A 3-D-Point-Cloud System for Human-Pose Estimation

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

***Abstract*—This paper focuses on human-pose estimation using a stationary depth sensor. The main challenge concerns reducing the feature ambiguity and modeling human poses in high-dimensional human-pose space because of the curse of dimensionality. We propose a 3-D-point-cloud system that captures the geometric properties (orientation and shape) of the 3-D point cloud of a human to reduce the feature ambiguity, and use the result from action classification to discover low-dimensional manifolds in human-pose space in estimating the underlying probability distribution of human poses. In the proposed system, a 3-D-point-cloud feature called viewpoint and shape feature histogram (VISH) is proposed to extract the 3-D points from a human and arrange them into a tree structure that preserves the global and local properties of the 3-D points. A nonparametric action-mixture model (AMM) is then proposed to model human poses using low-dimensional manifolds based on the concept of distributed representation. Since human poses estimated using the proposed AMM are in discrete space, a kinematic model is added in the last stage of the proposed system to model the spatial relationship of body parts in continuous space to reduce the quantization error in the AMM. The proposed system has been trained and evaluated on a benchmark dataset. Computer-simulation results showed that the overall error and standard deviation of the proposed 3-D-point-cloud system were reduced compared with some existing approaches without action classification.**

***Index Terms*—3-D-point-cloud feature, action classification, human-pose estimation.**

I. Introduction

HUMAN-pose estimation is the process of determining human-joint positions from a single or multiple sensors. It can benefit a wide range of applications such as human-motion analysis, healthcare, control and surveillance. For example, estimating human poses can provide valuable information, such as body-part positions, in correcting body postures in sport training and in identifying a potential fall in healthcare facilities. Traditionally, human poses are estimated using one or more charge-coupled device (CCD) cameras. Visual features, such as colors, edges, silhouettes, and textures, are used to represent human poses. However, those features are ambiguous because depth information is lost during the 3-D-to-2-D projection. For example, the same silhouette could correspond to different human poses. Thus, visual features are commonly used to form geometric features by estimating depth information using stereo vision [1]. Common geometric features for human bodies are 3-D surfaces [2], superellipsoids [3], [4], volumetric body scans [5], and metaballs [6]. Although visual features are commonly used and are useful for reconstructing depth information, they are ambiguous under different illumination conditions. For example, the color of an image of a person may change drastically under different lighting conditions. Thus, the precision of estimating depth information is decreased when visual features are ambiguous. The estimated depth information will then affect the performance of estimating human poses.

The availability of depth sensors allows us to obtain depth information directly and with less ambiguity. The depth sensor outputs a 3-D point cloud, which is a set of 3-D coordinates on the surface of objects. Thus, depth information will not be affected by the quality of visual features. Using a depth sensor, geometric features [7], [8] are extracted by comparing the depth at nearby pixels from the depth images captured by the depth sensor. Rusu *et al.* [9] proposed the viewpoint feature histogram (VFH), which measures the pan, tilt and yaw angles between 3-D points. Plagemann *et al.* [10] proposed the accumulative geodesic extrema (AGEX) by extracting the geodesic distances from pairs of points on a human body. Ye *et al.* [11] used a 3-D surface mesh to represent a human body, and to estimate human poses by pose detection and pose refinement. Since the geometric features are extracted from the whole human body, a body's properties, such as orientation and shape, in local regions are understated. It motivates us to investigate and derive a 3-D-point-cloud feature that captures the body's properties, namely orientation and shape, in both global and local regions of the 3-D point cloud of a human body. Throughout the paper, we will use the terms global and local properties to represent the body's properties in global and local regions, respectively.

Once a feature is extracted from the 3-D point cloud of a human, a human-pose model that represents the underlying probability distribution of human poses in human-pose space is trained. Due to the curse of dimensionality [12], the amount of training data needed grows exponentially with the dimensionality of human-pose space. However, the amount of training data is usually limited (fixed). Thus, for a fixed amount of training data, discovering low-dimensional manifolds that

Manuscript received March 31, 2013; revised August 16, 2013 and November 30, 2013; accepted December 21, 2013. Date of publication July 14, 2014; date of current version October 13, 2014. This work was supported by the National Science Foundation under Grant CNS-0958487, Grant CNS-0960061, and Grant IIS-0916807. Any opinion, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. This paper was recommended by Associate Editor V. Piuri.

The authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: chan56@purdue.edu; chengkok@purdue.edu; csglee@purdue.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2014.2329266

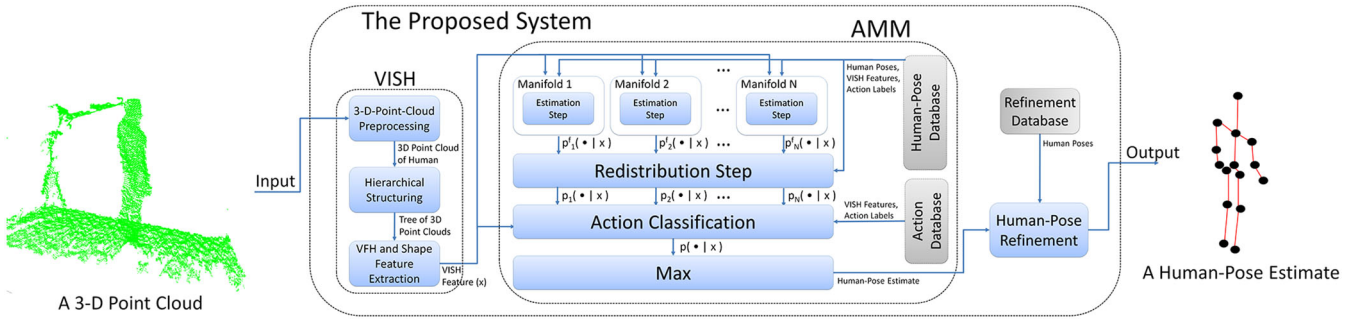

Fig. 1. Proposed 3-D-point-cloud system for human-pose estimation. The system uses the proposed VISH feature as input to capture the global and local properties of 3-D points from a human. It then uses the proposed nonparametric AMM and the refinement model (a kinematic model) to represent the human-pose space using multiple low-dimensional manifolds and to refine human-pose estimates, respectively. The action and refinement databases are prepared in advance for training the action classifier and the kinematic model, respectively. The human-pose database is prepared in advance to represent the discrete space of human poses in each manifold.

represent human-pose space could produce a more accurate human-pose model. For example, low-dimensional manifolds can be discovered by dimensionality-reduction methods, such as locality preserving projection algorithm [13]. Furthermore, human poses in high-dimensional space can be projected to low-dimensional manifolds for estimating human poses [14]–[16]; unfortunately, the spatial relationship in human poses could be lost during the projection.

Instead of using dimensionality-reduction methods, prior knowledge, such as body shape [17]–[20] and kinematic relationship [21]–[23], can be used to discover the low-dimensional manifolds in human-pose space. Furthermore, human action has been used recently as prior knowledge [24]. Gall *et al.* [25] proposed a model for estimating human poses by determining the prior probability of actions from action classification. The human pose in a previous frame was used as an initialization for finding the optimal human pose and action jointly using a particle-based annealing optimization scheme. Yao *et al.* [24] proposed the appearance-based and pose-based features to classify actions using Hough forest algorithm [26]. Yao *et al.* [27] further extended the model of estimating the prior probability of actions in [25] by incorporating the action classification in [24] into a single framework. Since, in general, a human pose can appear in more than one action, for example, the human pose of hand waving can appear in the actions of standing and raising both arms, we extend this concept in our system using an action-mixture model (AMM).

In this paper, we propose a 3-D-point-cloud system that uses a 3-D-point-cloud feature as input to capture the global and local properties of the 3-D point cloud of a human. It then uses the result from action classification and a kinematic model to represent human-pose space using low-dimensional manifolds in estimating human poses and to refine human-pose estimates, respectively. To avoid accumulating errors from previous frames, our proposed system uses temporal information only for training but not for testing. We propose a 3-D-point-cloud feature, called VISH [28], which is derived based on the VFH feature. The novelty in our proposed feature is the spatial ordering of orientation and shape in both global and local regions of the 3-D points from a human. The spatial ordering is important in resolving the ambiguity of symmetric human poses. A nonparametric AMM is proposed to model human poses using the concept of distributed representation. The human-pose space is represented by low-dimensional manifolds, each of which corresponds to one action. The proposed model is different from the previous work [24], [25] in that a human pose may appear in more than one action. In the kinematic model, the angle of each body part is parameterized by a quaternion [29] to explicitly represent the spatial relationship between body parts. The proposed system was tested on the Stanford TOF Motion Capture Dataset [23]. Computer simulations showed that the proposed system reduced the overall error and standard deviation of human-pose estimates compared with some existing approaches.

The contributions of this paper include the following.

1) Deriving the VISH feature from a 3-D point-cloud to capture the global and local properties of 3-D points from a human.
2) Proposing a nonparametric AMM that represents human-pose space using low-dimensional manifolds associated with actions to yield a better estimate of the underlying probability distribution of human poses.
3) Estimating human poses automatically without temporal information to avoid accumulating errors of human-pose estimates in previous frames.

The structure of this paper is as follows. Section II describes the proposed system framework. Section III discusses the proposed VISH feature, which represents the 3-D point-cloud of a human. Section IV presents the proposed nonparametric AMM. Section V describes a kinematic model in refining the human poses estimated by the AMM. Computer-simulation results are discussed in Section VI, and conclusions are presented in Section VII.

II. System Framework

The proposed 3-D-point-cloud system takes a 3-D point cloud as input and produces a human-pose estimate as output. Fig. 1 shows the proposed system framework. From the 3-D-point-cloud input, a human, and the corresponding 3-D points are detected. The proposed VISH feature is extracted from the 3-D point cloud of the human. During the extraction, the 3-D point cloud is partitioned and replicated into 3-D regions in a tree to capture the spatial ordering of orientation

and shape in the 3-D point cloud. The orientation and shape are represented by the VFH and shape features, respectively. As the features are obtained based on histograms, coarse-level details are highlighted in large regions and fine-level details are highlighted in small regions. The VISH feature is then formed by combining the VFH and shape features. Details of the proposed VISH feature are described in Section III.

Once the VISH feature is extracted, human poses are modeled to estimate their underlying probabilities in human-pose space. As humans are highly articulated, the amount of training data needed to model human poses grows exponentially with the dimensionality of human-pose space. To better model human poses with a fixed amount of training data, we use the proposed nonparametric AMM to represent human-pose space using multiple low-dimensional manifolds based on the concept of distributed representation, where each manifold corresponds to one action. A human-pose database, which contains the ground-truth human poses, the corresponding action labels, and VISH features, is created to represent all possible human poses in the human-pose space. In each manifold, the probability distribution of human poses is estimated in two steps: estimation and redistribution steps. The estimation step calculates the Euclidean distance between the VISH-feature input and the VISH features from the human-pose database, and produces a probability distribution of human poses as output in each manifold. The redistribution step assigns weights to the probability distributions estimated from the estimation step according to the frequency of actions associated with the manifolds. Then, action classification is performed on the VISH-feature input to determine the weighting coefficients in combining the low-dimensional manifolds to represent the underlying probability distribution of human poses in the human-pose space. The human pose with the highest probability will be considered as the human-pose estimate corresponding to the VISH-feature input. The proposed nonparametric AMM is described in more details in Section IV.

In human-pose refinement, the human pose estimated by the proposed AMM is refined in continuous space to reduce the quantization error in the proposed AMM. The spatial relationship between body parts is used to refine human-pose estimates in a kinematic model. The body parts are modeled using their position and orientation, which can be computed based on the joint positions of the human-pose estimate. A refinement database is created to train the kinematic model that represents the spatial relationship between body parts. A more detailed description of the kinematic model is given in Section V.

III. Viewpoint and Shape Feature Histogram (VISH)

Histograms of oriented gradients (HOG) [30] is a common visual feature, based on the orientation histograms obtained from 2-D color/grayscale images, for human detection and human-pose estimation [31], [32]. Since HOG is computed based on gradients, it is to some extent invariant to the change in illumination. The proposed VISH measures the orientation and shape responses from a 3-D point cloud and hence can be considered as a 3-D adaptation of HOG. From the 3-D-point-cloud input of the proposed system, VISH features are extracted to represent the distributions of 3-D points of the person of interest. To handle the large number of 3-D points in a 3-D point cloud, the distribution of the 3-D points is summarized into a nonparametric distribution using VFH and shape extractors. The summarization is performed on overlapping 3-D regions of the 3-D point cloud so that the spatial ordering can be captured.

There are three steps to extract VISH from a 3-D point cloud, namely 3-D-point-cloud preprocessing, hierarchical structuring, and feature extraction. In the preprocessing of 3-D point cloud, 3-D points corresponding to a human are extracted and outliers are removed to retain the 3-D points of interest. This step is important to reduce the number of 3-D points by keeping only the 3-D points from the human for further processing. In the hierarchical structuring, the preprocessed 3-D point cloud is partitioned and replicated into a tree structure as nodes. VFH and shape features are extracted from each node in the tree to provide a descriptor to represent each node. Therefore, the features from the point cloud in the tree can capture coarse level to fine level information.

A. Three-Dimensional-Point-Cloud Preprocessing

We assume that the person, whose human pose is to be estimated, performs different actions in a predefined 3-D region. Therefore, we first remove the 3-D points outside the region because they are not from the person. The removal process is done by examining the 3-D coordinates of points to determine if points lie in the region.

Let $\mathcal{P}_A$ be the set of 3-D points in the region. Because of the measurement noise from the depth sensor, there are outliers in the 3-D points cloud $\mathcal{P}_A$. Smoothing or filtering techniques [33] can be used to remove the outliers by smoothing the spatial and temporal information of the 3-D points. Since our system does not consider the temporal information of the person during testing, 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 3-D point cloud $\mathcal{P}_A$ follows a Gaussian distribution. Thus, 3-D points, whose deviations from the mean of the normal distribution are larger than one standard deviation, are considered as outliers and removed. The resulting point cloud after removing the outliers is denoted as $\mathcal{P}_H$.

B. Hierarchical Structuring

The 3-D point cloud of a human in the predefined region $\mathcal{P}_H$ is organized into a tree structure such that a node in the tree represents a set of 3-D points in a 3-D region and an edge represents the process of duplicating the 3-D points. In this paper, the 3-D region is represented by a rectangular region (cuboid) but it can be generalized to other shapes such as spheres [34]. Nodes at different levels of the tree capture the spatial ordering of sets of 3-D points. At the zeroth level (root level), the root node represents the 3-D point cloud $\mathcal{P}_H$.

Let M^0 be the set of the smallest cuboid that contains the points in the root node at the zeroth level. Let $S_n(\cdot)$ be the

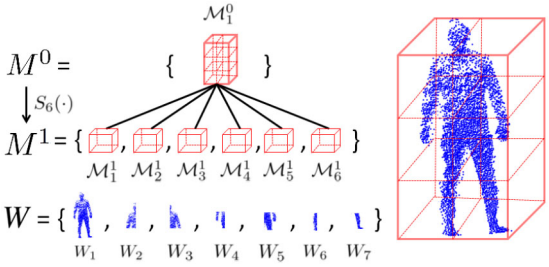

Fig. 2. Example of hierarchical structuring with the height of the tree is one and the number of cuboids split by $S_{n_1}(\cdot)$ is six. The smallest cuboid, which contains the 3-D point cloud $\mathcal{P}_H$ from the person, is shown on the right side. The cuboid is divided into six smaller sub-cuboids with equal volume by $S_{n_1}(\cdot)$. All the cuboids are arranged into a tree structure as shown on the left upper region. Points from the 3-D point cloud $\mathcal{P}_H$ are extracted from each cuboid and grouped together in W for feature extraction.

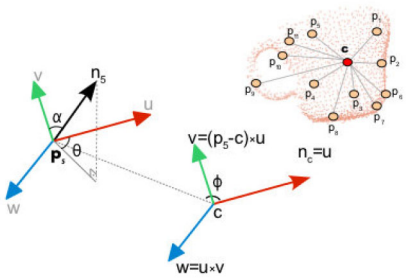

Fig. 3. Relative pan, tilt, and yaw angles between two points in the extraction of VFH feature [9].

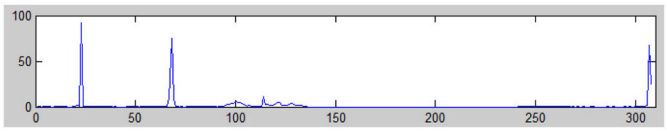

Fig. 4. Example of the VFH feature.

function that splits a cuboid into a set of n exhaustive, continuous, mutually exclusive, and equal-sized cuboids. Let M^i be the set of cuboids returned by the function $S_{n_i}(\cdot)$ at the ith level, where n_i is the number of cuboids returned by the function $S_{n_i}(\cdot)$ at the ith level and n_0 is preset to 1. The jth element in M^i is denoted as $\mathcal{M}_j^i$. The set M^i can be derived by the following recursive formula:

$$M^i = \bigcup_{j=1}^{|M^{i-1}|} S_{n_i}\left(\mathcal{M}_j^{i-1}\right), \quad \forall i = 1, 2, \ldots, h \tag{1}$$

where $\bigcup$ is the set union, $|\cdot|$ is the cardinality of the input set, h is the height of the tree with the convention that the height of the root level is 0.

In each cuboid $\mathcal{M}_j^i$, $i = 0, \ldots, h$, $j = 1, \ldots, |M^i|$, the 3-D points from the 3-D point cloud $\mathcal{P}_H$ are extracted for feature extraction. Let $P_H(\cdot)$ be the function that takes such cuboid as input and outputs a set of 3-D points from the 3-D point cloud $\mathcal{P}_H$ that lies in the input cuboid. In a breadth-first fashion, we apply the function $P_H(\cdot)$ to the cuboid at each level of the tree structure to obtain a set of collections of 3-D points, denoted as W, as follows:

$$W = \bigcup_{i=0}^{h} \bigcup_{j=1}^{|M^i|} \left\{P_H\left(\mathcal{M}_j^i\right)\right\}. \tag{2}$$

To illustrate the idea, Fig. 2 shows an example of the hierarchical structuring when the height of the tree is one and the number of cuboids split (n_1) is six. In the figure, the right side shows the smallest cuboid that contains the 3-D point cloud $\mathcal{P}_H$ from a person. The root node in the tree represents the smallest cuboid as shown on the upper left region. M^0 is the set that contains the cuboid. The cuboid is then divided into six sub-cuboids with equal volume by the function $S_6(\cdot)$. The sub-cuboids are the children of the root node in the tree. M^1 contains the six sub-cuboids. The points from the 3-D point cloud $\mathcal{P}_H$ in each node of the tree are extracted and stored in W for feature extraction. In the example, W has seven elements, denoted as W_i, $i = 1, \ldots, 7$.

C. Feature Extraction

Two features, namely VFH and shape features, are extracted from the set of 3-D points in each cuboid. The VFH feature is originally designed to estimate the six-degree-of-freedom pose of rigid objects. Each 3-D point in the 3-D point cloud in each cuboid is first assigned a direction. The direction of a 3-D point is found by three steps: 1) find the k nearest neighbors of the 3-D point; 2) find the normal to the 2-D plane that contains the 3-D point and minimizes the average perpendicular distance between the nearest neighbors and the plane; and 3) set the normal to the 2-D plane pointing toward the depth sensor as the direction of the 3-D point. The relative pan, tilt, and yaw angles between every pair of 3-D points in the k-nearest neighbors are then computed. Fig. 3 shows an example of the three angles between two points. The angles in the 3-D point cloud are collected in a histogram with 308 bins, as suggested in [9], to form the VFH feature for the 3-D point cloud. Fig. 4 shows an example of a VFH feature. A more detailed description of the VFH feature can be found in [9].

The shape feature measures the pattern or outline of a region from a human. It is measured around the 3-D centroid of the 3-D point cloud in each cuboid. A range image, denoted as I, is created by projecting the 3-D point cloud $\mathcal{X}$ on a 2-D plane such that each pixel value in I represents the distance between the corresponding 3-D point in $\mathcal{X}$ and the depth sensor, with the convention that pixel values outside I are zeros. A window with size $(2w + 1)$ pixels by $(2w + 1)$ pixels is centered at the centroid position in the range image to extract the set of pixels within the window. The shape feature is formed by dividing the pixels within the window by the maximum value among the pixels so that the feature is depth invariant. Fig. 5 shows an example of the shape feature with $w = 8$. When $w = 8$, the dimension of the shape feature in this example is $(2 \times 8 + 1)^2 = 289$. The color at each pixel represents the ratio of the distance between the corresponding 3-D point in the 3-D point cloud and the depth sensor to the maximum pixel value (distance) in I within the window.

Let $\mathbf{f}(\cdot)$ be the mapping from a set of 3-D points to the row vector of VFH and shape features. VISH can be arranged into a row vector $\mathbf{F}$ and is found by concatenating all the features extracted from each element of W; that is

$$\mathbf{F} = (\mathbf{f}(W_1), \mathbf{f}(W_2), \ldots, \mathbf{f}(W_{|W|})) \tag{3}$$

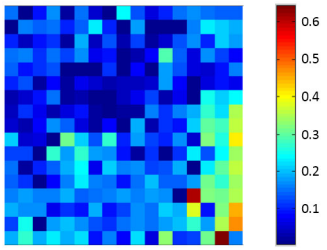

Fig. 5. Example of the shape feature.

where W_i is the ith element in W, and $|W|$ is the size/cardinality of W.

Note that in the tree structure, the 3-D points in a child node are duplicated from its parent node. It provides an illusion that 3-D points are redundant among different levels of the tree. However, VFH and shape features summarize the 3-D points in cuboids at different levels. The features extracted from different levels of cuboids provide different levels of details in the summarization. In addition, the order of the set W provides useful information of spatial ordering of the 3-D points from the human surface.

The time complexity of extracting the VFH, shape, and VISH features is derived as follows. As shown in [9], the time complexity of extracting the VFH feature is $O(n)$, where n is the number of 3-D points. When extracting the shape feature, the parameter w is typically small compared to n. Thus, the time complexity of extracting the shape feature is $O(n)$. As each 3-D point appears exactly $(h+1)$ times in the tree, the time complexity of extracting the VISH feature is $O((h+1)n)$. Computer-simulation results showed that h is typically small compared to n. Thus, the time complexity of the VISH feature extraction can be rewritten as $O(n)$. VISH is asymptotically more efficient than AGEX [10], which has a time complexity $O(n \log n)$.

The VISH feature captures the global and local properties of the 3-D point cloud of a human. It can be used to estimate human poses through a human-pose model such as k-nearest neighbor algorithm (k-NN) [35] and support-vector machine (SVM) [36]. As shown in computer simulations, the VISH feature can reduce the ambiguity of symmetric human poses (Figs. 7 and 8). However, the dimensionality of human-pose space is high and thus the k-NN and SVM do not perform well because of the curse of dimensionality. The proposed nonparametric AMM to be discussed in the next section further increases the accuracy of human-pose estimation by representing the human-pose space using low-dimensional manifolds associated with human actions.

IV. Nonparametric Action-Mixture Model

After extracting the VISH feature, human poses are modeled by a nonparametric AMM as shown in Fig. 1. The AMM is derived from an instance-based learning algorithm [37] without making stronger assumptions about the nature of the human-pose distributions compared with other parametric models such as an exponential family of distributions. An instance contains a ground-truth human pose, the corresponding VISH feature and action label. The instances are collected in advance to form a database called human-pose database, denoted as $\mathcal{D}$. Thus, the human-pose database contains all possible human poses generated at the output. In the human-pose database, the human poses from the ith action are grouped to form a subset of the human-pose database, denoted as $\mathcal{D}_i$.

Let x be a VISH-feature input and y be a human pose from the human-pose database. Using the concept of distributed representation, a human pose is estimated from multiple low-dimensional manifolds, where each manifold corresponds to one action indexed by i; that is, a human pose can be represented by a linear combination of multiple actions. Since a human pose may appear more often in some actions than the others, a human pose belonging to multiple actions should be weighted with different weights represented by its weighting functions $g_i(x)$'s. Thus, in estimating a human pose, we consider it to be represented by multiple actions weighted by its weighting functions. For each action, we determine the likelihood of a human pose from the human-pose database being the human pose of the VISH-feature input. We call the likelihood a base distribution, denoted by $p_i(y \mid x)$. Hence, the probability of a human pose y in the human-pose database being the human pose of a VISH-feature input x can be expressed as a linear combination of the base distributions $p_i(y \mid x)$'s with weights $g_i(x)$'s. Mathematically, the probability is given by

$$p(y \mid x) = \sum_{i=1}^{N} g_i(x) p_i(y \mid x) \tag{4}$$

where x is a VISH-feature input, $g_i(\cdot)$ is a weighting function for the ith manifold (action), $p_i(y \mid x)$ is the conditional probability (base distribution) of a human pose in the ith manifold and N is the number of manifolds. Note that, when a new VISH-feature comes as an input, the weights $g_i(x)$'s and the base distributions $p_i(y \mid x)$'s will be computed again. The human pose associated with the VISH-feature input x is then estimated by finding the human pose with the highest conditional probability; that is

$$y^* = \arg\max_{y \in Y} p(y \mid x) \tag{5}$$

where y^* is the human-pose estimate for the VISH-feature input x, and Y is the set containing all possible human poses in the human-pose database.

The AMM involves modeling two terms—the weighting function $g_i(\cdot)$ for each manifold and the base distribution $p_i(y \mid x)$ of human-pose estimation in each manifold. The weighting function is represented by a probability mass function (PMF) of action classification. The PMF is estimated using an action database, which contains VISH features and the corresponding action labels. We use the bootstrap aggregating algorithm (i.e., bagging) [38] to train an action classifier because it has been shown successful in classifying actions. Details on how to train the action classifier can be found in [38]. After training, the weighting function of each manifold is represented by the PMF to determine a weighting coefficient for each manifold for every VISH-feature input x. The base distribution is modeled in two steps, namely estimation and redistribution steps, as follows.

A. Estimation Step

The estimation step measures the likelihood of human poses spatially. As the spatial information of a 3-D point cloud is represented by the proposed VISH feature, the likelihood can be calculated based on the Euclidean distance between VISH features. If the distance is small, human poses are close spatially. In addition, since we use the concept of distributed representation, we expect that the human pose of a VISH-feature input should be similar to some human poses from the human-pose database with the same action. Thus, we increase the likelihood of all human poses from the same action by reducing the likelihood of human poses from other actions. The likelihood is reduced by introducing a weighting coefficient that is less than one. Mathematically, the unnormalized probability of a human pose in the ith action, where i is an element in the set $\{1, \ldots, N\}$, is defined as

$$\widetilde{p}_i^f(x' \mid x) = \begin{cases} \frac{1}{\|x-x'\|_2+z} & \text{when } x' \in A_i \\ \frac{1}{N} \cdot \frac{1}{\|x-x'\|_2+z} & \text{when } x' \notin A_i \end{cases} \tag{6}$$

where x' is the VISH feature of a human pose in the human-pose database, i is the index of an action, A_i is the set of features, each of which corresponds to a human pose in $\mathcal{D}_i$, z is a small constant to avoid division by zero, $\|\cdot\|_2$ is the Euclidean norm and $\frac{1}{N}$ is the weighting coefficient to reduce the likelihood of human poses from other actions. Hence, the unnormalized probability of a human pose is larger if the human pose comes from the same action as the VISH-feature input and the VISH-feature input is closer to the VISH feature of the human pose in the Euclidean space. The probability of a human pose in the ith action is then given by

$$p_i^f(x' \mid x) = \frac{\widetilde{p}_i^f(x' \mid x)}{\sum_{x' \in A} \widetilde{p}_i^f(x' \mid x)} \tag{7}$$

where A is the union of the feature sets $A_1, A_2, \ldots, A_N$.

B. Redistribution Step

The redistribution step measures the likelihood of human poses based on the frequency of actions. The frequency of actions is defined as the portion of time a human pose in each action. If the frequency of an action is higher than the other actions, that action is expected to be observed more often and thus the likelihood of human poses in that action is higher. In the redistribution step, we consider that human poses could be similar in different actions. Thus, observing a human pose from an action does not only change the frequency of that action but also the frequency of some other actions. We model this phenomenon using a continuous-time Markov chain as follows.

Mathematically, the probability distributions estimated from the estimation step are weighted according to the frequency of actions. Assume that, given the present action in a sequence of actions, the rest of the past actions is irrelevant for predicting the future actions. The weight is estimated by the stationary probability in a continuous-time Markov chain, which is trained using the VISH features in the human-pose databases. Let $X(t)$ be the continuous-time Markov chain with the state space $I = \{1, 2, \ldots, N\}$ for $t \geq 0$. The state space contains the indices of N actions. Assume the Markov chain is temporally homogeneous. The (i, j) entry of the transition probability matrix, denoted as Q, of the Markov chain $X(t)$ is defined as

$$Q(i,j) = \begin{cases} \lim\limits_{h \to 0} \frac{p(X(h)=j \mid X(0)=i)}{h} & \text{when } i \neq j \\ \lim\limits_{h \to 0} \frac{p(X(h)=i \mid X(0)=i)-1}{h} & \text{when } i = j. \end{cases} \tag{8}$$

For $i \neq j$, the transition probability measures the jump rate of the Markov chain from state i to j. For $i = j$, the transition probability is the negation of the rate at which the Markov chain leaves state i. The jump rate is estimated by modeling the transition of actions in a Poisson process [39] using the temporal information in the human-pose database.

Let λ_{ij} be the arrival rate of a state from i to j. The arrival rate between two actions is calculated by the normalized dynamic time warping algorithm (DTW) [40] that measures the similarity between two actions; that is

$$\lambda_{ij} = \begin{cases} \frac{z_1}{DTW(i,j)+r} & \text{when } DTW(i,j) < \tau \\ 0 & \text{when } DTW(i,j) \geq \tau \end{cases} \tag{9}$$

where z_1 is a constant, τ is a predefined threshold, $DTW(i, j)$ is the distance between the ith and jth actions calculated by the normalized DTW and r is a uniform random variable between 0 and 1.

The normalized DTW is used because actions may be different in time or speed. Given two actions, the normalized DTW calculates their matching cost under the optimal alignment by warping the two actions. The matching cost of a pair of frames in two actions is defined as the Euclidean distance between the VISH features extracted from the 3-D point clouds in the two frames.

As one kind of action can be changed to another action at any time, we assume the arrival of a state is equally likely at all time. Thus, if one unit of time is divided into m intervals, the probability of the arrival of state j from state i in each interval is $\frac{\lambda_{ij}}{m}$. The probability of the first arrival after time t can be approximated by

$$(1 - \frac{\lambda_{ij}}{m})^{tm} \xrightarrow[m \to \infty]{} e^{-\lambda_{ij} t}. \tag{10}$$

Therefore, the interarrival time is exponentially distributed with rate λ_{ij}. Hence, the (i, j) entry of the transition probability matrix Q is given by

$$Q(i,j) = \begin{cases} \lambda_{ij} & \text{when } i \neq j \\ -\sum_{k=1, k \neq i}^{N} \lambda_{ik} & \text{when } i = j. \end{cases} \tag{11}$$

The state space I is partitioned into a minimum number, denoted as M, of mutually exclusive and exhaustive sets such that the continuous-time Markov chain with any of the M partitioned sets is irreducible. Let $X_k(t)$ be the continuous-time Markov chain with the kth partitioned set I_k, where $k = 1, 2, \ldots, M$. The transition probability matrix, denoted as Q_k of the Markov chain $X_k(t)$, can be formed from the transition probability matrix Q by deleting its rows and columns of the corresponding actions that are not in the state space of the Markov chain $X_k(t)$. If the Markov chain $X_k(t)$ is positive recurrent, then the stationary distribution of the Markov chain

$X_k(t)$, denoted as π_k, can be found by solving $\pi_k Q_k = 0$; otherwise, the stationary distribution is set to be uniform to indicate equal importance of each action.

The probability distributions estimated from the estimation step are then weighted according to the stationary distribution as follows:

$$p_i^a(x' \mid x) = \pi_k(i) \sum_{j \in I_k} p_j^f(x' \mid x), \quad i \in I_k \tag{12}$$

where $\pi_k(i)$ is the stationary distribution of state i in the Markov chain $X_k(t)$.

The unnormalized base distribution, denoted as $\tilde{p}_i(y \mid x)$, is defined as the combination of the outputs from the two steps; that is

$$\tilde{p}_i(y \mid x) = u^f p_i^f(x' \mid x) + u^a p_i^a(x' \mid x) \tag{13}$$

where x' is the VISH features associated with the human pose y, u^f and u^a are user-defined constants. If u^f is larger (smaller) than u^a, the probability distribution estimated from the estimation step will have more (less) influence on the unnormalized base distribution. The base distribution $p_i(y \mid x)$ is then derived by normalizing the unnormalized base distribution as follows:

$$p_i(y \mid x) = \frac{\tilde{p}_i(y \mid x)}{\sum_{y \in D_i} \tilde{p}_i(y \mid x)}. \tag{14}$$

Using the proposed nonparametric AMM, human poses in each manifold is first modeled in the estimation step. The probability distributions of human poses among the manifolds are redistributed according to the frequency of actions in the redistribution step. The classification result then aggregates the base distributions from all actions. As we will show in computer simulations, the proposed AMM can increase the accuracy of human-pose estimates. However, human poses are estimated in discrete space. Quantization error is induced in the human-pose estimates. In the next section, we will describe a kinematic model that reduces the quantization error.

V. Kinematic Model

As the human poses estimated by the AMM are in discrete space, we use a kinematic model as shown in Fig. 6 to describe the spatial relationship between body parts of the human poses estimated from the AMM for reducing the quantization error in the AMM. Assume there is an underlying probability distribution governing the position and orientation of body parts. It is represented by a directed acyclic graph $G = (V, E)$ where V corresponds to a set of vertices and E corresponds to a set of edges. The vertex 1 is a softmax random variable [37] of the human pose estimated by the AMM. Let $V^- = V \setminus \{1\}$. Each vertex $s \in V^-$ corresponds to a body part and there is a random variable, denoted as O_s, representing the orientation of the body part with respect to its parent in the kinematic chain. The parent of the torso is set to be null and the orientation of the torso is measured with respect to the normal of the floor plane. The length of each body part is assumed to be fixed and the orientation is represented by a quaternion. As a result, O_s lies in a 4-D manifold. A (stochastic) configuration, C, of a human is a collection of body parts; that is,

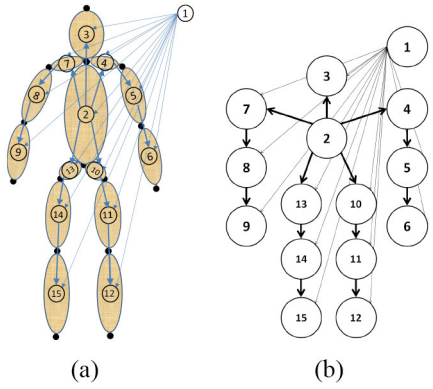

Fig. 6. (a) Kinematic model of a human. (b) Graph representation of a human kinematic model. Vertices are: 1, human pose estimated by the proposed nonparametric AMM; 2, torso; 3, head; 4, left shoulder; 5, left upper arm; 6, left lower arm; 7, right shoulder; 8, right upper arm; 9, right lower arm; 10, left hip; 11, left upper leg; 12, left lower leg; 13, right hip; 14, right upper leg; 15, right lower leg. The arrows represent the dependencies between vertices.

$C = \{O_2, O_3, \ldots, O_{15}\}$. Let $c, o_2, \ldots, o_{15}$ be the realization of the random variables $C, O_2, \ldots, O_{15}$, respectively. Given the graph G, the probability of a configuration c can be written as follows:

$$\begin{aligned} p(c) &= p(c \mid v_1) = p(\{o_2, o_3, \ldots, o_{15}\} \mid v_1) \\ &= p(o_2, o_3, \ldots, o_{15}) \end{aligned} \tag{15}$$

where the parentheses and the conditioning event v_1 are removed for notational simplicity.

Modeling the probability distribution $p(C)$ is generally intractable because of the high dimensionality. However, by exploiting the dependencies in the graph G, the probability distribution $p(C)$ can be rewritten as

$$p(C) = \prod_{s \in V^-} p(O_s \mid pa(O_s)) \tag{16}$$

where $pa(\cdot): V^- \mapsto V^-$ is a mapping from a vertex to its parent in the kinematic chain except that the torso in V^- is mapped to $\emptyset$ (null).

The probability distribution in (16) is tractable. For each body part, $p(O_s \mid pa(O_s))$ is assumed to be a Gaussian distribution with a mean vector μ_s and a positive-definite variance matrix Σ_s; that is

$$p(O_s \mid pa(O_s)) = \mathcal{N}(O_s \mid \mu_s, \Sigma_s) \tag{17}$$

where $\mathcal{N}(\cdot \mid \mu_s, \Sigma_s)$ is a Gaussian distribution.

To find the parameters μ_s and Σ_s in the kinematic model, a refinement database, denoted as $\mathcal{T}$, is created. The parameters can then be determined by maximizing the log-likelihood

$$\sum_{n=1}^{|\mathcal{T}|} \log p(c^n) = \sum_{n=1}^{|\mathcal{T}|} \sum_{s \in V^-} \log p\left(O_s^n \mid pa(O_s^n)\right) \tag{18}$$

where c^n is the nth configuration in the refinement database $\mathcal{T}$ and O_s^n is the orientation of a body part at vertex s in the nth configuration.

When refining a human pose using the kinematic chain, body parts are divided into observable and unobservable groups. We define body parts in the observable group as

nonoccluded body parts and body parts in the unobservable group as occluded body parts. Inference is made based on the body parts in the observable (nonoccluded) group to estimate/refine the body parts in the unobservable (occluded) group. The orientation of a body part in the observable group can be determined by finding the orientation of the line joining the joint positions at the two ends of the body part. The joint positions are obtained from the human pose estimated by the AMM. Mathematically, the set of vertices in the graph G is divided into evidential (observable) and non-evidential (unobservable) sets; that is

$$V^- = V_e^- \cup V_n^- \quad \text{and} \quad V_e^- \cap V_n^- = \emptyset \tag{19}$$

where V_e^- is the set of evidential vertices and V_n^- is the set of non-evidential vertices.

Given the orientation of the body parts in V_e^-, the orientation in the non-evidential vertices can be estimated by the most likely configuration, denoted as c^*, of the probability distribution of configuration; that is

$$c^* = \underset{c \in C,\ s \in V_e^-}{\arg\max}\ p(c \mid o_s). \tag{20}$$

Based on the factorization structure of the distribution in the kinematic model, (20) can be calculated efficiently using belief propagation [41]. The orientation of body parts that maximizes (16) are converted to a human pose. The human pose is then averaged with the human pose estimated from the AMM to provide the refined human-pose estimate, which is close to the human-pose estimate from the AMM, in continuous space.

VI. Computer-Simulation Results

The proposed 3-D-point-cloud system was implemented using the point cloud library (PCL) [42] and tested on the Stanford TOF Motion Capture Dataset [23], which contains 28 video sequences. Each video sequence corresponds to one action. The human pose of the subject has 15 degrees-of-freedom (joints), namely head, neck, left/right shoulder, left/right elbow, left/right wrist, hip center, left/right hip, left/right knee, and left/right ankle. The ground-truth 3-D joint locations of the subject were recorded by a commercial motion-capturing system. When evaluating the performance of human-pose estimation, frames with missing ground-truth 3-D joint locations were ignored. The error metric, ζ, for each video sequence was defined as

$$\zeta = \frac{1}{N_f} \sum_{s=1}^{N_f} \frac{1}{N_s} \sum_{i=1}^{N_s} \left\| \mathbf{j}_{s,i} - \tilde{\mathbf{j}}_{s,i} \right\|_2 \tag{21}$$

where N_f is the number of frames of the video sequence for testing, N_s is the number of 3-D joint locations measured by the motion-capturing system in the sth frame, $\mathbf{j}_{s,i}$ and $\tilde{\mathbf{j}}_{s,i}$ are the ground-truth and the estimated 3-D location of the ith joint in the sth frame, respectively, and $\|\cdot\|_2$ is the Euclidean norm.

A. Evaluation of VISH

Two existing geometric features, namely VFH and AGEX, were implemented for comparison. The k-NN and SVM were used to estimate human poses based on the geometric features. The features were evaluated using 5-fold cross-validation.

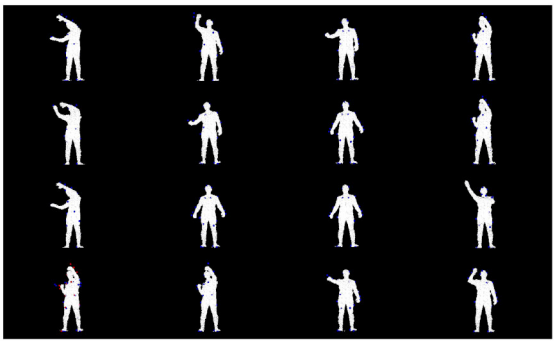

Fig. 7. Ambiguity of symmetric human poses represented by the VFH feature exists in some close matches using k-NN. The 3-D point cloud at the bottom left is the query pose. The others are the 3-D point clouds returned by k-NN. The returned 3-D point clouds are ranked from left to right, and bottom to top.

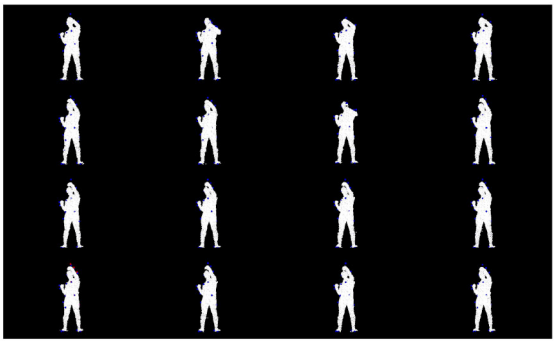

Fig. 8. Ambiguity of symmetric human poses is greatly reduced using the VISH feature.

In the preprocessing step, the predefined 3-D region was set to be the 3-D space with the depth ranging from -3 m to -2 m. Fifty closest 3-D points were used to estimate the average distance of each 3-D point.

Two hierarchical levels were used in the tree structure. Six cuboids were split from the smallest cuboid containing the root node. To estimate the direction of a 3-D point, 3-D points within 0.01 m from that 3-D point were used. The size of the 2-D region used for shape feature was set to be 17 pixels $\times$ 17 pixels.

1) Comparison Between VISH and VFH: As the proposed VISH feature is derived from the VFH feature, we first compare the performance of the two features qualitatively. Human poses were estimated by k-NN using the two features. Figs. 7 and 8 show the human poses estimated by k-NN using VFH and VISH features, respectively. In the figures, the subject at the bottom left was raising his/her left arm. The human-pose estimates (matches) returned by k-NN were ranked from left to right, and bottom to top. For example, the closest match was at the second column in the last row. Some of the matches were raising the other arm in Fig. 7. It showed that the VFH feature gave a similar description among symmetric 3-D human poses and could not distinguish them. When the subject pose was represented by the VISH feature, this ambiguity was greatly reduced.

The main difference between the VISH and VFH features was the spatial ordering of 3-D point cloud. When the VISH

TABLE I
QUANTITATIVE RESULT OF FEATURE EVALUATION USING k-NN WHEN $k = 3$. NUMBERS ON THE LEFT AND INSIDE THE PARENTHESES ARE ERRORS AND STANDARD DEVIATIONS (IN METERS), RESPECTIVELY

Test case	VFH	AGEX	VISH
0	0.030 (0.021)	0.042 (0.023)	0.015 (0.008)
1	0.039 (0.027)	0.073 (0.037)	0.016 (0.012)
2	0.028 (0.019)	0.053 (0.030)	0.013 (0.008)
3	0.026 (0.013)	0.084 (0.051)	0.021 (0.012)
4	0.032 (0.014)	0.047 (0.024)	0.015 (0.008)
5	0.039 (0.023)	0.051 (0.029)	0.014 (0.009)
6	0.038 (0.021)	0.055 (0.029)	0.012 (0.008)
7	0.033 (0.022)	0.054 (0.027)	0.016 (0.012)
8	0.038 (0.025)	0.055 (0.026)	0.012 (0.010)
9	0.040 (0.028)	0.065 (0.036)	0.015 (0.008)
10	0.033 (0.023)	0.078 (0.059)	0.019 (0.010)
11	0.024 (0.016)	0.056 (0.030)	0.013 (0.007)
12	0.025 (0.019)	0.077 (0.049)	0.012 (0.008)
13	0.026 (0.026)	0.068 (0.038)	0.012 (0.010)
14	0.025 (0.019)	0.085 (0.057)	0.020 (0.016)
15	0.034 (0.020)	0.061 (0.031)	0.014 (0.008)
16	0.065 (0.042)	0.071 (0.034)	0.017 (0.012)
17	0.058 (0.034)	0.086 (0.043)	0.023 (0.015)
18	0.034 (0.027)	0.079 (0.030)	0.015 (0.011)
19	0.042 (0.030)	0.071 (0.048)	0.016 (0.011)
20	0.040 (0.025)	0.082 (0.058)	0.020 (0.013)
21	0.043 (0.023)	0.085 (0.045)	0.022 (0.014)
22	0.042 (0.022)	0.073 (0.046)	0.014 (0.009)
23	0.048 (0.043)	0.089 (0.069)	0.015 (0.011)
24	0.049 (0.059)	0.141 (0.121)	0.021 (0.022)
25	0.032 (0.025)	0.096 (0.070)	0.020 (0.020)
26	0.030 (0.030)	0.097 (0.058)	0.018 (0.012)
27	0.060 (0.057)	0.177 (0.090)	0.023 (0.019)
Overall	**0.038 (0.027)**	**0.077 (0.046)**	**0.017 (0.012)**

feature was used, the VFH and shape features were extracted from the 3-D point clouds in a tree. However, the VFH feature was extracted from the input 3-D point cloud directly. As the VFH feature was a histogram-based feature, extracting the feature from the input 3-D point cloud directly could only capture the global properties of the point cloud. Any local properties (fine details) were suppressed. On the other hand, the VISH feature could capture both the global and local properties by partitioning the input 3-D point cloud into regions.

2) Comparison Using k-NN: The goal is to evaluate the accuracy of 3-D human-pose estimation using the proposed VISH feature in a discriminative model. k-NN was used to learn the mapping from a geometric feature to a 3-D human pose in the training database. The k-nearest human poses were then averaged to give the final estimate of human pose.

Table I shows the quantitative result of the feature evaluation when $k = 3$. A bar chart of the errors using different features is shown in Fig. 9. The first bar (in blue color) corresponding to the error of VISH was significantly lower than the other two bars and it showed the strength of spatial ordering of 3-D point cloud in VISH. The second bar (in red color) showed the error of using VFH. The error was higher than that of VISH because VFH suffered from the limitation of describing the local properties of input 3-D point cloud. The third bar (in green color) showed the error of using AGEX. In AGEX, when the limbs of the person were moved, the geodesic extrema of the limbs might be switched or lost. It failed to look up the human poses in the dataset. The overall error of the proposed VISH feature was 0.017 m. In VFH and AGEX, the overall

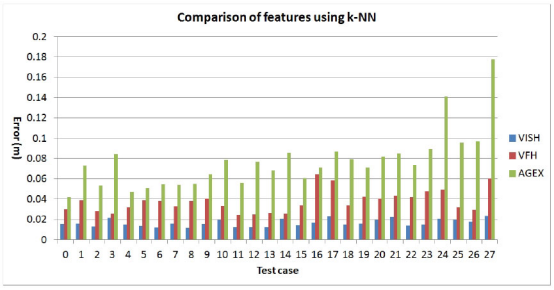

Fig. 9. Evaluation of VISH, VFH, and AGEX features using k-NN (when $k = 3$).

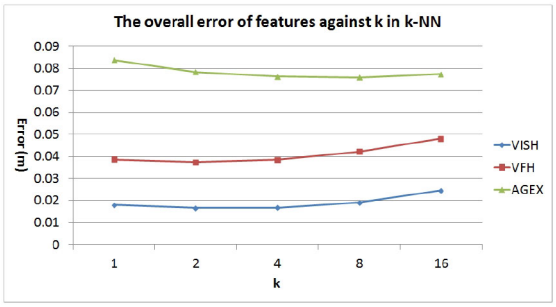

Fig. 10. Comparison of VISH, VFH, and AGEX features using k-NN under different values of k.

errors were about twice and 4.6 times as much as that of the proposed VISH feature, respectively. It showed that the VISH feature was more discriminative than the other two features.

The overall error of features under different values of k in k-NN is plotted in Fig. 10. The overall error of VISH was the lowest among the three features when the value of k was changed. When the value of k started to increase from one, the error was decreased because the noise in the point cloud $\mathcal{P}_H$ was averaged out. As the value of k increased, the error was first decreased, but started to increase because the details from the point cloud $\mathcal{P}_H$ were also averaged out when k was too big. Overall, the VISH feature was robust across a wide range of k in k-NN.

3) Comparison Using SVM: The accuracy of the proposed VISH feature was also evaluated using SVM. One thousand and five hundred human-pose prototypes were used to model 3-D human poses. They were found by running the k-means algorithm [35] on the training video sequences.

Table II shows the quantitative result of the feature evaluation. A bar chart of the errors using different features is shown in Fig. 11. The overall errors of VFH and AGEX features were about 1.5 times and 2.4 times as much as that of VISH, respectively. It assured that the proposed VISH feature performed better than the other two features. Note that the overall error and standard deviation using SVM was higher than the error using k-NN. The main reason was the quantization error in the human-pose prototypes when modeling the 3-D human poses in high-dimensional space.

B. Evaluation of the Proposed 3-D-Point-Cloud System

The dataset was divided into 20% for establishing the human-pose database, 30% for building the refinement

TABLE II
QUANTITATIVE RESULT OF FEATURE EVALUATION USING SVM. NUMBERS ON THE LEFT AND IN THE PARENTHESES ARE ERRORS AND STANDARD DEVIATIONS (IN METERS), RESPECTIVELY

Test case	VFH	AGEX	VISH
0	0.038 (0.051)	0.069 (0.053)	0.036 (0.043)
1	0.126 (0.091)	0.122 (0.083)	0.072 (0.076)
2	0.033 (0.055)	0.074 (0.044)	0.031 (0.047)
3	0.077 (0.080)	0.101 (0.053)	0.071 (0.072)
4	0.064 (0.045)	0.103 (0.047)	0.021 (0.035)
5	0.065 (0.061)	0.075 (0.069)	0.039 (0.044)
6	0.065 (0.070)	0.101 (0.067)	0.040 (0.060)
7	0.047 (0.064)	0.133 (0.065)	0.034 (0.063)
8	0.104 (0.085)	0.164 (0.056)	0.055 (0.059)
9	0.040 (0.062)	0.123 (0.051)	0.019 (0.034)
10	0.075 (0.078)	0.138 (0.066)	0.059 (0.075)
11	0.022 (0.049)	0.164 (0.051)	0.015 (0.035)
12	0.056 (0.069)	0.146 (0.067)	0.030 (0.047)
13	0.030 (0.052)	0.099 (0.040)	0.030 (0.050)
14	0.054 (0.070)	0.103 (0.050)	0.048 (0.055)
15	0.056 (0.063)	0.144 (0.069)	0.027 (0.043)
16	0.082 (0.064)	0.084 (0.061)	0.023 (0.043)
17	0.128 (0.098)	0.202 (0.085)	0.079 (0.085)
18	0.071 (0.065)	0.077 (0.055)	0.056 (0.068)
19	0.085 (0.088)	0.151 (0.093)	0.051 (0.076)
20	0.114 (0.089)	0.140 (0.074)	0.071 (0.077)
21	0.130 (0.094)	0.173 (0.076)	0.101 (0.097)
22	0.136 (0.109)	0.194 (0.067)	0.064 (0.094)
23	0.133 (0.117)	0.143 (0.082)	0.071 (0.094)
24	0.173 (0.185)	0.204 (0.138)	0.138 (0.158)
25	0.106 (0.102)	0.149 (0.094)	0.098 (0.107)
26	0.078 (0.106)	0.158 (0.105)	0.072 (0.104)
27	0.139 (0.163)	0.201 (0.132)	0.135 (0.144)
Overall	**0.083 (0.083)**	**0.133 (0.071)**	**0.057 (0.071)**

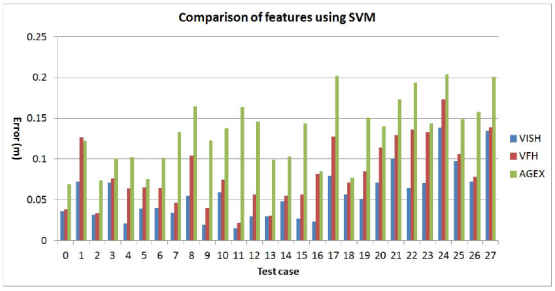

Fig. 11. Evaluation of VISH, VFH, and AGEX features using SVM.

database, 40% for building the action database, and 10% for testing. To reduce the bias in dividing the dataset, the dataset was randomly divided ten times with different random seeds in each trial. When the dataset was randomly divided, the proposed system was evaluated using 5-fold cross-validation. In the estimation step, the constant z in (6) was set to 0.1. In the redistribution step, the constant z_1 in (9) was set to 1. The threshold τ was calculated by subtracting the standard deviation of the values given by DTW from the mean of the values. When estimating the unnormalized base distribution, the constants u^f and u^a were set to 1. The non-evidential set in the kinematic model contained the vertices of lower arms and legs.

To evaluate the performance of the proposed AMM and the kinematic model in the proposed system, we considered the following three cases.

1) VISH: A human pose was estimated by the nearest neighbors using VISH features.

TABLE III
ERRORS (IN METERS) OF HUMAN-POSE ESTIMATION. NUMBERS ON THE LEFT AND IN THE PARENTHESES ARE THE ERRORS AND STANDARD DEVIATIONS OF HUMAN-POSE ESTIMATION, RESPECTIVELY

Trial	VISH	VISH+AMM	Proposed System
1	0.0304 (0.0387)	0.0275 (0.0334)	0.0251 (0.0295)
2	0.0289 (0.0308)	0.0265 (0.0306)	0.0244 (0.0280)
3	0.0287 (0.0365)	0.0263 (0.0276)	0.0241 (0.0247)
4	0.0304 (0.0391)	0.0269 (0.0342)	0.0247 (0.0311)
5	0.0276 (0.0307)	0.0255 (0.0289)	0.0236 (0.0264)
6	0.0294 (0.0354)	0.0280 (0.0349)	0.0255 (0.0309)
7	0.0290 (0.0333)	0.0273 (0.0325)	0.0248 (0.0284)
8	0.0287 (0.0352)	0.0264 (0.0330)	0.0245 (0.0306)
9	0.0293 (0.0368)	0.0271 (0.0372)	0.0248 (0.0335)
10	0.0285 (0.0343)	0.0268 (0.0341)	0.0246 (0.0306)
Overall	**0.0291 (0.0351)**	**0.0268 (0.0326)**	**0.0246 (0.0294)**

2) VISH+AMM: A human pose was estimated by the nonparametric AMM using VISH features.
3) VISH+AMM+Kinematic Model: A human pose was estimated by the proposed 3-D-point-cloud system.

Table III shows the errors and standard deviations of human-pose estimation incurred in these three cases. When comparing VISH+AMM with VISH, the overall error and standard deviation in VISH+AMM were reduced compared with that in VISH. The reduction of the overall error and standard deviation were about 7.9% and 7.1%, respectively. It showed that the result of action classification in the proposed AMM was useful in reducing the errors of human-pose estimates. Using the kinematic model, the proposed system further reduced the overall error and standard deviation. When comparing the proposed system with VISH+AMM, the overall error and standard deviation in the proposed system were reduced by 8.2% and 9.8%, respectively. When comparing the proposed system with VISH, the overall error and standard deviation of the proposed system were reduced by 15.5% and 16.2%, respectively. Thus, the kinematic model can reduce the quantization error of the human poses estimated by the AMM.

To test the effectiveness of the distributed representation of a human pose, the system was modified to control the number of actions (manifolds), denoted as N_a, being considered in the base distribution for human-pose estimation. The number of actions N_a was varied from one (one action) to 28 (all actions). The 28 actions were first sorted in a list in descending order according to the PMF of action classification. Then, the N_a actions were selected from the first N_a actions in the sorted list.

Fig. 12 shows the changes of the overall error and standard deviation of human-pose estimation. In the figure, when the number of actions N_a increased initially, the overall error and standard deviation incurred by the proposed system were decreased, showing that multiple actions should be considered in the system to yield a better representation of human poses and hence increase the accuracy/precision of human-pose estimation. As the number of actions N_a further increased, the overall error and standard deviation stopped to decrease.

The results of the proposed system were compared with the results reported in some existing works using the Stanford TOF Motion Capture Dataset [23]. The overall errors and standard deviations are shown in Table IV. The proposed system incurred the lowest overall error and standard deviation,

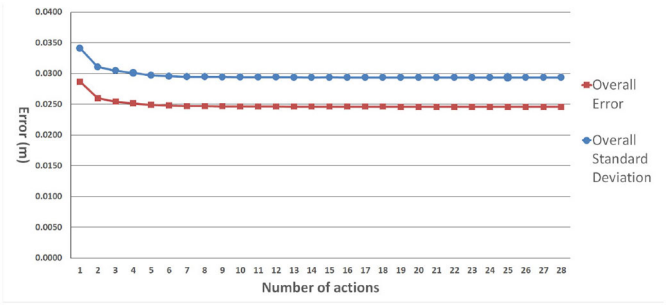

Fig. 12. Changes of the overall error and standard deviation of human-pose estimation using different numbers of actions.

TABLE IV
OVERALL ERRORS AND STANDARD DEVIATIONS (S.D.) OF HUMAN-POSE ESTIMATION IN THE STANFORD TOF MOTION CAPTURE DATASET [23]. THE ERRORS AND STANDARD DEVIATIONS OF THE EXISTING WORKS WERE OBTAINED FROM THEIR PAPERS

	Overall Error (m)	Overall S.D. (m)
HC and EP Method [23]	0.1	N.A.
Data-driven Hybrid Method [43]	0.0618	0.0424
Exemplar Method [11]	0.038	N.A.
Proposed 3D-Point-Cloud System	0.0246	0.0294

showing that the result from action classification and the kinematic model could reduce the errors in human-pose estimation. Note that the overall standard deviation in the proposed system was larger than the overall error because the human poses in the human-pose database for estimating the base distribution could not fully describe the human poses in the test dataset. Thus, for some human poses estimated by the proposed system, the errors were larger than the errors of other human-pose estimates.

VII. CONCLUSION

In this paper, we have proposed a 3-D-point-cloud system that uses a static depth sensor to estimate human poses. The proposed system is a general framework that captures the geometric properties of the 3-D point cloud of a human, and combines the action-classification result and a kinematic model in human-pose estimation. Since the proposed system only uses temporal information in the training phase, it does not accumulate errors from previous frames. In the proposed system, we proposed a 3-D-point-cloud feature VISH that highlights the orientation and shape features in both the global and local regions in 3-D point clouds to reduce feature ambiguity. The VISH feature was composed of 3-D-point-cloud preprocessing, hierarchical structuring, and VFH and shape feature extraction. In the preprocessing step, region-based thresholding and pseudo-residual were used to extract the 3-D points from a person. The 3-D points were then organized into a tree structure. The VFH and shape features were extracted separately from each node in the tree. The VISH feature was formed by combining all the features and therefore preserved the spatial ordering of the 3-D point cloud. The spatial ordering can capture the global and local properties of 3-D points in the 3-D point cloud. These properties can greatly remove the ambiguity of symmetric human poses. Two existing geometric features, namely VFH and AGEX, were implemented and compared with the proposed VISH feature. Both k-NN and SVM showed that the proposed feature incurred less errors than the other two features, suggesting that the proposed feature can describe the 3-D point cloud more accurately for 3-D human-pose estimation. The time complexity of the proposed feature extraction is asymptotically the same as that of VFH, which is $O(n)$, and VISH is more efficient than AGEX, which is of time complexity $O(n \log n)$.

In modeling human poses, we proposed a nonparametric AMM to represent human-pose space using multiple low-dimensional manifolds associated with actions for a more accurate estimation of the underlying probability distribution of human poses in human-pose space. The base distribution in the AMM for human-pose estimation was derived in two steps: estimation and redistribution steps. The estimation step calculated the Euclidean distance between the VISH-feature input and the VISH features from the human-pose database. It then produced a probability distribution of human poses in each manifold associated with an action. The redistribution step assigned weights to the probability distributions among all the manifolds from the estimation step according to the frequency of actions. The action of a VISH-feature input was classified by the bootstrap aggregating algorithm (bagging) to determine the weighting coefficients in combining the manifolds. As the human poses estimated by the proposed AMM were in discrete space, the kinematic model was used to model the spatial relationship of body parts in continuous space to reduce the quantization error in the AMM. Computer-simulation results showed that using multiple low-dimensional manifolds can represent the human-pose space and increase the accuracy and precision of human-pose estimates. The overall error and standard deviation of the proposed system were reduced compared with existing approaches without action classification.

REFERENCES

[1] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2004.
[2] J. Gall *et al.*, "Motion capture using joint skeleton tracking and surface estimation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1746–1753.
[3] R. Kehl, M. Bray, and L. van Gool, "Full body tracking from multiple views using stochastic sampling," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 129–136.
[4] R. Kehl and L. van Gool, "Markerless tracking of complex human motions from multiple views," *Comput. Vis. Image Underst.*, vol. 104, no. 2, pp. 190–209, Nov. 2006.
[5] M. Straka, S. Hauswiesner, M. Rüther, and H. Bischof, "Skeletal graph based human pose estimation in real-time," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 69.1–69.12.
[6] R. Plänkers and P. Fua, "Articulated soft objects for multiview shape and motion capture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1182–1187, Sep. 2003.
[7] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1297–1304.
[8] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Washington, DC, USA, Nov. 2011, pp. 415–422.
[9] R. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3D recognition and pose using the viewpoint feature histogram," in *Proc. Intell. Robots Syst. (IROS)*, Taipei, Taiwan, Oct. 2010, pp. 2155–2162.

[10] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Anchorage, AK, USA, May 2010, pp. 3108–3113.

[11] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3D pose estimation from a single depth image," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 731–738.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2003.

[13] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Scholkopf, Eds. Cambridge, MA, USA: MIT Press, 2004.

[14] S. Sedai, D. Huynh, and M. Bennamoun, "Supervised particle filter for tracking 2D human pose in monocular video," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Kona, HI, USA, Jan. 2011, pp. 367–373.

[15] Y. Tian, L. Sigal, H. Badino, F. De la Torre, and Y. Liu, "Latent Gaussian mixture regression for human pose estimation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Berlin, Germany, 2011, pp. 679–690.

[16] Y. Su, H. Ai, T. Yamashita, and S. Lao, "Human pose estimation using exemplars and part based refinement," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Queenstown, New Zealand, 2011, pp. 174–185.

[17] J. Gall, A. Fossati, and L. van Gool, "Functional categorization of objects using real-time markerless motion capture," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2011, pp. 1969–1976.

[18] Y. Zhu and K. Fujimura, "Bayesian 3D human body pose tracking from depth image sequences," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, vol. 5995. Xi'an, China, 2010, pp. 267–278.

[19] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2011, pp. 1705–1712.

[20] S. Zuffi, O. Freifeld, and M. Black, "From pictorial structures to deformable structures," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, Jun. 2012, pp. 3546–3553.

[21] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2004, pp. I.421–I.428.

[22] L. Sigal, M. Isard, H. Haussecker, and M. Black, "Loose-limbed people: Estimating 3D human pose and motion using nonparametric belief propagation," *Int. J. Comput. Vis.*, vol. 98, no. 1, pp. 15–48, May 2012.

[23] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, Jun. 2010, pp. 755–762.

[24] A. Yao, J. Gall, G. Fanelli, and L. van Gool, "Does human action recognition benefit from pose estimation?" in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 67.1–67.11.

[25] J. Gall, A. Yao, and L. van Gool, "2D action recognition serves 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Heraklion, Greece, 2010, pp. 425–438.

[26] J. Gall, A. Yao, N. Razavi, L. van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2188–2202, Nov. 2011.

[27] A. Yao, J. Gall, and L. van Gool, "Coupled action recognition and pose estimation from multiple views," *Int. J. Comput. Vis.*, vol. 100, no. 1, pp. 16–37, 2012.

[28] K. Chan, C. Koh, and C. Lee, "A 3D-point-cloud feature for human-pose estimation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2013, pp. 1615–1620.

[29] K. Shoemake, "Animating rotation with quaternion curves," *SIGGRAPH Comput. Graph.*, vol. 19, no. 3, pp. 245–254, 1985.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893.

[31] R. Poppe, "Evaluating example-based pose estimation: Experiments on the HumanEva sets," in *Proc. Comput. Vis. Pattern Recognit. (CVPR) 2nd Workshop Eval. Articulated Human Motion Pose Estimation*, 2007.

[32] S. Wang, H. Ai, T. Yamashita, and S. Lao, "Combined top-down/bottom-up human articulated pose estimation using Adaboost learning," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Istanbul, Turkey, Aug. 2010, pp. 3670–3673.

[33] D. Kuan, A. Sawchuk, T. Strand, and P. Chavel, "Adaptive noise smoothing filter for images with signal-dependent noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 2, pp. 165–177, Mar. 1985.

[34] G. Wasson, D. Kortenkamp, and E. Huber, "Integrating active perception with an autonomous robot architecture," *Robot. Auton. Syst.*, vol. 29, nos. 2–3, pp. 175–186, 1999.

[35] R. Duda, D. Stork, and P. Hart, *Pattern Classification and Scene Analysis. Part 1, Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, Nov. 2000.

[36] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[37] K. Murphy, *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning Series)*. Cambridge, MA, USA: MIT Press, Aug. 2012.

[38] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[39] J. F. C. Kingman, *Poisson Processes*, vol. 3. New York, NY, USA: Oxford Univ. Press, 1993.

[40] D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford, CA, USA: Center for the Study of Language and Information, Dec. 1999.

[41] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[42] R. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," presented at the *Int. Conf. Robot. Autom. (ICRA)*, Shanghai, China, May 2011, pp. 1–4.

[43] A. Baak, M. Muller, G. Bharaj, H. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1092–1099.

Kai-Chi Chan received the B.Eng. (Hons.) degree in computer engineering and the M.Phil. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA.

His current research interests include computer vision and machine learning.

Cheng-Kok Koh received the B.S. (Hons.) and the M.S. degrees, both in computer science, from the National University of Singapore, Singapore, in 1992 and 1996, respectively, and the Ph.D. degree in computer science from the University of California at Los Angeles, Los Angeles, CA, USA, in 1998.

He is currently a Professor of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His current research interests include physical design of VLSI circuits and modeling and analysis of large-scale systems.

Dr. Koh was a recipient of the ACM Special Interest Group on Design Automation Meritorious Service Award, the Distinguished Service Award, the National Science Foundation CAREER Award, and the Semiconductor Research Corporation Inventor Recognition Award.

C. S. George Lee (S'71–M'78–SM'86–F'93) received the Ph.D. degree from Purdue University, West Lafayette, IN, USA.

He is a Professor of Electrical and Computer Engineering at Purdue University. His current research interests include human-centered robotics, autonomous robots/systems, and neuro-fuzzy intelligent systems. He has published extensively in the above areas, including over 200 archival publications, two graduate textbooks, *Robotics: Control, Sensing, Vision, and Intelligence* (McGraw-Hill, 1986) and *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems* (Prentice-Hall, 1996), and 20 book chapters.

Prof. Lee was a recipient of the IEEE Third Millennium Medal Award, the Distinguished Service Award, and the Saridis Leadership Award from the IEEE Robotics and Automation Society.