Applied Intelligence

Real-time 3D Human Pose Recovery from a Single Depth Image Using Principal Direction Analysis

--Manuscript Draft--

Manuscript Number:	
Full Title:	Real-time 3D Human Pose Recovery from a Single Depth Image Using Principal Direction Analysis
Article Type:	Original Submission
Keywords:	3D human pose recovery; Depth image; Body part recognition; Principal direction analysis
Corresponding Author:	Tae-Seong Kim, Ph.D. Kyung Hee University Suwon, Seocheon-dong KOREA, REPUBLIC OF
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Kyung Hee University
Corresponding Author's Secondary Institution:	
First Author:	Dong-Luong Dinh, M.S
First Author Secondary Information:	
Order of Authors:	Dong-Luong Dinh, M.S
	Myeong-Jun Lim, M.S
	Duc-Thang Nguyen, Ph.D
	Sungyoung Lee, Ph.D
	Tae-Seong Kim, Ph.D.
Order of Authors Secondary Information:	

Noname manuscript No.
(will be inserted by the editor)

Real-time 3D Human Pose Recovery from a Single Depth Image Using Principal Direction Analysis

Dong-Luong Dinh · Myeong-Jun Lim · Nguyen Duc Thang · Sungyoung Lee · Tae-Seong Kim

Received: date / Accepted: date

Abstract In this paper, we present a novel approach to recover a 3D human pose in real-time from a single depth image using Principal Direction Analysis (PDA). 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 points in 3D, PDA is applied to estimate its principal direction. Finally, a 3D human pose gets recovered by mapping the principal directional vector to each body part of a 3D human body model. In our experiments, we have performed both quantitative and qualitative evaluations of the proposed 3D human pose recovering methodology. Our evaluation results show that the proposed approach performs reliably on a sequence of unconstrained poses and achieves an average reconstruction error of 7.07 degree in four key joint angles. In addition, our methodology runs at a speed of 20 FPS on a standard PC showing that our system could be suitable for real-time applications. Our 3D pose recovery methodology should be applicable to many areas such as human computer interactions and human activity recognition.

Keywords 3D human pose recovery · Depth image · Body part recognition · Principal direction analysis

Dong-Luong Dinh
Department of Computer Engineering, Kyung Hee University, 1 Seocheon-dong, Giheung-gu, Yongin-si,Gyeonggi-do, Republic of Korea
E-mail: luongdd@oslab.khu.ac.kr

Myeong-Jun Lim
Department of Biomedical Engineering, Kyung Hee University, 1 Seocheon-dong, Giheung-gu, Yongin-si,Gyeonggi-do, Republic of Korea
E-mail: mjlim@khu.ac.kr

Nguyen Duc Thang
Department of Biomedical Engineering, International University, Ho Chi Minh City, Vietnam
E-mail: ndthang@hcmiu.edu.vn

Sungyoung Lee
Department of Computer Engineering, Kyung Hee University, 1 Seocheon-dong, Giheung-gu, Yongin-si,Gyeonggi-do, Republic of Korea
E-mail: sylee@oslab.khu.ac.kr

Tae-Seong Kim
Department of Biomedical Engineering, Kyung Hee University, 1 Seocheon-dong, Giheung-gu, Yongin-si,Gyeonggi-do, Republic of Korea
Tel.: +82-31-201-3731
E-mail: tskim@khu.ac.kr

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 [13]. In the conventional systems, human body poses are reconstructed by solving inverse kinematics using the motion information of optical markers 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 [15]. 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 RGB image or multi-view RGB images [6,16,17].

Recently, with an introduction of depth imaging devices, 3D human pose recovery from a single depth image without optical markers and multi-view RGB images has become an active research topic in computer vision. Some studies have exploited novel approaches in human pose estimation methodologies based on the depth information [6]. In [18,19], depth data was used to build a graph-based representation of an human body silhouette and from which the geodesic distance map of the body parts was computed to find the primary landmarks such as the head, hands, and feet. Fitting a skeleton body model to the landmarks recovered human pose in 3D. In [4], using the information of primary landmarks as the features of each pose, the best matching pose was found from the set of poses coded in the hierarchy tree structure. In [14,8,9], depth data was presented as 3D surface meshes and then a set of geodesic feature points such as head, hands, and feet was found for tracking human pose. These approaches are generally based on the alternative representation of the depth human body silhouette and the detection of the body parts.

Another approach in 3D body pose recovery utilizes a learning methodology by which each body part gets recognized. From the information of the recognized body parts, its corresponding 3D pose gets reconstruction. In [23], 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 points 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. However, the speed of the algorithm was slow to be realized in real-time due to high computational cost for labeling. In [20], a new approach was developed to human pose recognition in parts from a single depth image. 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 Database (BD) 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 the mean-shift algorithm [7] on the recognized human body parts to estimate the body joint positions. Human body pose was recovered from these joint points. However, joint position estimation via the mean-shift algorithm generally suffers from the following limitations: (i) the position of estimated joints depend on the shape and size of subject, (ii) the computed relative information concentrates on the surface of the body parts, whereas the position of joints are inside of the parts, (iii) the method requires the value of the parameters such as window size that is unspecified.

In this paper, to overcome the limitations of the previous approaches [20,23], we propose a novel algorithm to recover a 3D human pose in real-time via Principal Direction Analysis (PDA) on the recognized human body parts from a single depth image. In our work, human body parts of the depth silhouette are first recognized via the trained Random Forests (RFs) with our synthetic training DB (DB) [11]. Using PDA, principal directional vectors are estimated from the recognized body parts. Then the directional vectors are mapped to the each body part of the 3D human body model to make the recovered 3D human pose which is constrained by the kinematic chains to allow feasible body movements.

The rest of the paper is organized as follows. In section 2, we describe our overall system. Sections 3, 4 and 5 introduce the processes of the proposed methodology including synthetic DB creation, RFs for a pixel-based classification, body parts recognition, PDA, and reconstruction of 3D human pose model. Section 6 presents experimental results and comparisons. Conclusion and discussion remarks are given in Section 7.

2 Our methodology

Our work focuses on recovering a 3D human pose from a single human depth silhouette. Fig. 1 shows the key steps of our proposed 3D human pose recovering methodology. 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. Finally, these directions are mapped on to the 3D human body model, resulting in the recovered 3D human body pose.

3 Body parts recognition

As aforementioned, to recognize the body parts from a depth human silhouette, we utilize RFs as performed in [5,11,20]. This learning-based approach requires a training DB, therefore, we have created our own training DB synthetically [11]. More details are given in the following sub-sections.

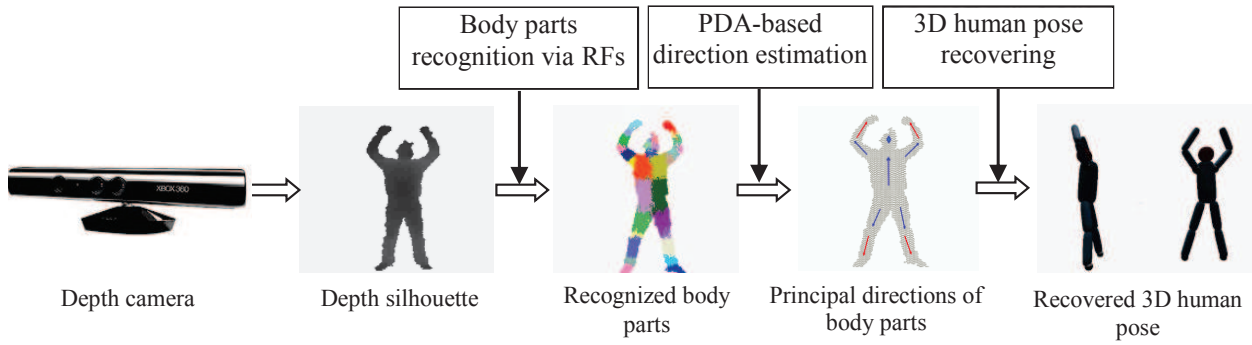


Fig. 1 The key processing steps of our proposed system. These steps consist of taking the depth image, removing backgrounds, labeling body parts, applying PDA of the body parts and finally recovering a 3D human pose.

(a) (b) (c)

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

3.1 A synthetic DB of depth maps and corresponding body parts labeled maps

In order to create the training DB, we have created synthetic human body models using 3Ds Max, a commercial 3D graphic package [1]. The body model consists of a total 31 body parts [11]. To create various poses, motion information from Carnegie Mellon University (CMU)'s motion DB [2] 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. Fig. 2 shows a set of samples of the human body model, a map of the labeled body parts, and its corresponding depth silhouette respectively. The size of images in the DB is 320 x 240 with the 16-bit depth values.

3.2 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 [12,20]

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

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

3.3 RFs for body parts labeling

In our work, we utilize RFs for body parts recognition. 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 order to create the trained RFs, we used an ensemble of five decision trees. The maximum deep of trees was twenty. Each tree in RFs was trained with different pixels randomly sampled from the synthetic depth silhouettes and their corresponding body part indices. A subset of 2,000 training sample pixels was drawn randomly from each synthetic depth silhouette in the DB. A sample pixel was extracted to get 2,000 candidate features as computed using Eq. (1). At each splitting node in the tree, a subset of 50 candidate features was considered. For pixel classification, each pixel of a tested depth silhouette was extracted to get the 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, goes to left and otherwise goes to right. The optimal threshold for splitting the node is determined by maximizing the information gain in the training process. At the leaf node reached in each tree, the probability distribution over 31 human body parts

is computed. Final decision to label each depth pixel for a specific body part is based on the voting result of all trees in RFs.

4 3D human pose proposals from the recognized body parts

In this part, we introduce the mean shift and PDA from which human poses are recovered in 3D.

4.1 Joint position proposal based on the mean shift

In [20], to recover a 3D human pose, a 3D human skeleton model of the joints is used where the joints are fitted from the recognized body parts using the mean-shift algorithm with a weighted Gaussian kernel. The mean shift algorithm is a nonparametric density estimation used for seeking the nearest mode of a point sample distribution [24]. The technique is commonly used in image segmentation and object tracking fields of computer vision [7,21]. Given n data points x_i, $i = 1, ..., n$ on a d-dimensional space R^d, the multivariate density obtained with the kernel $K(x)$, window radius of kernel h and weight function w is

$$\hat{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} w_i K\left(\frac{x - x_i}{h}\right). \tag{2}$$

The sample mean with kernel K $(G(x) = K'(x))$ at a point x is defined as

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i w_i G\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} w_i G\left(\frac{x-x_i}{h}\right)}. \tag{3}$$

The difference between $m_h(x)$ and x is called the mean shift. The mean shift vector always points toward the direction of the maximum value of the density. Therefore, the mean shift procedure is guaranteed to converge to a point where the gradient of the defined density function approaches zero. The mean shift algorithm process is illustrated in Fig. 3. Starting on the data point in cyan, the mean shift procedure is performed to find the stationary point in red of the density function. In order to optimize the parameters and improve the efficiency of the mean shift, in [20], the size of window h was replaced by b_c which is a learned per-part bandwidth and the weight w_{ic} is from the probability distribution of each pixel in a class C and the given depth $d_I(xi)$. The formula is written as

$$w_{ic} = P(x|I, x_i).d_I(x)^2. \tag{4}$$

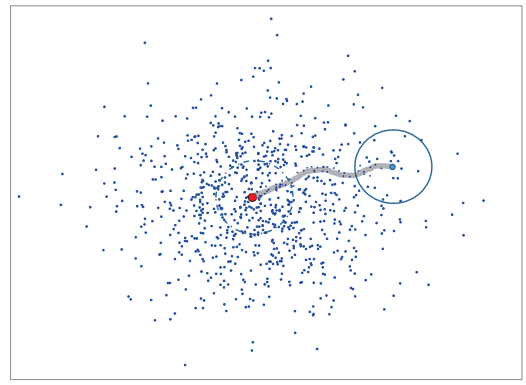


Fig. 3 Mean shift iteration process to find the centroid of a cloud.

To reconstruct and visualize an estimated 3D human pose, a skeleton model is presented by joint points estimated from the recognized body parts using the mean-shift. There are some limitations: the optimal window size that is difficult to find, so that an inappropriate window size can cause the modes to be merged; the position of estimated joints depends on the shape and size of the recognized body parts and only computed on the surface of the body parts, whereas the position of joints are inside of the parts. In order to overcome the limitations of this approach, we propose a PDA algorithm presented in the following section 4.2.

4.2 Principal direction analysis of the recognized body parts

In this section, our objective is to find principal direction 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^m consisting of the n 3D points $P^m = \{x_i\}_{i=1}^{n}$, the value of n changes with the size of body parts. The 3D point clouds $\{P^m\}_{m=1}^{M}$ 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.

4.2.1 Outlier removal

The recognized body parts which are represented as clouds of points contain some outliers and mislabeled points. These points can hinder PDA, resulting in inaccurate directional vectors of the body parts. Therefore, before applying PDA, we have devised a technique to select only interested points from each labeled point cloud which are subject to PDA. In order to select these points from the cloud, we have devised a technique to estimate the weight value of all points in the selected

cloud utilizing a logistic function and the Mahalanobis distance.

The logistic function of the population w can be written as

$$w(t_i) = \frac{L}{1 + e^{\alpha(t_i - t_0)}} \tag{5}$$

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

$$t_i = \sqrt{(x_i - \mu)^T (S)^{-1} (x_i - \mu)} \tag{6}$$

where, x_i is the ith point in the cloud, μ the mean vector of the cloud, and S the covariance matrix of the cloud which is computed as

$$S = \sum_{i=1}^{n} \frac{(x_i - \mu)(x_i - \mu)^T}{n}. \tag{7}$$

Our proposed approach is illustrated in Fig. 4. The selected points, subject to PDA, are shown in red that are used to determine the direction vector. While the points in green are regarded as outliers. The size or population of the region containing the points is controlled by the threshold parameter t_0, while the parameter α is used to control the weight value of points in the cloud. This means that if we assume that the weight value of function $w \in [0,1]$ then the weight of the points in red near from the centroid of the cloud is approximately 1, while the weight of the points in green far from the centroid is approximately 0. The weight of points around the threshold value t_0 is approximately 0.5.

4.2.2 PDA

This part presents how to estimate the directional vectors V_d from the selected point cloud P^m. We apply a statistical approach to estimate the PDA mean vector μ^* and covariance matrix S^* by using the weight value of each point as in Eq. (5). The mean vector and the covariance matrix of PDA are calculated as follows

$$\mu^* = \frac{\sum_{i=1}^{n} w(t_i^2) x_i}{\sum_{i=1}^{n} w(t_i^2)}, \tag{8}$$

$$S^* = \frac{\sum_{i=1}^{n} w(t_i^2)(x_i - \mu^*)(x_i - \mu^*)^T}{\sum_{i=1}^{n} w(t_i^2) - 1}. \tag{9}$$

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

$$V_d(E_k) = \underset{\{E_k\}_{k=1}^{3}}{\arg\max}(E_k^T S^* E_k) \tag{10}$$

where, E is an eigen-vector matrix of S^*.

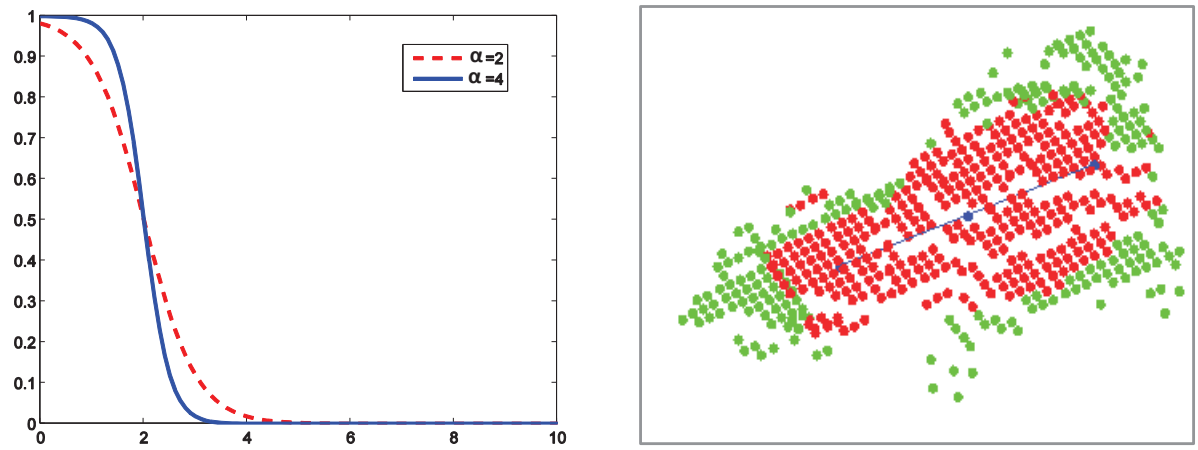
Algorithm 1 Principal Direction Analysis (PDA)

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, as in Eq. (7).
Step 2. Compute the Mahalanobis distance of all points in the cloud P^m with its mean vector μ and covariance matrix S in Eq. (6).
Step 3. Assign the weight value for all points in the cloud P^m using logistic function and the vector of determined Mahalanobis distance, as in Eq. (5).
Step 4. Compute the PDA mean vector μ^* and PDA covariance matrix S^* of the point cloud P^m using the assigned weight value of each point as in Eqs. (8) and (9).
Step 5. Find the eigen-vector corresponding to the largest value of eigen-value computed from the covariance matrix S^* in Eq. (10). The eigen-vector is a determined principal direction vector V_d.

We apply PDA to estimate the directional vectors of body parts on the 3D point clouds. Note that a 3D point cloud P^m, is presented as an n by 3 matrix, where n denotes the number of the 3D points in the cloud P^m and each point consists of three x, y, and z coordinates, respectively. To find the direction vector V_d of the cloud P^m, PDA starts with a covariance matrix S determined from the P^m=(P_x^m, P_y^m, P_z^m) and a vector of values with mean $\mu = (\mu_x, \mu_y, \mu_z)$. From the covariance matrix and mean vector, Mahalanobis distance of each point in the cloud is computed in Eq. (6). The result of Eq. (5) provides the weight vector corresponding to the points in the cloud P^m. Based on the weight vector of each pixel in the cloud, a PDA covariance matrix S^* and mean vector μ^* are determined by Eqs. (8) and (9). Finally, a direction unit vector V_d of the cloud P^m is estimated. The details of the PDA algorithm are presented in Algorithm 1. Some comparison results of the PDA were performed on the point clouds with outliers as illustrated in Fig. 5. The results of the estimated principal directions shown as lines in blue were directly drawn on the clouds.

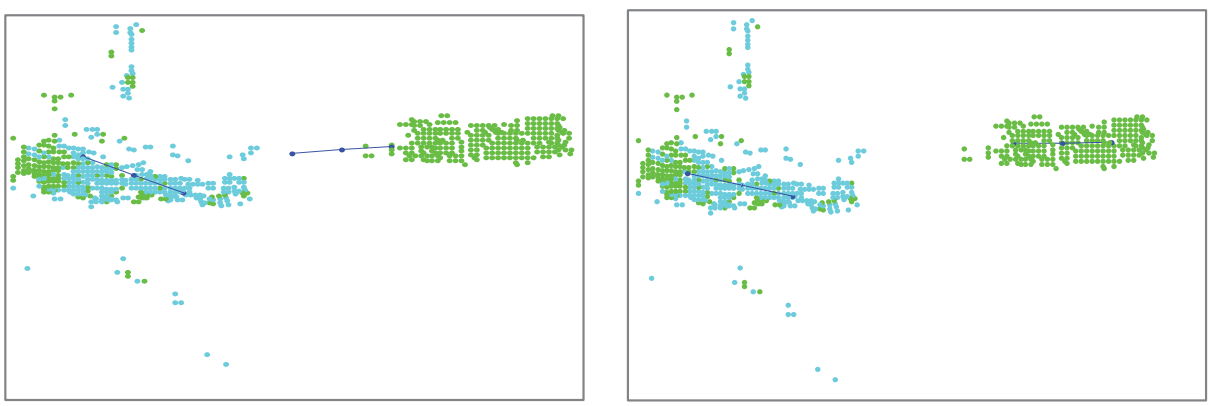
5 3D human pose representation

To represent a recovered 3D human pose, we utilize a 3D synthetic human model that is created by a set of



(a) (b)

Fig. 4 (a) Logistic function with $t_0 = 2$ and $\alpha = 2$, $\alpha = 4$. (b) Effect of the parameters t_0 and α on threshold value of 3D point clouds to eliminate outliers ($t_0 = 2$, $\alpha = 4$).



(a) (b)

Fig. 5 Comparison results of PDA (a) without outlier removal and (b) with outlier removal. The resultant principal directions are blue lines superimposed on the point clouds. Two set of 3D point clouds indicate an upper arm part (left, cyan) and a lower arm part (right, green) with some outliers.

(a) (b)

Fig. 6 3D synthetic human model. (a) Orientation model and (b) 3D model with super-quadrics shapes.

super-quadrics. The joints of the model are connected with a kinematic chain and parameterized with rotational angles at each joint [22,23]. Our 3D synthetic human body model is defined in the 4-D projective space as

$$m_e(X) = X^T V_\theta^T Q^T D Q V_\theta X - 2 = 0 \tag{11}$$

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) as shown in Fig. 6. In Fig. 6(a),

the dash line and its arrow superimposed on the model show the results of PDA and Fig. 6(b) is the result of its corresponding recovered 3D human pose with the 3D model of super-quadrics.

6 EXPERIMENTAL RESULTS

In this section, we have evaluated our proposed methodology through the quantitative and qualitative assessments using synthetic and real data as well as though comparison against the previous works [20, 23].

6.1 Experimental settings

In order to evaluate quantitative assessments, we utilized synthetic depth silhouettes and the ground-truth information extracted from the synthetic 3D body pose. For each synthetic 3D human pose, we measured joint angles of four major joints including the left-right elbows and knees from the 3D human body model and saved as the ground truth. Then, each recovered 3D human pose from the corresponding body depth silhouettes were recognized via the trained RFs and estimated the principal directions by PDA. Finally, We derived the same joint angles from the recovered 3D pose and compared them against the ground truth. For qualitative assessments on real data, we utilized the depth silhouettes that were captured by a depth camera [3]. These directions were finally mapped on to the 3D human body model, resulting in the recovered 3D human body pose. To assess on real data, visual inspection between the results of the recovered 3D human poses and RGB images was performed. Pose recovery was run on the standard desktop PC with Intel Pentium Core i5, 3.4 GHz CPU, and 8GB RAM.

6.2 Experimental results with synthetic data

We performed a quantitative evaluation using a series of 500 depth silhouettes containing various unconstrained movements. In this experiment, the evaluation results with the synthetic poses of our proposed methods are provided in Figs. 7 and 8. At each plot of Fig. 8 corresponds to an estimated joint angle by PDA. The solid and dashed lines indicate the PDA estimated and its ground truth joint angles, respectively.

Based on the results of estimated joint angles and the ground truth joint angles, we have computed the average reconstruction error as

$$\epsilon_\theta = \frac{\sum_{i=1}^{n_f} \mid \theta_i^{est} - \theta_i^{grd} \mid}{n_f}, \tag{12}$$

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. To assess the reconstruction errors, we performed the another experiment on the four different sequences of swimming, boxing, cleaning, and dancing activities. Each sequence contains 100 frames. The average errors at four considered joint angle of the second experiment are given in Table 1. The average reconstruction error of the four different sequences at four considered joint angle is $7.07 degree$.

6.3 Experimental results with real data

In the evaluation with real data, we asked three subjects to perform unconstrained movements. Two experiments were performed. In the first experiment, we examined the principal direction estimation using PDA on one subject. Fig. 9 shows the results in which the principal directions are shown as lines superimposed on the subject's poses. In the second experiment, we assessed the movements of the elbows and knees with arbitrary poses (some simple and complex poses). The results of the experiments with arm movements and leg movements of the first subject are shown in Fig. 10. The 2nd and the 3rd rows are results of the 3D human poses reconstruction in the front and side views. With the real data, since the ground truth joint angles are not available, only qualitative assessments were performed by visual inspection between the results of the 2nd, 3rd rows and RGB images at the 1st row. Fig. 11 shows the qualitative assessments on two other subjects who were different in body size and shape.

6.4 Comparisons against conventional methods

We have evaluated the performance of our proposed methodology by comparing against the conventional methods in [20, 23].

In comparison against the mean shift method [20], we implemented the real-time human pose recognition system as done in [20]. Our own synthetic DB was used to train RFs in this system. We evaluated the mean shift method through the quantitative and qualitative assessments using synthetic and real data. Table 2 provides the comparison results of quantitative assessments on the same tested synthetic data, the average reconstruction error at the four considered joint angles of our method is 7.07 degree compared to 9.79 degree from the mean shift method. We also performed a qualitative assessment on the same real data. The results of the recovered 3D human poses are represented on the same 3D synthetic pose model as shown in Fig. 12. As can

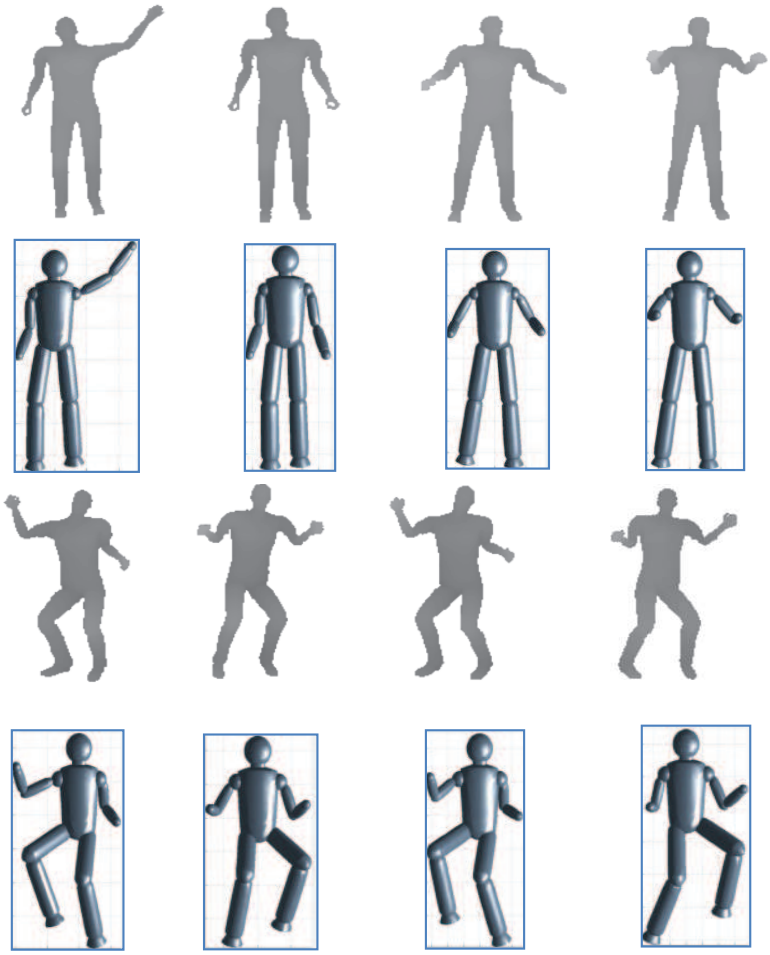


Fig. 7 Sample results of our proposed 3D human pose estimation on our synthetic data. The 1st and 3rd rows: the synthetic depth map. The 2nd and 4th rows: the estimated 3D human poses.

Table 1 The average reconstruction error of the evaluated angles joint angle in degree of 100 frames per each activity

Activities	Left elbow	Right elbow	Left knee	Right knee
Swimming	5.11	5.12	8.34	8.67
Boxing	6.78	6.57	8.12	9.24
Cleaning	5.45	5.62	7.56	7.67
Dancing	5.42	5.19	8.86	9.34
Average reconstruction error (^{o})	**5.69**	**5.63**	**8.22**	**8.73**

be seen in Fig. 12, our proposed methodology has significantly improved accuracy compared with the pose reconstruction based on the mean shift method. In particular, our proposed method has proved more robust than the mean shift method in some poses of overlapped or intersected body parts.

In comparison against the EM method [23], we used the average reconstruction errors, which is computed from the four experiments as given in [23] at left-right elbows and knees, were 7.50, 7.63, 8.03, and 13.81 degree compared to 5.69, 5.63, 8.22, and 8.73 degree from our proposed method as shown in Table 2. The obtained average reconstruction error of the proposed system in [23] are higher than our proposed system.

7 Conclusion and discussion

A novel method to recover a 3D human pose from a single depth silhouette has been proposed. The technique estimates the principal direction vectors from the recognized body parts by PDA. The quantitative assessments indicate the average reconstruction error of 7.07 degree, whereas the conventional approach of the mean shift and the EM methods produce 9.79 degree and 9.24

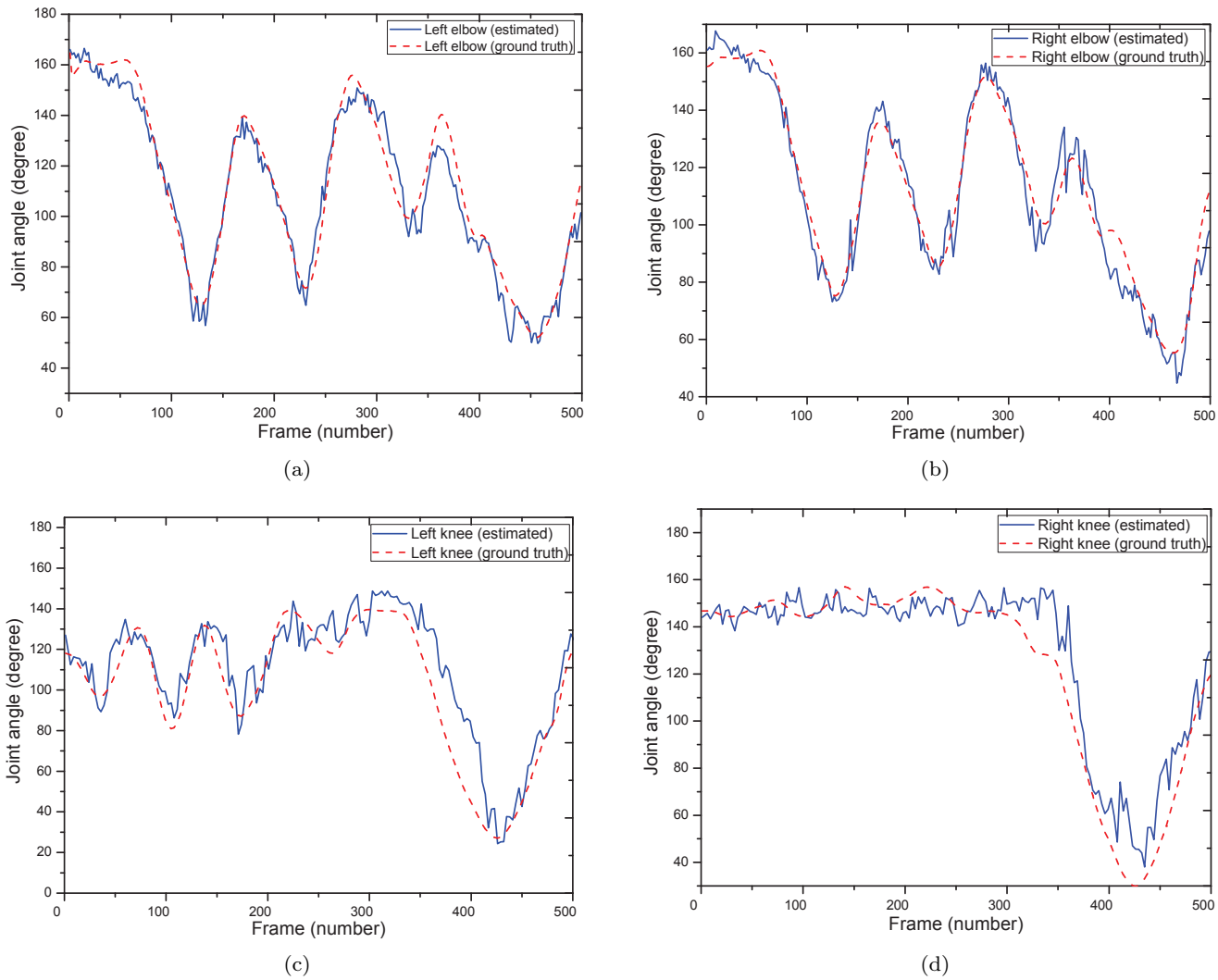


Fig. 8 A comparison between the ground-truth and the estimated joint angles in synthetic data: (a) joint angle of left elbow, (b) joint angle of right elbow, (c) joint angle of left knee, and (d) joint angle of right knee.

Fig. 9 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.

Fig. 10 Sample results of our proposed 3D human pose estimation on four different poses of arm and leg movements: the 1st row shows RGB images of four different poses, the 2nd and 3rd rows show the results of estimated 3D human poses in the front and side views respectively.

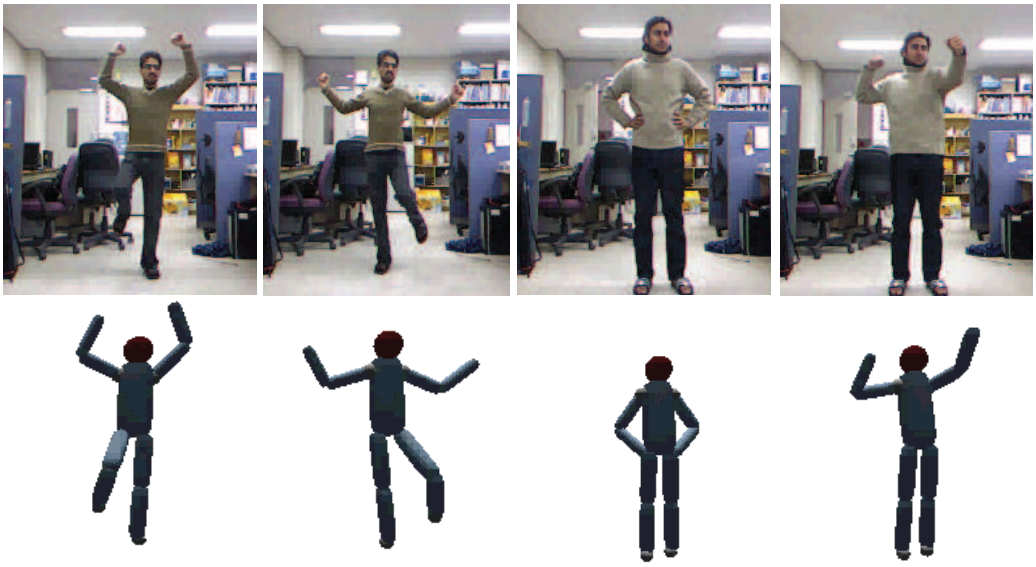


Fig. 11 Sample results of our proposed 3D human pose estimation on four different poses of difference shape subjects: the 1st row shows RGB images, the 2nd row shows the results of estimated 3D human poses from two different subjects.

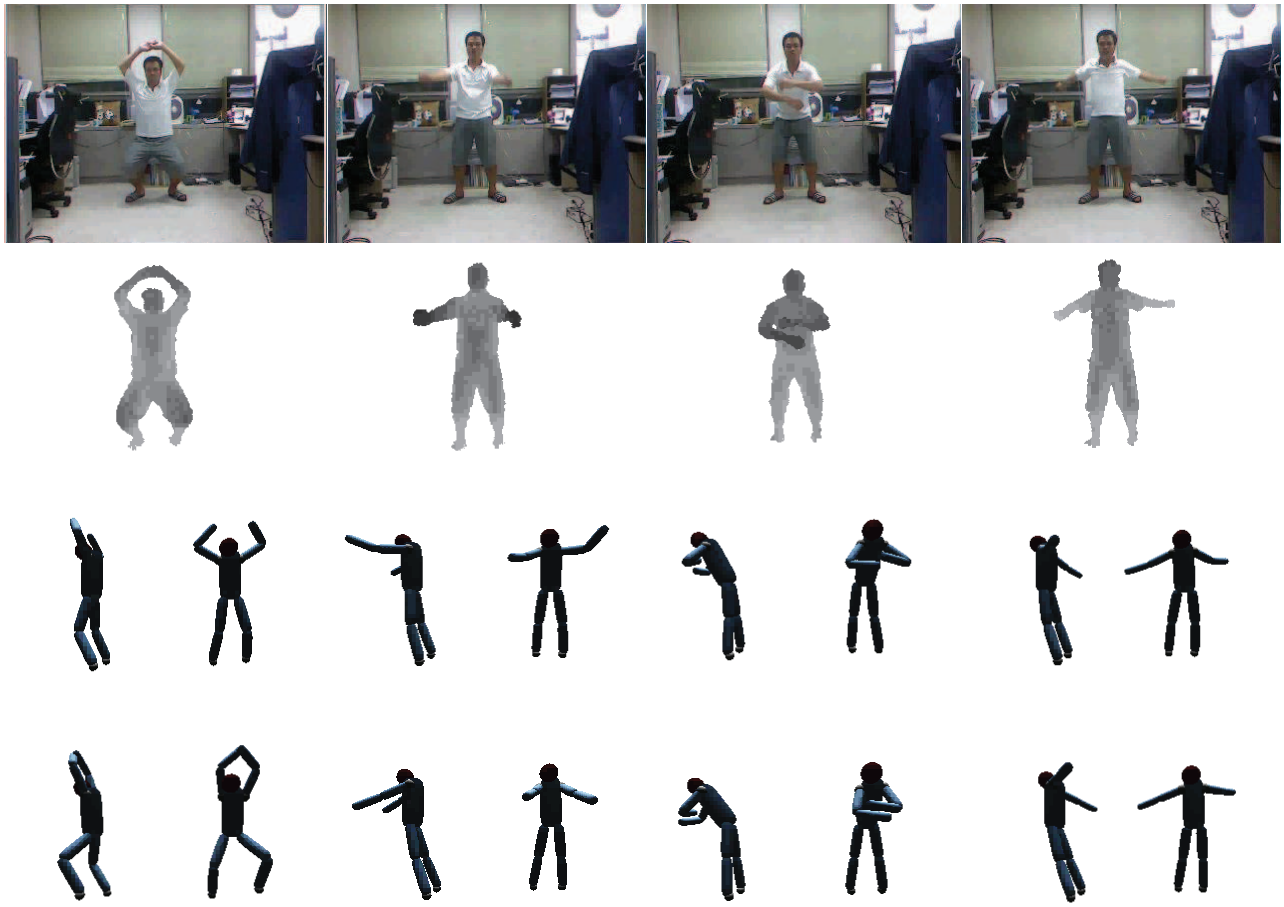


Fig. 12 Comparison results against [20] of four different poses: the 1st row shows RGB images, the 2nd row shows the depth silhouettes, the 3rd row shows the results obtained from the mean shift algorithm and the 4th row shows the results obtained from our proposed PDA algorithm.

Table 2 A comparison about the average reconstruction error (^{0})

Evaluated angles	Left elbow	Right elbow	Left knee	Right knee	Average error of the four joints
The method proposed by [23]	7.50	7.61	8.03	13.81	9.24
The method proposed by [20]	9.24	9.41	10.15	10.34	9.79
Our proposed method	5.69	5.63	8.22	8.73	7.07

degree in the four key joint angles, respectively. Our methodology runs at a speed of 20 FPS on a standard PC showing that our system is be suitable for real-time human activity recognition and human computer interaction applications for personal life-care and health-care service of the elderly and disabled people. The experiments on real data show that our system reliably performs on sequences containing unconstrained movements of various appearance and differently shaped subjects.

Acknowledgements 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)). This work also was supported by the Industrial and Strategic technology development program, (10043977, Development of a Sustainable and Practical Wellness System using Emotion Mechanism and Smart Media for the Elderly) funded by the Ministry of Knowledge Economy(MKE, Korea).

References

1. Autodesk 3Ds MAX, 2012.
2. CMU motion capture database. `http://mocap.cs.cmu.edu`
3. PrimeSense Ltd. `http://www.primesense.com`
4. Baak, A., Mller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: Data-driven approach for real-time full body pose reconstruction from a depth camera. In: ICCV'11 ICCV

'11 Proceedings of the 2011 International Conference on Computer Vision, pp. 1092–1099 (2011)
5. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
6. Chen, L., Wei, H., Ferryman, J.: A survey of human motion analysis using depth imagery. Pattern Recognition Letters (2013)
7. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5), 603–619 (2002)
8. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 755–762 (2010)
9. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real-time human pose tracking from range data. In: ECCV'12 Proceedings of the 12th European conference on Computer Vision, pp. 738–751 (2012)
10. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2008)
11. Jalal, A., Sharif, N., Kim, J.T., Kim, T.S.: 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 (2013)
12. Lepetit, V., Lagger, P., Fua, P.: Randomized trees for real-time keypoint recognition. In: CVPR '05 Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 775–781 (2005)
13. Moeslund, T.B., Hilton, A., Krger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding **104**(2), 90–126 (2006)
14. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 3108–3113 (2010)
15. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding **108**(1-2), 4–18 (2007)
16. Rosenhahn, B., Kersting, U.G., Smith, A.W., Gurney, J.K., Brox, T., Klette, R.: A system for marker-less human motion estimation. Lecture Notes in Computer Science **3663**, 230–237 (2005)
17. Rosenhahn, B., Schmaltz, C., Brox, T., Weickert, J., Cremers, D., Seidel, H.P.: Markerless motion capture of man-machine interaction. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 23–28 (2008)
18. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N.: 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, pp. 700–706 (2011)
19. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N.: Human skeleton tracking from depth data using geodesic distances and optical flow. Journal Image and Vision Computing **30**(3), 217–226 (2012)
20. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR '11 Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1297–1304 (2011)
21. Shuang, Z., Yu-ping, Q., Hao, D., Gang, J.: Analyzing of mean-shift algorithm in extended target tracking technology. Lecture Notes in Electrical Engineering **144**, 161–166 (2012)
22. Sundaresan, A., Chellappa, R.: Model-driven segmentation of articulating humans in laplacian eigenspace. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(10), 1771–1785 (2008)
23. Thang, N.D., Kim, T.S., Lee, Y.K., Lee, S.: 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 (2011)
24. Vilaplana, V., Marques, F.: Region-based mean sift tracking: Application to face tracking. In: IEEE International Conference on Image Processing, pp. 2712–2715 (2008)