2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Selecting Best Viewpoint for Human-Pose Estimation

Kai-Chi Chan, Cheng-Kok Koh and C. S. George Lee

***Abstract*— Estimating human poses is an important step towards developing robots that can understand human motion. Since a human is highly articulated, changing viewpoints of sensors on robots can improve the accuracy of human-pose estimation. We propose a two-phase approach that determines the best viewpoint of a depth sensor for human-pose estimation. The proposed approach measures the quality of potential viewpoints and selects one of them as the best viewpoint for each human pose. Based on the quality of viewpoints, human poses can be directly mapped to the best viewpoint without reconstructing the human body. Thus, the proposed approach provides a discriminative mapping to determine the best viewpoint for estimating different human poses. To measure the quality of a potential viewpoint, the viewpoint is first instantiated by representing the depth sensor of the viewpoint using the finite projective camera model. The quality of the viewpoint is expressed in terms of the error of human-pose estimates. A mapping is derived by minimizing the error in a human-pose estimate among different viewpoints. The proposed two-phase approach has been evaluated on a benchmark database. Experimental results showed that the best viewpoint for a human pose could be determined by evaluating the quality of potential viewpoints. The mean error and standard deviation of human-pose estimates were reduced by using the best viewpoint determined by the proposed two-phase approach.**

I. INTRODUCTION

Estimating human poses is important in human-motion analysis, which is beneficial to the development of robots' cognitive capabilities. Most previous and related work about human-pose estimation are based on stationary sensors [1]–[5]. In many applications, robots equipped with sensors can be considered to be moving sensors, and the mobility of robots with sensors could improve the accuracy of estimating human poses by determining the best viewpoint of sensors. Figure 1 shows a Hubo-II+ humanoid robot and a P3-DX mobile robot in our ARTLab, each mounted with an RGB-D sensor, and both can be considered to be moving sensors.

Using a moving robot equipped with an RGB-D sensor, the process of human-pose estimation involves iterations of three stages — human-pose prediction, best-viewpoint determination, and moving the robot to the best-viewpoint location. When a robot observes a person at time t, the future pose of the person at $t+\delta t$ is predicted, where δt is the time to account for the movement of the robot. The best viewpoint for the predicted human pose is determined. The robot can then be moved to the best viewpoint in δt. The process is then repeated. In this paper, we focus on the problem in the second stage; that is, for a given predicted human pose, we determine the best viewpoint, which includes the position and orientation, of a depth sensor for human-pose estimation. The problems of human-pose prediction and moving the robot to the best viewpoint will be our future work.

(a) A Hubo-II+ humanoid robot. (b) A P3-DX mobile robot.

Fig. 1: A Hubo-II+ humanoid robot and a P3-DX mobile robot in our ARTLab, each mounted with an RGB-D sensor.

Past research work in finding the best viewpoint has mainly focused on rigid and stationary objects [6], [7]. An information gain [8], [9] has been derived to represent the amount of new information for each viewpoint. The viewpoint that maximizes the information gain is deemed the best and selected. Unfortunately, an information gain is not directly applicable in human-pose estimation for two main reasons. First, a *change* in human posture alters an information gain. The best viewpoint selected based on the information gain of a previous human posture does not, in general, maximize the amount of new information under the current human posture. Second, computing an information gain involves object reconstruction. Object (i.e., human-body) reconstruction is affected by human posture because some body parts may be occluded. Hence, an information gain may not accurately represent the amount of new information for a viewpoint.

In this paper, instead of using information gain and estimating human poses from human-body reconstruction,

Kai-Chi Chan, Cheng-Kok Koh and C. S. George Lee are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, U.S.A. {chan56, chengkok, csglee}@purdue.edu

†This work was supported in part by the National Science Foundation under Grants CNS-0958487 and IIS-0916807. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

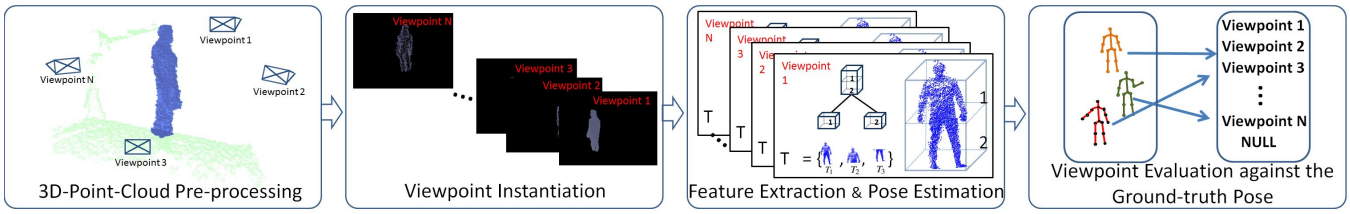

Fig. 2: The overall schematic flow of the proposed viewpoint-evaluation framework using N potential viewpoints in the training phase.

we propose a two-phase approach to determine the best viewpoint for estimating human poses, assuming that the number of potential viewpoints is fixed and the potential viewpoints are given. Hence, the best viewpoint can be determined by examining all the potential viewpoints. Following the proposed approach, an object (human) whose pose being estimated can be rigid or articulated. Also, object reconstruction is not required. Experimental results showed that the proposed two-phase approach could determine the best viewpoint for each human pose, and hence reduce the mean error and standard deviation in human-pose estimation.

II. Proposed Two-Phase Approach

The proposed two-phase approach consists of training and estimation phases. In the training phase, we propose a viewpoint-evaluation framework to train the mapping between a human pose directly to the best viewpoint of a depth sensor. The best viewpoint is determined based on the accuracy of human-pose estimates, without reconstructing a human body. Multiple depth sensors at potential viewpoints are used to capture the 3D point cloud of a person. The quality of a potential viewpoint for human-pose estimation is evaluated in terms of the accuracy of human-pose estimates by matching the geometric VISH feature [10] from different viewpoints. The best viewpoint is then determined by selecting the viewpoint with the lowest average error in human-pose estimation.

In the estimation phase, we use a single depth sensor that is mounted on a humanoid robot. At the current viewpoint, a human pose is first estimated. The human-pose estimate is then used as input to the mapping that has been trained in the training phase to determine the best viewpoint for that human-pose estimate. Once the best viewpoint is determined, the robot can then be moved to the best viewpoint to estimate the human pose. We shall describe the two phases in more detail next.

A. Training Phase: Viewpoint-Evaluation Framework

Due to the mobility of robots/sensors, existing methods that are based on human-body reconstruction require an update of human-body shape in calculating an information gain. Since a human is highly articulated, modeling human-body shape directly is intractable. Simply assuming independence among body parts, however, cannot reflect the reality, and hence can result in modeling errors. Thus, instead of generating a human-body shape for each update, we propose a viewpoint-evaluation framework to measure the quality of potential viewpoints and directly establish the mapping between a human pose and the best viewpoint of a depth sensor for that human pose using the viewpoint quality. As the proposed framework is derived based on a supervised learning algorithm, we build a human-pose database that consists of a set of human poses and the corresponding 3D point clouds at potential viewpoints. The mapping can be considered as a discriminative/conditional relationship that does not require modeling the joint distribution of human-body shape and the best viewpoint.

Figure 2 shows the overall schematic flow of the proposed viewpoint-evaluation framework, which has four main stages: 1) 3D-point-cloud pre-processing, 2) viewpoint instantiation, 3) feature extraction and pose estimation, and 4) viewpoint evaluation. The first three stages process 3D point clouds to estimate human poses, and the final stage determines the mapping between a human pose and the best viewpoint of a depth sensor. We next describe the function of each stage.

1) Pre-processing of a 3D Point Cloud: 3D point clouds are pre-processed such that the 3D points corresponding to a human are extracted and outliers are removed to retain only 3D points that are of interest. We assume that the person, whose human pose is to be estimated, performs different actions in a predefined 3D region. Therefore, we first remove the 3D points outside the region, and the removal process is done by examining the 3D coordinates of points to determine if the points lie in the region. Because of the measurement noise from the depth sensor, we apply spatial smoothing by removing the outliers based on the Euclidean distance between neighboring points. We assume that the Euclidean distance between two neighboring points in the 3D point cloud follows a Gaussian distribution, and 3D points that deviate from the mean of the Gaussian distribution more than one standard deviation are considered as outliers and are removed. The resulting point cloud after removing the outliers is denoted as $\mathscr{P}_H$.

2) Viewpoint Instantiation: As the depth images of the 3D point clouds after pre-processing are different from the depth images directly obtained from the depth sensors, depth images from each potential viewpoint are instantiated for feature extraction. We use the finite projective camera model [11] to represent each depth sensor. The focal length, denoted as f, the skew parameter, denoted as s and the principal point, denoted as (p_x, p_y), of a depth sensor are

measured and stored in the camera calibration matrix, denoted as K, as follows:

$$K = \begin{pmatrix} fm_x & s & m_x p_x \\ 0 & fm_y & m_y p_y \\ 0 & 0 & 1 \end{pmatrix}, \quad (1)$$

where m_x and m_y are the numbers of pixels per unit distance in image coordinates along the x- and y-direction respectively.

The orientation and the center of the camera coordinate frame of the depth sensor can be represented by a 3×3 rotation matrix and a 3×1 position vector, respectively, in the world coordinate frame. Then, the mapping between a point in the world coordinate frame and an image point in the camera coordinate frame can be represented as follows:

$$\mathbf{x} = KR \begin{pmatrix} 1 & 0 & 0 & -x_c \\ 0 & 1 & 0 & -y_c \\ 0 & 0 & 1 & -z_c \end{pmatrix} \mathbf{X}, \quad (2)$$

where $\mathbf{x}$ is the homogeneous coordinates of an image point in the camera coordinate frame and $\mathbf{X}$ is the homogeneous coordinates of a point in the world coordinate frame, K is the camera calibration matrix, R is the 3×3 rotation matrix, and (x_c, y_c, z_c) are the coordinates of the camera center in the world coordinate frame.

Figure 2(b) shows an example of viewpoint instantiation. In the figure, depth images of the point clouds $\mathscr{P}_H$ at potential viewpoints are shown. For each point in the point cloud $\mathscr{P}_H$, the corresponding image point in a depth image from a viewpoint can be calculated using Eq. (2). Each depth image from a potential viewpoint is generated by aggregating the image points with intensity values equal to the distance between the points in the point cloud $\mathscr{P}_H$ and the depth sensor at that viewpoint along the principal axis of the sensor. For those points in the point cloud $\mathscr{P}_H$ that project on the same image point, only the point closest to the viewpoint is selected.

3) Feature Extraction and Pose Estimation: The geometric VISH feature [10] is extracted from the depth image of the point cloud $\mathscr{P}_H$ for human-pose estimation. While extracting the VISH feature, the point cloud $\mathscr{P}_H$ is partitioned and replicated into a tree structure as nodes. VFH [12] and shape features are extracted from each node in the tree to provide a descriptor to represent each node. As the features are obtained based on histograms, coarse-level detail is highlighted in large regions and fine-level detail is highlighted in small regions. Therefore, the features from the point clouds in the tree can capture coarse-to-fine information. A human pose is then estimated by matching the VISH feature with the VISH features of other depth images from the same viewpoint in the human-pose database using the k-nearest neighbor algorithm (k-NN) [13].

4) Viewpoint Evaluation: The mapping between a human pose and the best viewpoint, denoted as M, is a function that is defined as follows:

$$M : P \rightarrow V, \quad (3)$$

where P is a set of human poses and V is a set of potential viewpoints.

The quality of a viewpoint v for a human pose p, denoted as $Q_{v,p}$, depends on the costs from all the joints in the human pose p. The cost of a joint j is calculated based on the Euclidean distance between the joint position estimated by k-NN and the ground-truth position of that joint. As the feature of the human pose p is less similar to the features of the human-pose estimates with lower rank in k-NN, the cost is weighted to reflect the ranking of human-pose estimates. Mathematically, the weighted cost of the joint j in the human pose p at viewpoint v is given by

$$c_{j,p,v,r} = \frac{1}{r} \|\mathbf{j}_p - \tilde{\mathbf{j}}_{p,v,r}\|_2, \quad (4)$$

where r is the rank of the human pose estimated by k-NN, $\mathbf{j}_p$ is the ground-truth position of the joint j in the human pose p, $\tilde{\mathbf{j}}_{p,v,r}$ is the position of the joint j in the rth human-pose estimate of the human pose p found by k-NN using the features extracted at viewpoint v, and $\|\cdot\|_2$ is the Euclidean norm. Hence, the weighted cost can measure the accuracy of human-pose estimates and reflect the quality of a viewpoint.

For a more accurate human-pose estimate at a particular viewpoint, the weighted cost is lower compared with the weighted costs of joints from the human-pose estimates with the same rank at other viewpoints. Therefore, the quality of a viewpoint $Q_{v,p}$ is defined in terms of the weighted costs as follows:

$$Q_{v,p} = -E[c^2_{j,p,v,r}], \quad (5)$$

where $E[\cdot]$ is the expectation operator over all the joints in the human poses estimated by k-NN. The quality $Q_{v,p}$ is the negation of the mean-squared error of the weighted costs. It can be expressed as

$$Q_{v,p} = -E[c_{j,p,v,r}]^2 - Var(c_{j,p,v,r}), \quad (6)$$

where $Var(\cdot)$ is the variance operator.

The weighted cost is a function of the observations of the depth images generated from different viewpoints. Given a viewpoint, we can compute the quality of the viewpoint using Eq. (6). All the viewpoints with quality higher than or equal to a predefined threshold are compared and the best viewpoint of a human pose is then found by maximizing the quality of viewpoints. Hence, the mapping between the human pose p and the best viewpoint is given by

$$M(p) = \begin{cases} \arg\max_{v\in V} Q_{v,p}, & \text{if } \max_{v\in V} Q_{v,p} \geq \tau; \\ null, & \text{otherwise,} \end{cases} \quad (7)$$

where τ is a predefined threshold and *null* means that the best viewpoint does not exist for that human pose as no viewpoint has met the threshold.

The derivation of the mapping M is summarized in the supervised learning algorithm as illustrated in Algorithm 1.

Algorithm 1 Derivation of the mapping between a human pose and the best viewpoint.

Require:
P: Set of Human Poses
V: Set of Potential Viewpoints
PC: Set of Point Clouds
for $i := 1$ to $|P|$ **do**
  $p := P(i)$
  $pc := PreProcess(PC(i))$
  $Pose := EmptyList()$
  for $v \in V$ **do**
    $I := DepthImageGeneration(pc, v)$
    $f := FeatureExtraction(I)$
    $Pose(v) := PoseEstimate(f)$
  end for
  $BestViewpoint(Pose, p, V)$
end for

B. Estimation Phase: Best-Viewpoint Determination

During the estimation phase, a single depth sensor, which is mounted on a robot, is used to estimate human poses. The best viewpoint of the depth sensor can be determined by the mapping M in which the output varies among different human poses. Therefore, a human pose is first estimated based on the VISH features. The best viewpoint is then found by the mapping M using the estimated human pose.

To estimate human poses, a 3D point cloud captured by the depth sensor mounted on a robot is pre-processed to remove the outliers in the 3D point cloud and extract the 3D point cloud corresponding to the human. The VISH feature is then extracted from the 3D point cloud of the human and matched with the VISH features from the human-pose database which contains both the 3D point clouds of the human and the corresponding ground-truth human poses. Mathematically, the human-pose estimate, denoted as $\hat{p}$, can be found by

$$\hat{p} = \arg\min_{p \in P} \|f - f_{extract}(\mathscr{C}^p)\|_2, \tag{8}$$

where f is the VISH feature extracted from the 3D point cloud of the human, $\mathscr{C}^p$ is a 3D point cloud in the human-pose database, $f_{extract}(\cdot)$ is a function that takes a 3D point cloud $\mathscr{C}^p$ as input and outputs the VISH feature extracted from the 3D point cloud $\mathscr{C}^p$, p is the ground-truth human pose corresponding to the 3D point cloud $\mathscr{C}^p$, P is the set that contains all ground-truth human poses in the human-pose database, and $\|\cdot\|_2$ is the Euclidean norm.

As the human-pose estimate $\hat{p}$ may not be in the domain of the mapping M, the human-pose estimate $\hat{p}$ is represented by the closest match of human poses in the domain of the mapping M. Let $\tilde{p}$ be the closest match which is found by

$$\tilde{p} = \arg\min_{p \in P^M} \|\hat{p} - p\|_2, \tag{9}$$

where p is a human pose in the domain of the mapping M and P^M is the domain of the mapping M.

The human-pose estimate $\tilde{p}$ can then be used to determine the best viewpoint through the mapping M. The best viewpoint, denoted as v_{best}, is given by

$$v_{best} = M(\tilde{p}). \tag{10}$$

III. Experimental Results

The proposed two-phase approach was evaluated on the Berkeley Multimodal Human Action Database (MHAD) [14]. In the database, a subject performing an action was captured by two Microsoft Kinect sensors. Only depth images, which were captured by the depth sensors in the Microsoft Kinect sensors, were used throughout the experiment. The depth images were captured at 30 frames per second at a resolution of 640×480 pixels. The ground-truth 3D joint locations of the subjects were recorded by a motion-capture system. The error metric, ζ, for each image is defined as

$$\zeta = \frac{1}{N_f} \sum_{s=1}^{N_f} \frac{1}{N_s} \sum_{i=1}^{N_s} \|\mathbf{l}_{s,i} - \tilde{\mathbf{l}}_{s,i}\|_2, \tag{11}$$

where N_f is the number of depth images for testing, N_s is the number of 3D joint locations measured by the motion-capture system at the sth image, $\mathbf{l}_{s,i}$ is the ground-truth 3D location of the ith joint at the sth image and $\tilde{\mathbf{l}}_{s,i}$ is the estimated 3D location of the ith joint at the sth image.

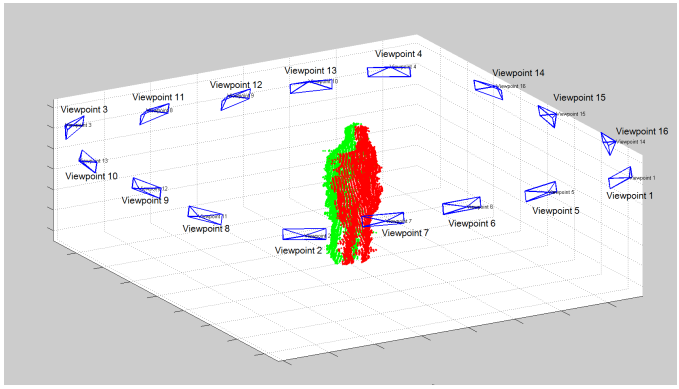

Fig. 3: The viewpoints of depth sensors in the training phase.

In the pre-processing stage, the predefined 3D region with size $17.5 \times 20 \times 7.5\ m^3$ was set. In the feature-extraction stage, two levels were used in the tree structure, and six child nodes were split from the root node in the tree. The size of the 2D region for extracting the shape features was set to be 17×17 pixels. To evaluate the performance of the proposed two-phase approach, the domain of the mapping M was set to the set of ground-truth human poses in the human-pose database. The range of the mapping M was set as a collection of the sixteen viewpoints as shown in Fig. 3. Two Microsoft Kinect sensors were located at viewpoints 1 and 3 in the training phase. The other viewpoints were constructed using the finite projective camera model [11]. Under this setting, viewpoints were located around the subject as shown in the figure. The red points and the green points represented the 3D points captured from the front and the back of the subject

from viewpoints 1 and 3, respectively. From each viewpoint, human poses were estimated using k-NN and k was set to 1. One percent of the human-pose database were used as the samples in calculating the weighted costs. The threshold τ in the mapping M was set to be the summation of the mean of the weighted costs and three times the standard deviation of the weighted costs.

We assumed that a humanoid robot with a depth sensor was located at one of the sixteen viewpoints at the beginning. As this work did not involve human-pose prediction and moving the robot to the best viewpoint, we assumed that the humanoid robot could move instantaneously to any of the other viewpoints. In other words, the human pose remained the same while the humanoid robot moved to different viewpoints. A human pose was first estimated at the current viewpoint so that it could be used as input to the mapping M to determine the best viewpoint for that human-pose estimate. The humanoid robot was then "moved" to the best viewpoint instantaneously to estimate the human pose. The error between the human-pose estimate at the best viewpoint and the ground-truth human pose was then calculated using Eq. (11). Each test was verified using 5-fold cross-validation.

Table I shows the mean error and standard deviation of human-pose estimation on a subject using a fixed viewpoint and the best viewpoints determined by the proposed two-phase approach. The humanoid robot was located at an original viewpoint at the beginning. Using the fixed viewpoint, the humanoid robot remained stationary at the original viewpoint and human poses were estimated using the VISH features [10]. The errors of human-pose estimates were calculated using Eq. (11). The second and third columns of Table I showed the mean errors and standard deviations of the human-pose estimates at different fixed viewpoints. Using the two-phase approach, the humanoid robot was "moved" to the best viewpoint based on the mapping M. Thus, human poses were estimated at their best viewpoints. The fourth and fifth columns showed the mean errors and standard deviations of the human-pose estimates at the best viewpoints. Both mean errors and standard deviations were decreased for all the viewpoints. The maximum reduction of mean error occurred at viewpoint 1. The mean error was reduced by about 26%. The maximum reduction of standard deviation occurred at viewpoint 5. The standard deviation was reduced by about 26%. The statistical results showed that the proposed two-phase approach could select the viewpoints with more accurate human-pose estimates among the sixteen viewpoints and reduce the error in human-pose estimation.

Figure 4 shows the best viewpoints determined by the mapping M in the two-phase approach for some human poses estimated at four viewpoints (viewpoints 1, 2, 3 and 4) for illustration. The leftmost column shows the 3D point clouds of a subject performing different actions and the corresponding ground-truth 3D human poses. The other columns show the depth images of the 3D point clouds from different viewpoints and the corresponding human-pose estimates. The depth images and human-pose estimates

TABLE I: The mean errors and standard deviations (in centimeters) of human-pose estimates for a subject after applying the proposed two-phase approach.

Original Viewpoint	Fixed Viewpoint		Best Viewpoint	
	Mean Error	Std. Dev.	Mean Error	Std. Dev.
1	44.36	16.44	32.88	12.50
2	38.79	13.82	31.83	11.52
3	42.78	15.11	31.72	11.56
4	39.19	13.58	31.81	10.60
5	39.11	14.21	30.60	10.46
6	38.30	12.50	32.46	13.80
7	37.63	12.57	31.75	10.62
8	39.36	13.97	32.77	12.15
9	36.91	11.96	31.41	11.05
10	38.71	13.88	31.64	11.69
11	41.15	15.59	35.03	13.57
12	37.04	12.47	35.02	14.31
13	35.94	11.65	32.26	11.03
14	37.14	12.04	31.65	11.98
15	36.71	12.86	31.87	11.13
16	38.18	13.20	33.07	12.19

from the best viewpoints determined by the mapping M are highlighted. For comparison purpose, ground-truth 3D human poses and human-pose estimates are translated and rotated such that all joint positions are referenced from the same coordinate frame. In the figure, some of the human-pose estimates for the human poses in the first two rows were in actions different from the ground-truth human poses. For example, the ground-truth human pose in the first row was raising both arms but the human-pose estimate from viewpoint 3 was lowering both arms downwards. The human pose in the second row was raising one arm but the human-pose estimates from viewpoints 1 and 4 were raising both arms. It showed that the accuracy of human-pose estimates was affected by the human poses of the subject and the viewpoints of the sensor. Thus, switching among the viewpoints is important in reducing the error in human-pose estimation.

In the first row, viewpoint 1 was selected as the best viewpoint. Among the four viewpoints, the human-pose estimates from the viewpoints 1, 2 and 4 were similar to the ground-truth human pose. The error of the human-pose estimate from viewpoint 1 was the lowest. It showed that the mapping M in the two-phase approach was able to select the best viewpoint. In the second row, the human-pose estimates from the viewpoints 2 and 3 were similar to the ground-truth human pose. The error incurred by the human-pose estimate from viewpoint 2 was the lowest, and the proposed approach selected viewpoint 2 as the best viewpoint.

In the third row, all the human-pose estimates were similar to the ground-truth human pose. Among the four human-pose estimates, the error of the human-pose estimate from viewpoint 3 was the lowest. The mapping M could identify the human-pose estimate and selected viewpoint 3 as the best viewpoint. In the last row, the human-pose estimate from viewpoint 4 was closest to the ground-truth human pose. However, the mapping M selected both viewpoints 1 and 4 as the best viewpoint for the ground-truth human pose because the nearest neighbors of the ground-truth human pose from

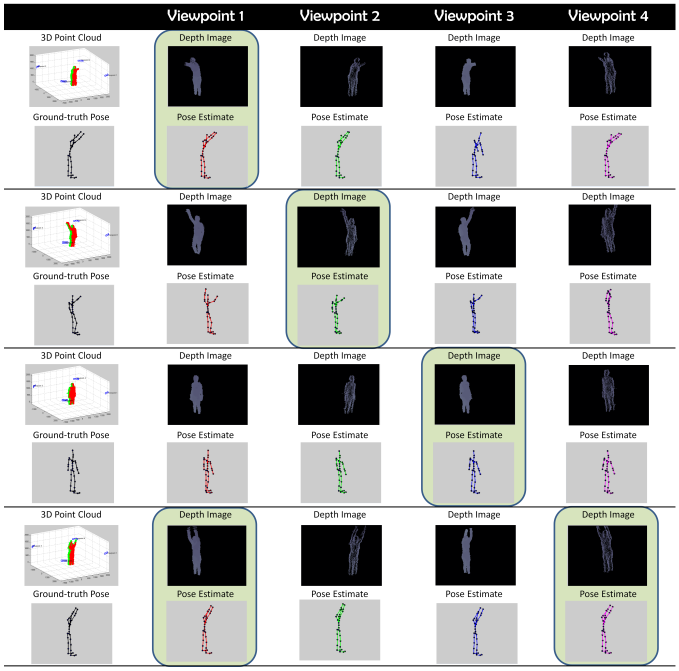

Fig. 4: Best viewpoints of some human poses determined by the mapping M in the proposed viewpoint-evaluation framework using the two-phase approach.

the two viewpoints in the training data were different. In addition, the best viewpoints determined by the mapping M for the two nearest neighbors were viewpoints 1 and 4. It showed that the human-pose estimates from viewpoints 1 and 4 were very close to each other. Thus, for similar ground-truth human poses in the training data, the errors of the human-pose estimates from viewpoints 1 and 4 were similar. By allowing more than one viewpoint as the best viewpoint, there can be potentially more feasible solutions when combining the sensor movement constraint with other constraints such as motion planning.

IV. Conclusions and Future Work

In this paper, we proposed the two-phase approach to determine the best viewpoint for human-pose estimation. Since human posture could be changed over time, existing methods based on an information gain had difficulties in updating a human-body shape. The proposed approach can determine the best viewpoint for each human pose directly without human-body reconstruction.

In the training phase, we proposed the viewpoint-evaluation framework to measure the quality of potential viewpoints for human-pose estimation. The proposed framework consisted of four stages: 3D-point-cloud preprocessing, viewpoint instantiation, feature extraction and pose estimation, and viewpoint evaluation. The 3D point cloud of a person captured by depth sensors was first extracted and filtered. The viewpoints of the depth sensors were instantiated using the finite projective camera model. The VISH features were then extracted from the depth images generated from the instantiated viewpoints. Each viewpoint was evaluated for every human pose estimated by k-NN based on the matching of the VISH features.

In the estimation phase, we used a single depth sensor that was mounted on a humanoid robot. A human pose was first estimated at the current viewpoint and the estimate was then used as input to the mapping that determined the best viewpoint for that human-pose estimate. Once the best viewpoint was found, the robot was moved instantaneously to the best viewpoint to estimate the human pose.

Experimental results showed that different viewpoints would affect the accuracy of human-pose estimates. The mapping in the proposed framework could identify the best viewpoint for each human pose and hence reduce the mean error and standard deviation of the human-pose estimate. The maximum reductions of the mean error and standard deviation were about 26% and 26%, respectively.

In the future, we will determine the optimal number of potential viewpoints by studying the relationship between the accuracy of human-pose estimation and the number of potential viewpoints. We will also study how to predict human poses after the time delay for a robot to move among different viewpoints because a human could be disturbed and change his/her pose while a robot is moving to the optimal viewpoint. We will also extend the proposed approach with multiple sensors.

References

[1] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun. Real-time identification and localization of body parts from depth images. In *ICRA*, pages 3108–3113, May 2010.

[2] J. Gall, A. Yao, and L. van Gool. 2D action recognition serves 3D human pose estimation. In *Proceedings of the 11th European conference on computer vision*, pages 425–438, Berlin, Heidelberg, 2010. Springer-Verlag.

[3] L. Sigal, M. Isard, H. Haussecker, and M.J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *IJCV*, 98(1):15–48, May 2012.

[4] 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 *CVPR*, pages 1297–1304, Jun. 2011.

[5] K. C. Chan, C. K. Koh, and C. S. G. Lee. Using action classification for human-pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1176–1181, Nov. 2013.

[6] K.A. Tarabanis, P.K. Allen, and R.Y. Tsai. A survey of sensor planning in computer vision. *IEEE Trans. Robot. Autom.*, 11(1):86–104, 1995.

[7] W.R. Scott, G. Roth, and J.F. Rivest. View planning for automated three-dimensional object reconstruction and inspection. *ACM Computing Surveys*, 35(1):64–96, Mar. 2003.

[8] M.A. Sipe and D. Casasent. Feature space trajectory methods for active computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1634–1643, 2002.

[9] M. Krainin, B. Curless, and D. Fox. Autonomous generation of complete 3D object models using next best view manipulation planning. In *ICRA*, pages 5031–5037, May 2011.

[10] K. C. Chan, C. K. Koh, and C. S. G. Lee. A 3D-point-cloud feature for human-pose estimation. In *IEEE International Conference on Robotics and Automation*, pages 1615–1620, May 2013.

[11] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

[12] R.B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3D recognition and pose using the viewpoint feature histogram. In *IROS*, pages 2155–2162, Oct. 2010.

[13] R.O. Duda, D.G. Stork, and P.E. Hart. *Pattern classification and scene analysis. Part 1, Pattern classification*. Wiley, 2nd edition, Nov. 2000.

[14] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley MHAD: A comprehensive multimodal human action database. In *IEEE Workshop on Applications of Computer Vision*, pages 53–60, Jan. 2013.