kr

Hee-Sok Han
Dept. of Biomedical Engineering
Kyung Hee University
South Korea
hshan@khu.ac.kr

Hyun Jae Jeon
Dept. of Biomedical Engineering
Kyung Hee University
South Korea
hjjeon1106@khu.ac.kr

Sungyoung Lee
Dept. of Computer Engineering
Kyung Hee University
South Korea
sylee@oslab.khu.ac.kr

Tae-Seong Kim
Dept. of Biomedical Engineering
Kyung Hee University
South Korea
tskim@khu.ac.kr

ABSTRACT

Human pose estimation in real-time is a challenging problem in computer vision. In this paper, we present a novel approach to recover a 3D human pose in real-time from a single depth human silhouette using Principal Direction Analysis (PDA) on each recognized body part. In our work, the human body parts are first recognized from a depth human body silhouette via the trained Random Forests (RFs). On each recognized body part which is presented as a set of 3D points cloud, PDA is applied to estimate the principal direction of the body part. Finally, a 3D human pose gets recovered by mapping the principal directional vector to each body part of a 3D human body model which is created with a set of super-quadrics linked by the kinematic chains. In our experiments, we have performed quantitative and qualitative evaluations of the proposed 3D human pose reconstruction methodology. Our evaluation results show that the proposed approach performs reliably on a sequence of unconstrained poses and achieves an average reconstruction error of 7.46 degree in a few key joint angles. Our 3D pose recovery methodology should be applicable to many areas such as human computer interactions and human activity recognition.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Scene Analysis, Reconstruction—*Depth cues*; I.5 [**Pattern Recognition**]: Models; H.5 [**Information Interfaces and Presentation**]: HCI

SoICT '13, December 05-06 2013, Danang, Viet Nam

http://dx.doi.org/10.1145/2542050.2542071

General Terms

Design, Algorithms, Experimentation

Keywords

Human Pose Estimation, Depth Image, Body Part Recognition, Principal Direction Analysis

1. INTRODUCTION

Recovering 3D human body poses from a sequence of images in real-time is a challenging problem in computer vision. Many potential applications of this methodology in daily life include entertainment game, surveillance, sport science, health care technology, human computer interactions, motion tracking, and human activity recognition [12]. In the traditional systems, human body poses are reconstructed solving inverse kinematics using the information of optical markers that are attached to the human body parts and tracked by multiple cameras. These marker-based systems are capable of recovering accurate human body poses, but they are not suitable for real-life applications due to the sensor attachment, multiple camera installation, expensive equipment, and complicated setups [13]. In contrast to the marker-based approaches, some recent studies have focused on markerless-based methods which could be utilized in daily applications. Typically, this markerless system is based on a single image or multi-view images [14, 15, 18].

Recently, with an introduction of depth imaging devices, 3D human pose reconstruction from a single depth image without optical markers or multi-view images has become an active research topic in computer vision. Some studies have explored novel approaches in human pose estimation methodologies based on this depth information [7]. In [16, 17], depth data was used to build a graph-based representation of a depth human body silhouette and from which the geodesic distance map of the body parts was computed, finding the primary landmarks such as the head, hands, and feet. Finally, fitting a skeleton body model to the landmarks would provide a recovered human pose in 3D. In [4], using the information of primary landmarks as features of each pose, matching pose was found from the pose database

(DB) coded in the tree structure. In [6, 8, 9], depth data was presented as 3D surface meshes and then a set of geodesic feature points as head, hands, and feet was found to estimate a human pose. These approaches are generally based on the accurate alternative representation of the depth human body silhouette and accurate detection of the body parts.

Another approach in 3D body pose reconstruction utilizes a learning methodology by which each body part gets recognized, and then from the information of the recognized body parts, its corresponding 3D pose gets reconstruction. In [21], the authors developed a new algorithm based on expectation maximization (EM) with two-step iterations: namely, body part labeling (E-step) and model fitting (M-step). The depth silhouette and the estimated 3D human body model of this method were represented by a cloud of point in 3D and a set of ellipsoids, respectively. Each 3D point of the cloud was assigned and then fitted to one corresponding ellipsoid. This process was iterated by minimizing the discrepancies between the model and depth silhouette. The speed of the algorithm was slow to be realized in real-time due to high computational cost for labeling and fitting. In [19], a new approach was developed to efficiently predict 3D position of body joints from a single depth image. In their work, they treated human body part detection like an object classification task. Therefore, the human body part recognition of the depth image was inferred as a per-pixel classification via some randomized decision trees trained using a large DB of synthetic depth images. This allowed a real-time and efficient identification of human body parts: it could recognize up to 31 body parts from a single human depth silhouette. To model 3D human pose, they then applied mean-shift algorithm on the recognized human body parts to estimate joint positions. The result of estimated 3D human pose was reconstructed and visualized by the joint positions. However, joint position estimation from the recognized body parts via the mean-shift algorithm generally suffers from the following limitations: (1) the position of estimated joints depend on the shape and size of subject; (2) the computed modes lie on the surface of the body parts, whereas the position of joints are inside of the parts; (3) the methods require an arbitrary definition of body parts that roughly align with the body joints; and (4) the 3D human body model is a simple skeleton without constraints on the position of the joints.

In this paper, to overcome the limitations of the previous approaches [21, 19], we propose an improved real-time 3D human pose estimation algorithm based on Principal Direction Analysis (PDA) of each recognized body part from a single depth image. In our work, human body parts of the depth silhouette are first recognized via the trained RFs with our synthetic training DB and RFs [3]. To recover 3D human body pose, we propose a new algorithm of analyzing the recognized body parts for the principal direction vectors, improving the limitations of the mean-shift. The directional vectors are then mapped to the each body part of the 3D human body model to make the 3D estimated human pose. In addition, our 3D human body model also uses the kinematic chains with constraints to limit the movement. Our proposed methodology could be useful in human computer interactions and human activity recognition applications.

The rest of the paper is organized as follows. Section 2 introduces the processes of the proposed methodology including synthetic DB creation, RFs for pixel-based classification, body parts recognition, PDA, and reconstruction of 3D human pose model. Section 3 presents experimental setups and results. Conclusion remarks are given in Section 4.

(a) (b) (c)

Figure 2: (a) A 3D graphic human body model used in a silhouette DB generation, (b) a body part-labeled model, and (c) a depth silhouette in the synthetic DB.

2. METHODOLOGY

2.1 Overview of the proposed system

Our work focuses on estimating a 3D human pose from a single human depth silhouette. Figure 1 shows the key steps of our proposed 3D human pose estimation. In the first step, a single depth image gets captured by a depth camera. The human depth silhouette is then extracted by removing the background. In the second step, human body parts of the silhouette are recognized via the trained RFs. In the third step, the principal directions of the recognized body parts are estimated by PDA. In the finally step, these directions are mapped on to the 3D human body model, resulting in the estimated 3D human body pose.

2.2 Body part recognition

As aforementioned, to recognize the body parts from a depth human silhouette, we utilize RFs as performed in [5, 19]. However, this learning-based approach requires a training DB. In this work, we have created our own training DB synthetically. More details are given in the following subsections.

2.2.1 A synthetic DB of depth human pose silhouettes and corresponding body parts labeled maps

In order to create the training DB, we have created synthetic human body models: we have utilized 3Ds Max, a commercial 3D graphics package. The body model consists of a total 31 body parts [19]. To create various poses, motion information from CMU [1] is mapped to the model. Finally, a pair of depth silhouette and its corresponding body part-labeled map is saved into a DB: the DB contains 20.000 of depth maps and corresponding body parts labeled maps. Figure 2 shows a set of samples of the human body model, the maps of the labeled body parts, and the depth silhouette respectively. The size of images in the DB is 320 x 240 with the 16-bit depth values.

2.2.2 Depth feature extraction

In our work, the depth features are computed from the differences of a neighboring pixel pairs. The depth features f are extracted from a pixel x of the depth silhouette as done in [11, 19]

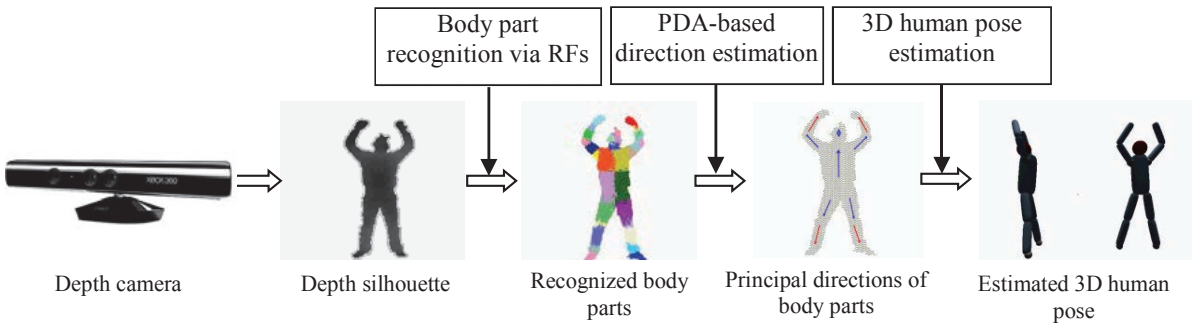

Figure 1: The key processing steps of our proposed 3D pose estimation system. These steps take the depth image, remove background, label human body parts, apply PDA of the human body parts, and finally estimate 3D human pose.

$$f_\theta(I,x) = \left[d_I\left(x + \frac{o_1}{d_I(x)}\right) - d_I\left(x + \frac{o_2}{d_I(x)}\right)\right] \quad (1)$$

where $d_I(x)$ is the depth value at pixel x in image I, and parameters $\theta = (o_1, o_2)$ describe offset o_1 and o_2. In our work, the maximum offset value of (o_1, o_2) pairs was 60 pixels corresponding to 3 meters distance from actor to camera. The normalization of the offset by $\frac{1}{d_I(x)}$ ensures that the features are distance invariant.

2.2.3 RFs for classification

RFs are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [5, 10].

In our work, to create the trained RFs, we used an ensemble of 5 decision trees in RFs. The maximum deep of trees was 20. Each tree in RFs was trained with a different pixel set of randomly synthetic depth human pose silhouettes and their corresponding body part indices. A subset of 2000 training example pixels was drawn randomly from each synthetic depth human pose silhouette in the DB. An example pixel was extracted to get 2000 candidate features as computed using Eq. 1. A subset of 50 candidate features was considered at each splitting node in the tree. For each depth pixel classification, each pixel of a tested depth silhouette was extracted to get 2000 candidate features. Based on all built trees in RFs, at each tree, starting from the root node, if the value of splitting function is less than a threshold of the node, go to left and otherwise go to right. The optimal threshold for splitting the node is determined by maximizing the information gain for particular features in training process. At the leaf node reached in each tree, the probability distribution over 31 human body parts is computed. Final recognized result of the pixel is based on the voting result of all trees in RFs.

2.3 Principal direction analysis (PDA)

In this section, our objective is to find principal directional vectors from the recognized body parts. If we denote the recognized body parts as $\{P^1,P^2,...,P^M\}$ where, M is the number of body parts. Each body part is a 3D point cloud P consisting of the n 3D points $P=\{y_i\}_{i=1}^n$, the value of n changes depending on the size of body parts. The 3D point clouds $\{P^i\}_{i=1}^M$ are used to determine principal direction vectors $\{ V_d^1, V_d^2 ..., V_d^M\}$ by the PDA algorithm. More details of PDA are given in the following sub-sections.

2.3.1 Outlier removal

The recognized body parts which are represented as clouds of points contain some outlier and mislabeled points. These incorrect points can hinder the PDA analysis, resulting in inaccurate directional vectors of the body parts. Therefore, before applying PDA, we have devised a technique to select only effective points from the cloud which are subject to PDA. In order to select effective points from the cloud, we have devised a technique to estimates the weight values of all points in the selected cloud utilizing of a logistic function and the Mahalanobis distance.

The logistic function of the population w can be written as

$$w(t_i) = \frac{C}{1 + e^{\alpha(t_i - t_0)}} \quad (2)$$

where, t_0 denotes the rough threshold value that is defined based on the size of radius of the clouds of points, α a constant value, and C the limiting value of the output (in our case $C = 1$). Here, t_0 and α are chosen based on the shape and size of each body parts. t_i is the Mahalanobis distance computed at pixel i^{th} in the point cloud and it is computed by the Mahalanobis distance written as

$$t_i = \sqrt{(y_i - \mu)^T (S)^{-1} (y_i - \mu)} \quad (3)$$

where, y_i is the i^{th} 3D point in the cloud, μ is the mean vector of the cloud and S is the covariance matrix of the cloud and it is computed as

$$S = \sum_{i=1}^{n} \frac{(y_i - \mu)(y_i - \mu)^T}{n}. \quad (4)$$

2.3.2 PDA algorithm

This part presents how to estimate the directional vectors V_d from the selected point clouds P^m. We apply a statistical approach to estimate the PDA mean vector μ^* and covariance matrix S^* using the weight value of each point from Eq. 2. The mean vector and the covariance matrix are calculated as follows

$$\mu^* = \frac{\sum_{i=1}^{n} w(t_i^2) y_i}{\sum_{i=1}^{n} w(t_i^2)}, \tag{5}$$

$$S^* = \frac{\sum_{i=1}^{n} w(t_i^2)(y_i - \mu^*)(y_i - \mu^*)^T}{\sum_{i=1}^{n} w(t_i^2) - 1}. \tag{6}$$

To estimate a direction vector V_d from a cloud P^m. The problem can be expressed as

$$V_d(E_k) = \underset{\|E_k\|_{k=1}^3}{\arg\max}(E_k^T S^* E_k) \tag{7}$$

where, E is an eigen-vector matrix of S^*. The details of the PDA algorithm are presented in Algorithm 1. Some comparison results of the PDA analysis without outlier removal and with outlier removal are shown in Figure 3.

Algorithm 1 Principal Direction Analysis (PDA)

Inputs: Given a 3D point cloud P^m

Outputs: A principal direction vector V_d

Method:

Step 1. Find the mean vector μ and the covariance matrix S of the point cloud P^m, Eq. 4.

Step 2. Compute the Mahalanobis distance of all points in the cloud P^m with its mean vector μ and covariance matrix S, Eq. 3.

Step 3. Assign the weight value for all points in the cloud P^m using logistic function and the vector of determined Mahalanobis distance, Eq. 2.

Step 4. Compute the PDA mean vector μ^* and PDA covariance matrix S^* of the point cloud P^m with using the assigned weight value of each point as Eqs. 5 and 6.

Step 5. Find the eigen-vector corresponding to the largest value of eigen-value computed from the covariance matrix S^* in Eq. 7. The eigen-vector is a determined principal direction vector V_d.

2.4 A 3D synthetic human model

To reconstruct and visualize an estimated 3D human pose, we utilize a 3D synthetic human model that is created by a set of super-quadrics. The joints of the model are connected with a kinematic chain and parameterized with rotational angles at each joint [21, 20]. Our 3D synthetic human body model is defined in the 4-D projective space as

$$m_e(X) = X^T V_\theta^T Q^T D Q V_\theta X - 2 = 0 \tag{8}$$

where X is the coordination of the 3D point on the surface of super-quadrics. D is a diagonal matrix containing the size of super-quadrics. Q locates the center of super-quadrics in the local coordination system. V_θ is a matrix containing relative kinematic parameters that is computed from the directional vectors V_d. Our model is composed of ten human body-parts (including head, torso, left and right upper arm and lower arm, left and right upper leg and lower leg) and nine joints (two knees, two hips, two elbows, two shoulders, and one neck). There is a total of 24 DOFs (including two DOFs at each joint and six free transformations from the global coordinate system to the local coordinate system at the hip).

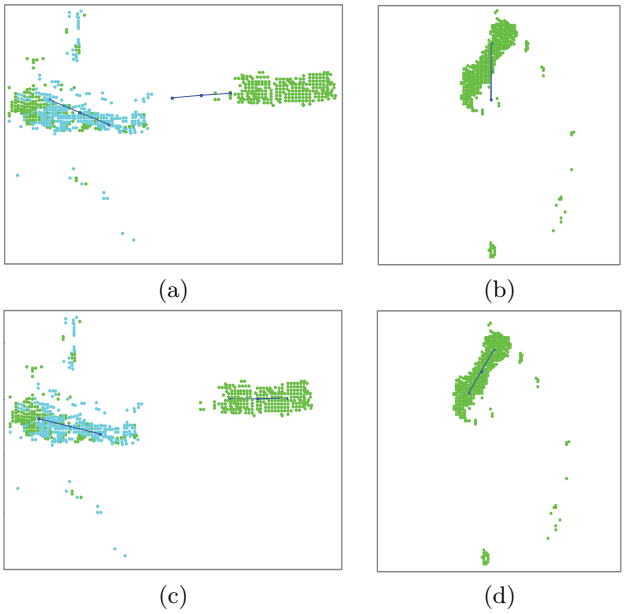

Figure 3: Comparison results of (a), (b) with PDA without outlier removal and (c), (d) with PDA with outlier removal. The results of PDA are blue lines superimposed on the clouds of points. (a), (c) two set of 3D point clouds indicate an upper arm part (left, cyan) and a lower arm part (right, green) with some mix outliers. (b), (d) a 3D point cloud of right upper arm part with some outliers.

3. EXPERIMENTAL RESULTS

We have evaluated our proposed methodology through the quantitative and qualitative assessments using synthetic and real data.

3.1 Experimental settings

To test our system, we used our own synthetic DB of 20,000 different poses for training RFs. The human body parts were recognized from a depth human body silhouette via the trained RFs. In order to evaluate quantitative assessments, we utilized synthetic depth silhouettes not from the training DB to test with the ground-truth information from the original 3D body model. At each estimated 3D human pose, we measured joint angles of a few joints from the 3D human body model and saved as the ground truth. Then we derived the same joint angles from the reconstructed 3D pose and compared them the ground truth. In our experiment, we only focus on the evaluation of the four main joints including left-right elbows and knees.

For qualitative assessment on real data, we utilized the depth silhouettes that were captured by a depth camera from Primesense [2]. Then, human body parts of the silhouettes were recognized via the trained RFs. The principal directions of the recognized body parts were estimated by

PDA. These directions were finally mapped on to the 3D human body model, resulting in the estimated 3D human body pose. The training process was run on supercomputer system. The testing process was run on the standard desktop PC with Intel Pentium IV Dual-core, 2.5 GHz CPU, and 3G RAM.

3.2 Experimental results with synthetic data

We performed a quantitative evaluation using a series of 500 depth silhouettes containing various unconstrained movements. From those test depth silhouettes, we got the labeled body parts using the trained RFs, applied the PDA algorithm, estimated the joint angles, and compared the estimated to the ground truth joint angles. In this experiment, we only focused evaluation on the four main joints including left-right elbows and knees corresponding to the eight directional vectors including lower-upper left arm, lower-upper right arm, lower-upper left leg, and lower-upper right leg. The evaluated results of our methods are provided in Figure 4. We have computed the average reconstruction error which is computed as

$$\epsilon_\theta = \frac{\sum_{i=1}^{n_f} \mid \theta_i^{est} - \theta_i^{grd} \mid}{n_f} \tag{9}$$

where n_f is the number of frames, i the frame index, θ_i^{grd} the ground-truth angle, and θ_i^{est} the estimated angle. Our quantitative evaluation produced the average reconstruction errors which of the left elbow, right elbow, left knee, and right knee are 5.52, 5.14, 8.86, and 10.34 degree, respectively. As can be seen in Figure 4 and the computed average reconstruction errors, the accuracy of estimated joint angles is better at elbows than at knees.

3.3 Experimental results with real data

In the evaluation with real data, we asked the subject to perform some unconstrained movements. Figure 5 shows the obtained results of PDA showing the principal direction as lines superimposed on the subject poses. The results of the experiments with arm movements and leg movements are given in Figures 6. The 2^{nd} and the 3^{rd} rows are results of the 3D human poses reconstruction in the front and side view point. With the real data, since we do not have the ground truth joint angles, only qualitative assessments are performed by visual inspection between the results of the 2^{nd}, 3^{rd} rows and RGB image at the 1^{st} row.

4. CONCLUSIONS

A novel method to recover a correct 3D human pose from a single depth silhouette has been proposed. The technique estimates the principal directional vectors from the recognized body parts by PDA. The quantitative assessments indicate the average reconstruction error of 7.46 degree in some key joint angles. Moreover, our methodology runs at a speed of 15FPS on a standard PC showing that our system could be suitable for real-time applications. Experiments of qualitative evaluations also show that our system is able to perform 3D human pose estimation with real data.

Acknowledgment

This research was supported by the MSIP (Ministry of Science, ICT&Future Planning), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency (NIPA-2013-(H0301-13-2001)).

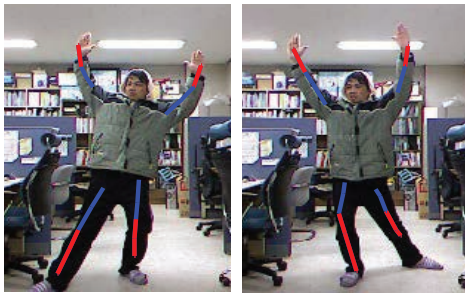

Figure 5: Sample results of PDA. The blue lines indicate the directions of the four body parts such as upper arms and legs. The red lines indicate the directions of the four body parts such as lower arms and legs.

5. REFERENCES

[1] CMU motion capture database. `http://mocap.cs.cmu.edu`.

[2] PrimeSense Ltd. `http://www.primesense.com/`.

[3] J. T. K. Ahmad Jalal, Naeha Sharif and T.-S. Kim. Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. *Journal of Indoor and built Environment*, 22:271–279, October 2013.

[4] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. Data-driven approach for real-time full body pose reconstruction from a depth camera. In *Proceedings of the 2011 International Conference on Computer Vision*, pages 1092–1099, 2011.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

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

[7] L. Chen, H. Wei, and J. Ferryman. A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, February 2013.

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

[9] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun. Real-time human pose tracking from range data. In *Proceedings of the 12th European conference on Computer Vision*, pages 738–751, 2012.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2008.

[11] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 775–781, 2005.

[12] T. B. Moeslund, A. Hilton, and V. Krüger. A survey

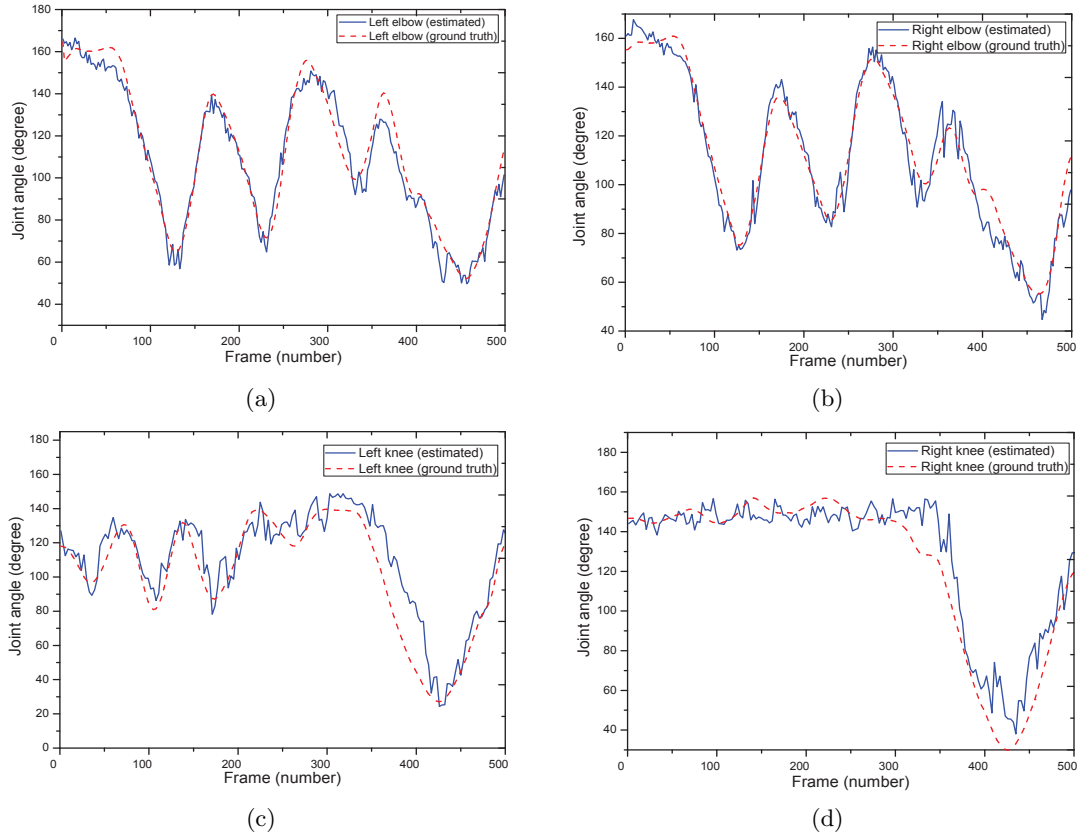

Figure 4: Comparison results between the ground-truth and the estimated joint angles using the synthetic BD: (a) joint angle of left elbow, (b) joint angle of right elbow, (c) joint angle of left knee, and (d) joint angle of right knee.

of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, November 2006.

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

[14] B. Rosenhahn, U. G. Kersting, A. W. Smith, J. K. Gurney, T. Brox, and R. Klette. A system for marker-less human motion estimation. *Lecture Notes in Computer Science*, 3663:230–237, 2005.

[15] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, D. Cremers, and H. P. Seidel. Markerless motion capture of man-machine interaction. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 23–28, 2008.

[16] 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 *IEEE conference on Automatic Face and Gesture Recognition*, pages 700–706, 2011.

[17] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab. Human skeleton tracking from depth data using geodesic distances and optical flow. *Journal Image and Vision Computing*, 30(3):217–226, March 2012.

[18] J. ShenAuthor, W. YangAuthor, and Q. Liao. Part template: 3d representation for multiview human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):1920–1932, July 2013.

[19] 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 *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

[20] A. Sundaresan and R. Chellappa. Model-driven segmentation of articulating humans in laplacian eigenspace. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1771–1785, October 2008.

[21] N. D. Thang, T.-S. Kim, Y.-K. Lee, and S. Lee. Estimation of 3-d human body posture via co-registration of 3-d human model and sequential stereo information. *Journal Applied Intelligence*, 35(2):163–177, October 2011.

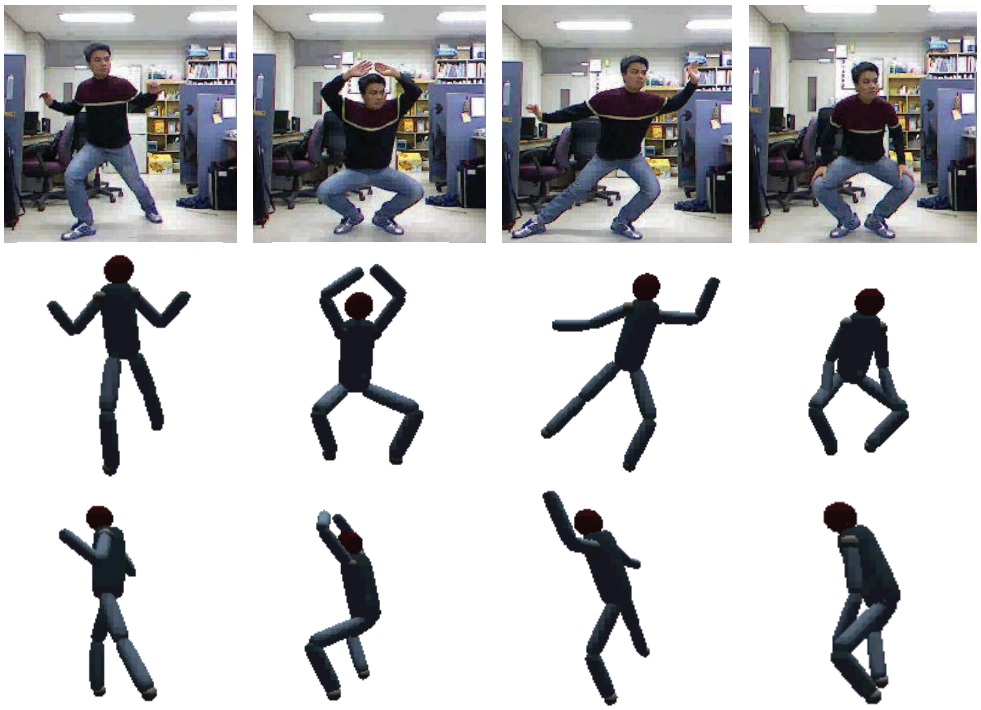

Figure 6: Sample results of our proposed 3D human pose estimation for arm and leg movements: the 1^{st} row shows RGB images of four different poses, the 2^{nd} and 3^{rd} rows show the results of estimated 3D human poses in the front and side views respectively.