2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Using Action Classification for Human-Pose Estimation

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

Abstract— **This paper presents a 3D-point-cloud system that extracts a 3D-point-cloud feature (VISH) from the observation of a depth sensor to reduce feature/depth ambiguity and estimates human poses using the result of action classification and a kinematic model. Based on the concept of distributed representation, a non-parametric action-mixture model is proposed in the system to represent high-dimensional human-pose space using low-dimensional manifolds in searching human poses. In each manifold, the probability distribution is estimated by the similarity of features. The distributions in the manifolds are then redistributed according to the stationary distribution of a Markov chain that models the frequency of actions. After the redistribution, the manifolds are combined according to the distribution determined by the action classification. In addition, the spatial relationship between human-body parts is explicitly modeled by a kinematic chain. Computer-simulation results showed that multiple low-dimensional manifolds can represent human-pose space. The 3D-point-cloud system showed reduction of the overall error and standard deviation compared with other approaches without using action classification.**

I. INTRODUCTION

Human-pose estimation has long been a challenging problem in robotics and computer vision. The main challenge concerns reducing feature/depth ambiguity and searching human poses in high-dimensional human-pose space. In our previous work [1], we used a depth sensor to reduce the depth ambiguity and proposed a 3D-point-cloud feature called VIewpoint and Shape feature Histogram (VISH) to capture the spatial ordering of orientation and shape of a human in a 3D point cloud to reduce the feature ambiguity. This work extends our previous work and focuses on incorporating an action-mixture model into the system to improve the performance of searching human poses in high-dimensional space.

In general, the search space grows exponentially with the dimensionality of human-pose space. Thus, searching human poses directly is intractable. Existing methods include dimensionality-reduction methods [2] [3] that project human poses in high-dimensional space to low-dimensional space. Human poses are then searched and estimated directly in the low-dimensional space. Unfortunately, the spatial relationship in human poses could be lost during the projection.

On the other hand, human poses are commonly assumed to lie in low-dimensional manifolds because of correlations between body parts. Thus, human poses can be estimated by discovering the low-dimensional manifolds and searching human poses in these manifolds. Prior knowledge, such as kinematic relationship [4]–[6], can be used to discover the manifolds. Furthermore, human actions have been used recently as prior knowledge [7]. Gall *et al.* [8] proposed a model for estimating the prior probability of an action from action classification to separate 3D-human poses into action-specific manifolds. Yao *et al.* [9] further extended the model of estimating the prior probability of action in [8] by using action classification [7].

In this paper, we extend our previous work on a 3D-point-cloud feature VISH [1] to a 3D-point-cloud human-pose estimation system by incorporating an action-mixture model to represent human poses and a kinematic model of a human into the system to improve the performance of searching human poses in high-dimensional space. A non-parametric action-mixture model (AMM) is proposed in the system to map a VISH-feature input to a corresponding human pose. It uses low-dimensional manifolds associated with human actions to represent human-pose space. Using the concept of distributed representation to represent human poses, a human pose may appear in more than one action, which is different from the previous models [8] [9]. For example, the human pose of hand-waving can appear in the actions of standing and raising both arms. The probability distribution of human poses in each manifold is estimated by the similarity of features. Then, the distributions in the manifolds are redistributed according to the stationary distribution of a Markov chain that models the frequency of actions. Finally, the action of a VISH-feature input is classified to determine the weighting coefficients in combining the low-dimensional manifolds. Since the proposed AMM is derived based on an instance-based learning algorithm, its human-pose estimates are in discrete space. We use a kinematic model of a human to further refine the human-pose estimates in continuous space to reduce the quantization error. Computer simulations showed that multiple low-dimensional manifolds can be used efficiently to represent the high-dimensional human-pose space. The overall error and standard deviation of human-pose estimates using the proposed 3D-point-cloud system were the lowest compared with other approaches without using the low-dimensional manifolds.

II. SYSTEM FRAMEWORK

The proposed 3D-point-cloud human-pose estimation system takes a 3D point cloud as input and produces estimates of joint positions of a human as output. Figure 1 shows the framework of the proposed system. From the 3D-point-

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, CNS-0960061 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.

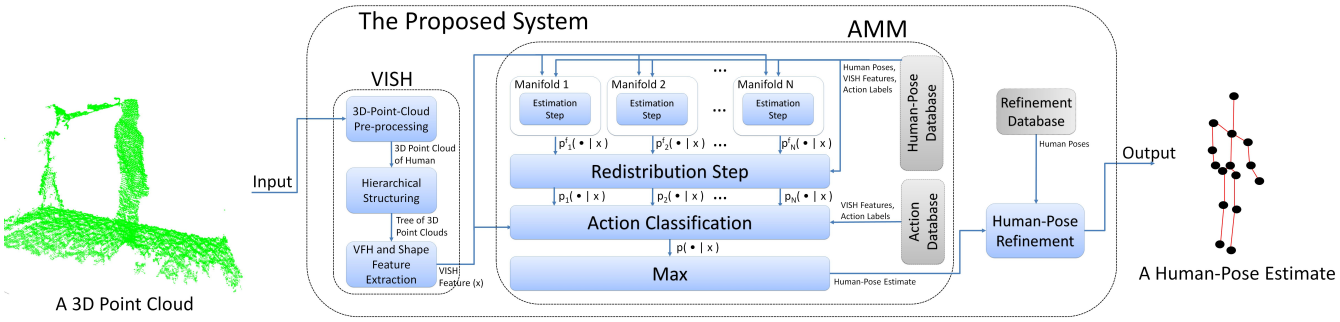

Fig. 1: The proposed 3D-point-cloud system for human-pose estimation. 3D point cloud is represented by a VISH feature to capture the spatial ordering of orientation and shape from a human. The proposed AMM then uses multiple low-dimensional manifolds to represent the human-pose space in searching the human pose corresponding to the VISH-feature input. The refinement model refines the human-pose estimate from AMM in continuous space to reduce the quantization error. The action and refinement databases are prepared in advance for training the action classifier and the refinement model (a kinematic model), respectively. The human-pose database is prepared in advance to represent the discrete space of human poses in each manifold.

cloud input, VISH feature is extracted to capture the spatial ordering of orientation and shape of a human. Details of the VISH feature and its extraction can be found in [1].

The proposed AMM represents human-pose space using low-dimensional manifolds, each of which represents the space of human poses in one human 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. The redistribution step assigns weights to the probability distributions estimated from the estimation step in the manifolds 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. The human pose with the highest probability will be considered as the human-pose estimate corresponding to the VISH-feature input. The proposed AMM is described in more details in Section III.

In human-pose refinement, the human pose estimated by the proposed AMM is refined in continuous space to reduce the quantization error in the AMM. The spatial relationship between body parts is used to refine human-pose estimates. 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 IV.

III. Action-Mixture Model

From a 3D point cloud, VISH feature is extracted as described in [1]. The action-mixture model (AMM) then maps a VISH-feature input to a corresponding human-pose estimate based on an instance-based learning algorithm [10]. The learning algorithm can train a non-parametric model (as shown in Eq. (1)) to represent the human-pose space 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 i-th action are grouped to form a subset of the human-pose database, denoted as $\mathcal{D}_i$.

Using the concept of distributed representation, a human pose can lie in multiple low-dimensional manifolds, where each manifold corresponds to one action. Mathematically, the conditional probability of a human pose y is given by

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

where x is a VISH-feature input, $g_i(\cdot)$ is a weighting function for the i-th manifold (action), $p_i(y \mid x)$ is the conditional probability (base distribution) of a human pose in the i-th manifold, N is the number of manifolds and Z is a normalizing constant.

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,

$$y^* = \arg\max_{y \in Y} p(y \mid x), \tag{2}$$

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 for each manifold and the base distribution of human-pose estimation in each manifold. In determining the weighting function, we use an unnormalized probability

mass function (pmf) of action classification to represent the weighting function. The unnormalized pmf is trained using an action database, which contains VISH features and the corresponding action labels. We use the bootstrap aggregating algorithm (i.e., bagging) [11] to train the action classifier because it has been shown successful in classifying actions. Details on how to train the action classifier can be found in [11]. After training, the weighting function of each manifold is represented by the unnormalized pmf to determine a weighting coefficient for each manifold for every VISH-feature input x.

The base distribution is modeled 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 in each manifold associated with an action. The redistribution step assigns weights to the probability distributions in the manifolds estimated in the estimation step according to the frequency of actions. Note that the redistribution step is different from the action classification that the redistribution step utilizes the temporal information of actions in the human-pose database.

A. Estimation Step

The probability distribution of human poses in the human-pose database is estimated in each manifold associated with an action based on the inverse of the Euclidean distance between the VISH-feature input and VISH features of the human poses. The unnormalized probability of a human pose in the i-th action, $\forall i = 1, 2, \ldots, N$, is defined as

$$\widetilde{p}_i^f(x' \mid x) = \begin{cases} \frac{1}{\|x-x'\|_2+z} & \text{when } x' \in A_i, \\ \frac{1}{N} \cdot \frac{1}{\|x-x'\|_2+z} & \text{when } x' \notin A_i, \end{cases} \quad (3)$$

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 VISH features, each of which corresponds to a human pose in $\mathcal{D}_i$, z is a small constant to avoid division by zero and $\|\cdot\|_2$ is the Euclidean norm. Hence, the unnormalized probability of a human pose is larger if the VISH-feature input is closer to the VISH feature of the human pose in the Euclidean space. We include the factor $\frac{1}{N}$ in Eq. (3) to penalize the case when the action associated with a VISH feature from the human-pose database is different from the action of the unnormalized probability. The probability of a human pose in the i-th action is then given by

$$p_i^f(x' \mid x) = \frac{\widetilde{p}_i^f(x' \mid x)}{\sum_{x' \in A} \widetilde{p}_i^f(x' \mid x)}, \quad (4)$$

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

B. Redistribution Step

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 past actions are irrelevant for predicting the future actions. The weight is formulated as the stationary probability in a continuous-time Markov chain, which is trained using the VISH features in the human-pose database. Let $X(t)$ be the continuous-time Markov chain with the state space $I = \{1, 2, ..., N\}$ 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

$$Q(i,j) = \begin{cases} \lim\limits_{h \to 0} \frac{p(X(h)=j \mid X(0)=i)}{h} & \text{when } i \neq j, \\ \lim\limits_{h \to 0} \frac{p(X(h)=i \mid X(0)=i)-1}{h} & \text{when } i = j. \end{cases} \quad (5)$$

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 [12] 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) [13] that measures the similarity between two actions; that is,

$$\lambda_{ij} = \begin{cases} \frac{z_1}{DTW(i,j)+r} & \text{when } DTW(i,j) < \tau, \\ 0 & \text{when } DTW(i,j) \geq \tau, \end{cases} \quad (6)$$

where z_1 is a constant, τ is a predefined threshold, $DTW(i, j)$ is the distance between the i-th and j-th 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 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 3D point clouds in the two frames.

As one 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

$$(1 - \frac{\lambda_{ij}}{m})^{tm} \xrightarrow[m \to \infty]{} e^{-\lambda_{ij} t}. \quad (7)$$

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

$$Q(i,j) = \begin{cases} \lambda_{ij} & \text{when } i \neq j, \\ -\sum_{k=1, k \neq i}^{N} \lambda_{ik} & \text{when } i = j. \end{cases} \quad (8)$$

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 k-th 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,

$$p_i^a(x' \mid x) = \pi_k(i) \sum_{j \in I_k} p_j^f(x' \mid x), \quad i \in I_k, \tag{9}$$

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,

$$\tilde{p}_i(y \mid x) = u^f p_i^f(x' \mid x) + u^a p_i^a(x' \mid x), \tag{10}$$

where x' is the VISH feature 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,

$$p_i(y \mid x) = \frac{\tilde{p}_i(y \mid x)}{\sum_{y \in D_i} \tilde{p}_i(y \mid x)}. \tag{11}$$

Using the proposed AMM, the probability distribution of human poses in each manifold is modeled in the estimation step. The probabilities of human poses are redistributed according to the frequency of actions in the redistribution step. The classification result then aggregates the base distributions from all actions. The human pose with the highest probability in the human-pose database will be considered as the human-pose estimate of the VISH-feature input.

IV. Kinematic Model

As the human poses estimated by the AMM are in discrete space, we use a kinematic model as shown in Figure 2(a) 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 $G = (V, E)$ as shown in Figure 2(b), where V corresponds to a set of vertices and E corresponds to a set of edges. The vertex 1 is a softmax random variable [10] 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 on a 4D manifold. A (stochastic) configuration, C, of a human is a collection of body parts; that is, $C = \{O_2, O_3, \ldots, O_{15}\}$. Let $c, o_2, ..., o_{15}$ 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,

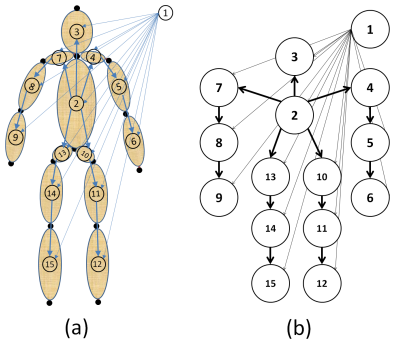

Fig. 2: (a) The kinematic model of a human. (b) The directed acyclic graph of the kinematic model. The vertices are: 1. human pose estimated by the proposed AMM, 2. torso, 3. head, 4. left shoulder, 5. left upper arm, 6. left lower arm, 7. right shoulder, 8. right upper arm, 9. right lower arm, 10. left 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.

$$\begin{aligned} p(c) = p(c \mid v_1) &= p(\{o_2, o_3, \ldots, o_{15}\} \mid v_1) \\ &= p(o_2, o_3, \ldots, o_{15}), \end{aligned} \tag{12}$$

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

$$p(C) = \prod_{s \in V^-} p(O_s \mid pa(O_s)), \tag{13}$$

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 Eq. (13) 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,

$$p(O_s \mid pa(O_s)) = \mathcal{N}(O_s \mid \mu_s, \Sigma_s), \tag{14}$$

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,

$$\sum_{n=1}^{|\mathcal{T}|} \log p(c^n) = \sum_{n=1}^{|\mathcal{T}|} \sum_{s \in V^-} \log p(O_s^n \mid pa(O_s^n)), \qquad (15)$$

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

When refining a human pose using the kinematic chain, body parts are divided into observable and unobservable groups. 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. Inference is made based on the orientation of the body parts in the observable group to estimate the orientation of the body parts in the unobservable group. Mathematically, the set of vertices in the graph G is divided into evidential (observable) and non-evidential (unobservable) sets; that is,

$$V^- = V_e^- \cup V_n^- \quad \text{and} \quad V_e^- \cap V_n^- = \emptyset, \qquad (16)$$

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,

$$c^* = \underset{c \in C,\ s \in V_e^-}{\arg\max}\ p(c \mid o_s). \qquad (17)$$

Based on the factorization structure of the distribution in the kinematic model, Eq. (17) can be calculated efficiently using belief propagation [14]. The orientation of body parts that maximizes Eq. (13) are then converted to a human pose and averaged with the human pose estimated from the AMM to output the refined human-pose estimate.

V. Computer-Simulation Results

The proposed 3D-point-cloud system was tested on the Stanford TOF Motion Capture Dataset [6]. There are 28 video sequences in the dataset. We considered each video sequence as one action. Some example actions are kicking and rotation. The ground-truth 3D joint locations of the subject were recorded by a commercial motion-capturing system. Frames with missing ground-truth 3D joint locations were ignored. The error metric, ζ, for each video sequence is defined as

$$\zeta = \frac{1}{N_f} \sum_{s=1}^{N_f} \frac{1}{N_s} \sum_{i=1}^{N_s} \|\mathbf{j}_{s,i} - \tilde{\mathbf{j}}_{s,i}\|_2, \qquad (18)$$

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

The dataset was divided into 20% for establishing the human-pose database, 30% for building the refinement database, 40% for building the action database and 10% for testing. To reduce the bias in dividing the dataset, the dataset was randomly divided 10 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 Eq. (3) was set to 0.1. In the redistribution step, the constant z_1 in Eq. (6) 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 neighbor using VISH features.
2) VISH+AMM: A human pose was estimated by the non-parametric AMM using VISH features.
3) VISH+AMM+Kinematic Model: A human pose was estimated by the proposed 3D-point-cloud system.

Table I 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.

TABLE I: The 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.

	VISH	VISH+AMM	Proposed System
Overall	0.0291 (0.0351)	0.0268 (0.0326)	0.0246 (0.0294)

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 1 (one action) to 28 (all actions). The 28 actions were first sorted in a list in descending order according to the unnormalized pmf

of action classification. Then, the N_a actions were selected from the first N_a actions in the sorted list.

Figure 3 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 ceased to decrease.

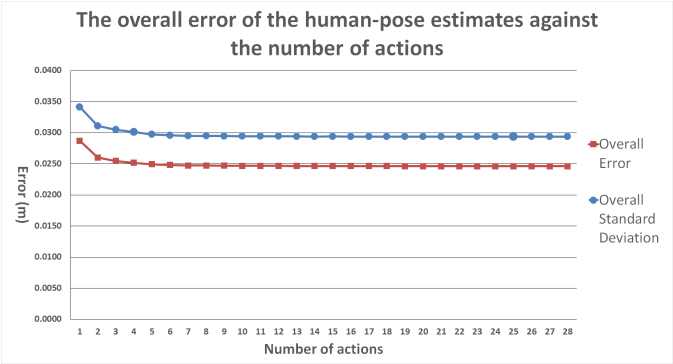

Fig. 3: The changes of the overall error and standard deviation of human-pose estimation using different numbers of actions.

The results of the proposed system were compared with the results reported in some existing works using the Stanford TOF Motion Capture Dataset [6]. The overall errors and standard deviations are shown in Table II. The proposed system incurred the lowest overall error and standard deviation, 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.

TABLE II: The overall errors and standard deviations (std. dev.) of human-pose estimation in the Stanford TOF Motion Capture Dataset [6].

	Error (m)	Std. Dev. (m)
HC and EP Method [6]	0.1	N.A.
Data-driven Hybrid Method [15]	0.0618	0.0424
Exemplar Method [16]	0.038	N.A.
Proposed 3D-Point-Cloud System	0.0246	0.0294

VI. Summary and Conclusions

In this paper, we have presented a 3D-point-cloud human-pose estimation system that extracts the 3D-point-cloud feature VISH from the observation of a depth sensor to reduce feature/depth ambiguity and estimates the corresponding human pose using the result of action classification and the kinematic model. Based on the concept of distributed representation, a non-parametric action-mixture model has been proposed to represent human-pose space using low-dimensional manifolds associated with actions in searching human poses.

In AMM, human poses in each manifold were modeled using the base distribution without making stronger assumptions about the nature of the human-pose distribution compared with other parametric models such as an exponential family of distributions. Instead, the base distribution was estimated using an instance-based learning algorithm that measured the similarity of VISH-features in the estimation step and frequency of actions in the redistribution step. The action of a VISH-feature input was then 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, a 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 could 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] K. C. Chan, C. K. Koh, and C. S. G. Lee, "A 3D-point-cloud feature for human-pose estimation," in *ICRA2013*, May 2013, pp. 1615–1620.

[2] X. He and P. Niyogi, "Locality preserving projections." Cambridge, MA: MIT Press, 2004.

[3] Y. Tian, L. Sigal, H. Badino, F. De la Torre, and Y. Liu, "Latent Gaussian mixture regression for human pose estimation," in *Asian Conference of Computer Vision*, 2011, pp. 679–690.

[4] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," in *CVPR*, vol. 1, Jun. 2004, pp. 421–428.

[5] 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*, vol. 98, no. 1, pp. 15–48, May 2012.

[6] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Computer Vision and Pattern Recognition*, Jun. 2010, pp. 755–762.

[7] A. Yao, J. Gall, G. Fanelli, and L. van Gool, "Does human action recognition benefit from pose estimation?" in *British Machine Vision Conference*, 2011, pp. 67.1–67.11.

[8] J. Gall, A. Yao, and L. van Gool, "2D action recognition serves 3D human pose estimation," in *ECCV*, 2010, pp. 425–438.

[9] A. Yao, J. Gall, and L. van Gool, "Coupled action recognition and pose estimation from multiple views," *IJCV*, vol. 100(1):16–37, 2012.

[10] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, Aug. 2012.

[11] L. Breiman, "Bagging predictors," in *Machine Learning*, vol. 24, 1996, pp. 123–140.

[12] J. F. C. Kingman, *Poisson processes*, New York, 1993, vol. 3.

[13] D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, 1999.

[14] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

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

[16] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3D pose estimation from a single depth image," in *International Conference on Computer Vision*, Nov. 2011, pp. 731–738.