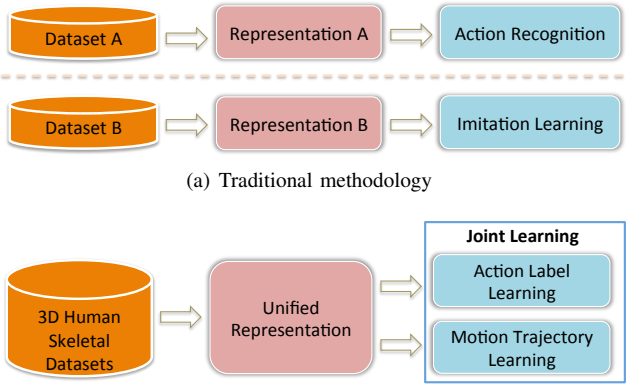
25th IEEE International Symposium on
Robot and Human Interactive Communication (RO-MAN)
August 26-31, 2016. Columbia University, NY, USA

Unified Robot Learning of Action Labels and Motion Trajectories from 3D Human Skeletal Data

Chi Zhang[1], Hao Zhang[2], Rui Guo[1], and Lynne E. Parker[1]

***Abstract*— Currently, robot learning of human activities is mainly studied in two largely disconnected domains: high level semantics understanding in human activity recognition, and low level motion trajectory reproduction in robot imitation learning. The critical problem of *human activity unified learning* (HAUL) was not well studied in previous work. One important challenge is the lack of a representation that can be learned from both levels. We introduce a novel approach to address this HAUL problem at the representation level, by simultaneously learning action labels and motion trajectories from publicly available 3D human skeletal datasets, thus avoiding additional human labor for data collection. Our approach builds a subject and body position independent shared skeleton, and extracts features of skeletal activities based on this model. Then the extracted features are encoded by the parameter set of Gaussian Mixture Models to construct the unified representations. The proposed compact representation can be directly applied to identify activity labels when combined with Support Vector Machines, and can be also employed to generate trajectories of the learned activities when combined with Gaussian Mixture Regression on a robot. Finally, an inverse kinematic mapping is developed to transfer human skeletal trajectories to joint angle sequences in the robot's embodiment. Empirical studies using simulation and real humanoid robots demonstrate that our approach achieves promising performance on robot unified learning of human action labels and motion trajectories, effectively addressing the HAUL problem.**



(a) Traditional methodology

(b) The proposed framework.

Fig. 1. Motivations of this work: traditional robot learning methodologies treat human behavior understanding and imitation learning as two separate problems, using different datasets and representations. In addition, previous research on imitation learning generally requires significant human efforts to collect demonstration data. In this paper, we introduce a novel unified learning approach based on compact human representation that can enable a humanoid robot to simultaneously learn both action labels and motion trajectories. To reduce human effort, we propose the use of existing whole-body 3D human skeletal data publicly available on the Internet to construct our representation and realize unified robot learning.

I. Introduction

The ability to automatically understand and imitate human actions is essential for intelligent robots to learn from people and to effectively assist and interact with humans in many human-centered robotics domains, including home assistance, service robotics, and search and rescue. Although human action recognition [1], [2] and robot imitation learning [3] have been widely studied for many years in robotics research, these two problems were addressed independently using separate representations of the human. The critical problem of simultaneous learning of human action labels and motion trajectories has not been previously well studied. We name this research problem as *human action unified learning* (HAUL), as illustrated in Fig. 1.

Previous methods of activity recognition allow robots to understand the semantics of human behaviors by representing the spatial information of global human activities using local features, silhouettes or even a point [4], [5], and by encoding the temporal information using histograms [5]; however, all these representations are inapplicable for reproducing their motion trajectories. Methods of imitation learning [6], [7] address the trajectory learning problem by providing a robot with human demonstrations, which are regarded as an initial guideline for performing human-like tasks. However, such approaches usually learn low level trajectories for a single activity, without the capability of understanding the semantic meanings of a large variety of human activities. Therefore, a unified learning is desired for intelligent robots to simultaneously recognize human activities and reproduce the activities.

Traditional imitation learning approaches collect data by manually moving the robot's body to perform activities and record data from the robot's sensors [3]. Such a dataset gathering method is often time-consuming and requires much human labor. Typically, only few activities (usually less than 5) are demonstrated by the human [6], [7], making them inapplicable for teaching robots a large number of activities. In addition, since the size of the demonstration data is usually small (e.g., human teaches an activity with 4-7 demonstrations [6]), fault tolerance cannot be guaranteed. As a result, the performance of robot imitation learning is inherently limited by the availability and quality of the demonstration data.

In recent decades, 3D sensors, including color-depth cameras, allow a cheap acquirement of whole-body human

This work is supported by the National Science Foundation under Grant No. IIS-1427004.

[1]Chi Zhang, Rui Guo and Lynne E. Parker are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA. {czhang24,rguo1,leparker}@utk.edu

[2]Hao Zhang is with the HCRobotics Lab in the Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA. hzhang@mines.edu

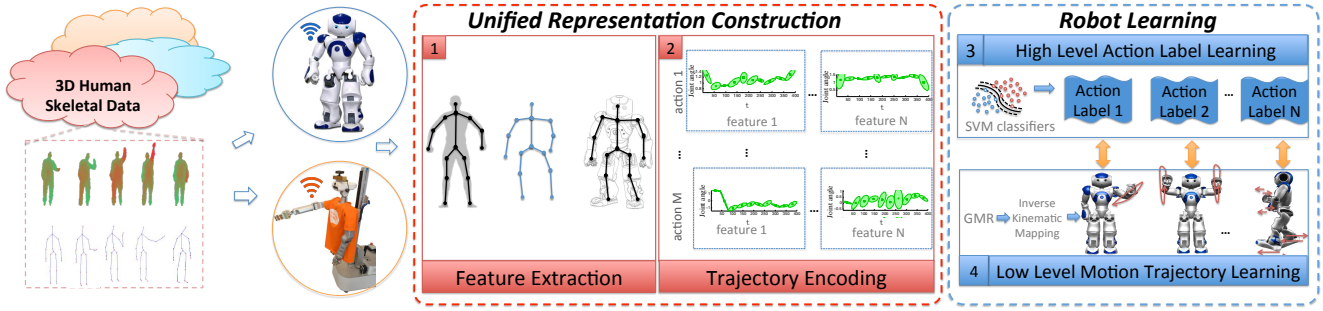


Fig. 2. Framework of unified robot learning of action labels and motion trajectories, based on the proposed representation, from existing 3D human skeletal data. After accessing the datasets, a robot learns by 1) extracting features from the skeletal data that are irrelevant to human body scales and positions based on a shared skeleton model, 2) encoding activity trajectories of each feature using probabilistic models to compactly represent human activities, 3) recognizing action labels by feeding the representation to a classifier, and 4) reproducing generalized motion trajectories of each activity using regression and mapping them to robot's joint space. Finally, the robot associates the learned action labels with motion trajectories to accomplish the united learning.

motion data. The datasets cover a wide range of human daily activities; a large amount of demonstrations of a certain activity can be collected due to the easy way of performing demonstrations (e.g., tens to hundreds activities, and hundreds to thousands of instances in total [8]). Accordingly, the gathered data are considerably larger than the traditional imitation learning methods can provide. Such datasets naturally become good resources for robot imitation learning. However, using the existing large-scale 3D human activity datasets to teach robots human activities has several challenges. First, substantial variations are inherently contained within an activity, especially when performed by different humans. Even the demonstrations performed by the same subject vary in speeds and poses. These variations make it difficult to generalize motion trajectories from large-scale human activity data. Moreover, the recorded activity data can be defective and ambiguous (e.g., self-occlusion can occur in human performances, reducing the quality of the recorded data). Third, to effectively assist and interact with people, a robot must perform activity recognition and trajectory reproduction in real time, which makes most of the state-of-the-art activity recognition methods less applicable in robotics applications.

To address the HAUL problem and the aforementioned challenges, we propose a novel unified learning approach based on the representation of humans that bridges the gap between human action understanding and robot imitation learning, as Fig. 2 shows. We first build a shared skeleton model to extract subject and body position independent features in the 3D human skeletal data. Then, the spatial-temporal patterns of extracted features are encoded using the compact set of Gaussian Mixture Models (GMMs) to construct unified representations of human activities. The proposed representation can be directly used by a robot to identify human activities and reproduce motion trajectories of the activities in real time. To reason about the meanings of human activities, the proposed representation can be combined with a classification method such as the Support Vector Machine (SVM). To reproduce the activities, the same representation can be directly used with Gaussian Mixture Regression (GMR) to generate the human motion trajectories, and finally project them onto the robot's embodiment using an inverse kinematics mapping. Note that given the human skeletal data, our approach learns motion trajectories (i.e., imitates the poses sequence) without interacting with real objects to accomplish the tasks.

The key contributions of this paper are twofold:

1) We identify the important HAUL problem and propose a novel unified learning approach based on human representation that enables a robot to simultaneously interpret human actions and reproduce the motions to address the HAUL problem in real time. The approach bridges the gap between human activity recognition and robot imitation learning.
2) We propose to adopt the publicly available 3D human skeletal data to address the demonstration data availability issue in the imitation learning methodology. An efficient incremental learning approach is also implemented to enable a robot to learn from the challenging, highly variable data. Together, this innovation has the potential to significantly reduce human efforts on collecting large-scale demonstration data for robot training.

The rest of the paper is structured as follows. In Section II, we overview related work on human activity recognition and imitation learning. Section III discusses our new approach in detail. Results of empirical studies are presented in Section IV. After discussing the characteristics of our approach in Section V, we conclude the paper in Section VI.

II. Related Work

Several methods for 3D human activity representation have been proposed in the past few years. Previous works that 1) use a 3D centroid trajectory to describe the human activity with a point [9], 2) represent global human activities with local features [10], or 3) extract human silhouettes [11] are not applicable to robot imitation learning because these representations cannot be used in robot motor learning. Due to the similarity between the human skeleton and the humanoid robot's embodiment, the approaches based on a 3D human model, (e.g., a 3D human skeleton model [12], [13], [14]), naturally fit the scope of robot imitation learning. Sequence of the Most Informative Joints (SMIJ) [4] is such an approach that represents human actions with the most informative joints in human skeleton, which interprets the physical meaning of different activities. The trajectories of

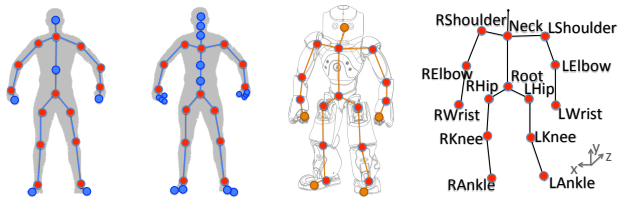
3D human skeleton are also described by a Histogram of Oriented Displacements (HOD) of three 2D projections in each joint for the action recognition purpose [5]. Even though these skeleton-based approaches yield good performance in activity classification, the discretization of the entire trajectory with a histogram loses certain details (e.g., order of motions) in the trajectories, and is thus unable to reproduce motion trajectories. In this paper, we tackle human activity recognition in the HAUL problem by proposing a new representation to describe the human activities, which can also be applied to learn the motion trajectories.

In robot imitation learning, how the demonstrations are recorded and what platform is used to acquire the data vary greatly across approaches. Examples range from a human holding and moving the robot's body parts, to the human teleoperating the robot. Since teleoperation (by either using a controller or by wearing motion detection sensors [15] on the human body), requires great human expertise, and the robot execution is often slow, building a dataset becomes tedious and expensive. Directly holding and moving a robot to perform activities [6], [16] is also difficult if the motion is high speed [16], or even impossible if many degrees of freedom (DoF) are involved (e.g., a whole body motion like walking). Recently, the development of perception systems such as cameras and depth sensors makes human activity datasets cheap and publicly available [12], [17]. In this paper, we leverage the existing massive 3D human activity data as the implicit demonstrations from which the robot learns.

A large number of works address robot imitation learning as a regression problem and solve it with probabilistic models. GMM [6] generalizes a smooth trajectory from human demonstrations and adapts to small variances as well. Hidden Markov Model (HMM) approach [7] encodes temporal-spatial information in trajectories and then multiple HMMs are employed to retrieve new generalized motions. Due to the efficiency of GMM in capturing characteristics embedded in demonstrated trajectories, it is adopted in this paper to encode the human activity data. While some prior approaches recognized explicitly demonstrated human activities and transferred the activities to robots [18], they did not use the existing large-scale 3D human activity data as implicit demonstration from which the robot can learn. Thus, we believe our proposed work is the first to develop a robot learning mechanism that utilizes the existing human activity data there are relatively in a large-scale, as implicit demonstrations for robot imitation learning. By considering constraints, a large body of in imitation learning works propose methods for object robot interactions, such as grasp planning [19], Dynamic Systems [20], and human-like motion generation [21], which can be easily combined with our approach to improve the motion trajectory learning.

III. Unified Robot Learning

To address the HAUL problem, we develop a novel unified robot learning framework, with an ultimate goal of allowing robots to efficiently learn from human activity datasets, which also tackles the problem of expensive human demonstrations in robot imitation learning. We begin the discussion by developing a shared skeleton model to extract the subject and body position independent features of human activities. Then, we present our representation approach to efficiently encode the extracted features of human activities. The constructed representation is then fed into classifiers to learn the semantic meaning of human activities. Finally, by learning the representation with an incremental training model, the generalized motion trajectories of human activities are mapped by an inverse kinematic model and reproduced by robots.



(a) Kinect SDK (b) MoCap (c) NAO (d) Shared skeleton

Fig. 3. Examples of skeletal kinematic human body models obtained from different 3D perception technologies. Skeleton data acquired from Microsoft Kinect contains 20 joints as shown in (a), and 31 joints from a MoCap system in (b). NAO robot has fewer joints than a human, as (c) shows. By using the joints in common (red dots), a shared model is built in (d).

A. Feature Extraction

Acquisition of the 3D human skeletal data can be achieved by various techniques, such as the MoCap systems and Kinect sensors. However, these 3D perception systems yield different human skeleton models, as Fig. 3(a) and 3(b) present. In addition, humanoid robots usually have different embodiments in terms of joints, DoFs, etc., than the human skeleton, as the NAO robot in Fig. 3(c) shows. To consistently process the 3D human skeletal data acquired by different perception systems and smoothly transfer them to the robot's embodiment, we extract the common joints in the human skeletons and robot skeleton to build a shared skeleton model, as showed in Fig. 3(d).

A commonly used feature to describe human skeletal actions is a time series collection of 3D positions, i.e., 3D trajectories of joints. However, this feature does not scale with various body position and orientation. To address these issues, we use the difference of each two adjacent limb vectors to depict human skeletal activities. The limb vectors are defined by each two adjacent joints. For example, given joint positions p_1, p_2, p_3 of the left shoulder, left elbow, and left wrist, the bone vector of the upper left arm can be represented as $\overrightarrow{p_2p_1}$, and the bone vector of the lower arm is $\overrightarrow{p_3p_2}$. Specifically, given these two vectors, our approach competes the subtraction:

$$\xi = \overrightarrow{p_2p_1} - \overrightarrow{p_3p_2} \tag{1}$$

where vector ξ starts from point $p2$, and indicates the change of direction and scale of two adjacent limbs, as showed in Fig. 4. Consequently, the sequences of ξ are features describing the human skeletal activities, which are invariant to the absolute body position, the initial body orientations, and camera view. Formally, the 3D trajectory performed by the shared model is $\{\xi_i^n\}, i = 1, ..., L, n = 1, ..., N$, where L is the length of the sequence, and N is the total number of subtraction vectors.

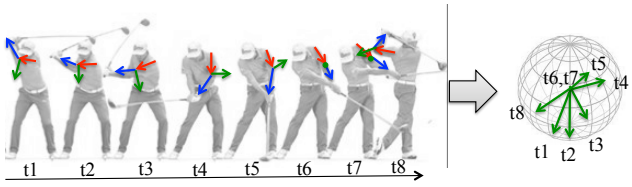


Fig. 4. An example of the extracted features in the golf-swing activity. The upper left arm vectors and lower left arm vectors are respectively denoted by red and blue arrows. The green arrows represent subtractions between the two limbs. Note that at t_6 and t_7, the computed vectors are zeros (singularities). Since the actual situation is that a subsequent limb and its preceding limb are in the same direction, we replace them by a vector having the same direction as the preceding vector(red), and the same length as the subsequent vector (blue).

B. Unified Representation Construction

In existing human activity datasets, spatial and temporal variations are common when multiple subjects perform activities for multiple times in different environments. In this paper, Dynamic Time Warping (DTW) with Euclidean distance metric is adopted to align the temporal difference, which has has been shown to be a robust and computationally efficient method [6]. By integrating the aligned temporal information with the spatial information, a data point in an activity sequence can be rewritten as $\xi_i^n = \{\xi_{t,i}^n, \xi_{s,i}^n\}$, where $\xi_{t,i}^n$ corresponds to the temporal information, i.e., time stamps, $\xi_{s,i}^n$ represents the spatial information, i.e., joint positions. The superscript n is omitted for the sake of simplicity in the rest of this section.

To extract the key characteristics of a human activity, we apply mixture modeling to encode the sequence of $\{\xi_i\}$ by a mixture of K Gaussians, and with the probability density function:

$$p(\xi_i) = \sum_{k=1}^{K} p(k)p(\xi_i|k) \tag{2}$$

where $p(k) = \pi_k$ is the prior, and $p(\xi_i|k) = \mathcal{N}(\xi_i; \mu_k, \sigma_k)$ is a Gaussian component denoted by:

$$p(\xi_i|k) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} \exp\{-\frac{1}{2}(\xi_i - \mu_k)^T\Sigma_k^{-1}(\xi_i - \mu_k)\} \tag{3}$$

The *Expectation Maximization* (EM) method [22] is used to obtain the maximum likelihood estimation of parameters in the GMM.

Then we use the temporal information as queries for each GMM to derive the spatial information (joint positions) from the mean and covariance matrix of the Gaussians:

$$\mu_k = \{\mu_{t,k}, \mu_{s,k}\} \tag{4}$$

$$\Sigma_k = \begin{pmatrix} \Sigma_{t,k} & \Sigma_{ts,k} \\ \Sigma_{st,k} & \Sigma_{s,k} \end{pmatrix} \tag{5}$$

Consequently, the representation of one feature in one training sample can be constructed as $\lambda = \{\mu_k, \Sigma_k, \pi_k\}, k = \{1, ..., K\}$. For the n-th feature of the i-th training sample labeled with m, we rewrite the representation as λ_{mn}^i, as the blue block in Fig. 5 shows. Then we concatenate models of all features as a compact representation to describe one training sample, as the highlighted red block in Fig. 5 presents. Usually the optimal number of Gaussian components K is decided by the *Bayesian Information Criterion* (BIC) [23] method to balance the fitness of the model and the complexity of training. In this paper, since the same length of input vectors are needed the SVM, a fixed number of K components for each GMM is experimentally selected.

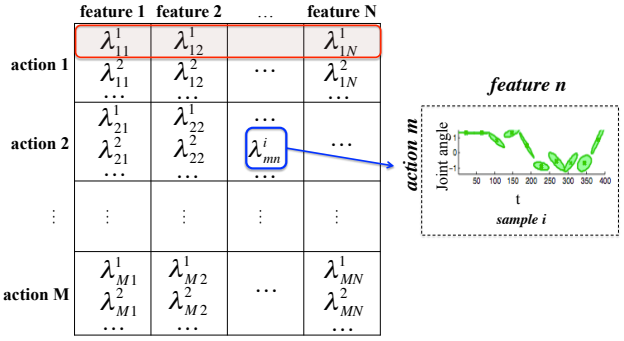


Fig. 5. The matrix of the proposed representations of human activities. Each feature of one human activity sample is encoded by a GMM (λ in the blue block). The encoded features are then combined to construct the representation of one human activity sample (red block).

C. Action Semantics Learning

After the dynamics of human activities are captured by the unified representation, the input vectors for an SVM classifier are built accordingly. Given a sample of one human activity, e.g., one row in Fig. 5, we use the $\mu_k, k = \{1, ..., K\}$ of each feature, i.e., spatial dimension $\xi^n, n = \{1, 2, ..., N\}$ and the temporal dimension t as the input vector to SVM, as showed in Eq. 6. Consequently, the length of an input vector is $2 \times K \times N + 1$ (label), where K is the number of Gaussian components and N is the number of features.

$$v = \{lab., t_1^{\xi_1}, \mu_1^{\xi_1}, ...t_K^{\xi_1}, \mu_K^{\xi_1}, ..., t_1^{\xi_N}, \mu_1^{\xi_N}, ...t_K^{\xi_N}, \mu_K^{\xi_N}\} \tag{6}$$

Given a large amount of training instances in the datasets, e.g., p instances performed by s subjects and labeled with M classes of activities, there are multiple ways to create the input vectors from the encoded GMMs for the SVM classification. In this paper, we design two modes to generate the vectors: 1) The *individual mode* takes each instance individually as one learning sample to build p compact representations, and associating each representation with the corresponding action label to generate a vector. Therefore, p vectors are created for training in the individual mode. 2) The *batch mode* combines instances from the same subject to obtain s samples and corresponding s representations for one class of action, leading to $s \times M$ vectors.

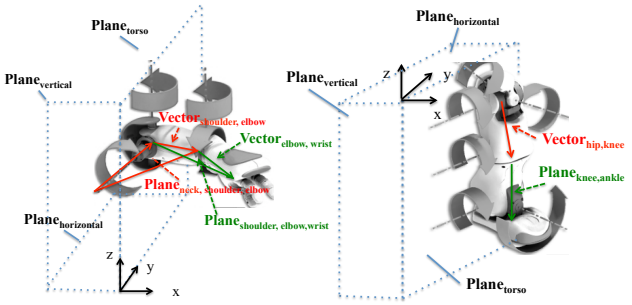
D. Motion Trajectory Learning

To learn the motion of human activities, we apply GMR to generalize a single trajectory from the learned GMM. Given the temporal information ξ_t, the conditional expectation $\hat{\xi}_{s,k}$ and estimated conditional covariance $\hat{\Sigma}_{s,k}$ are:

$$\begin{aligned} \hat{\xi}_{s,k} &= \mu_{s,k} + \Sigma_{st,k}(\Sigma_{t,k})^{-1}(\xi_t - \mu_{t,k}) \\ \hat{\Sigma}_{s,k} &= \Sigma_{s,k} - \Sigma_{st,k}(\Sigma_{t,k})^{-1}\Sigma_{ts,k} \end{aligned} \tag{7}$$

For each Gaussian component, the responsibility for ξ_t is:

$$\beta_k = \frac{p(\xi_t|k)}{\Sigma_{i=1}^{K} p(\xi_t|i)} \tag{8}$$



(a) Left arm (b) Left leg

Fig. 6. Mapping joint positions to robot's joint angles.

Combining Eqs. (7) and (8), the conditional expectation of ξ_s and the conditional covariance Σ_s can be computed:

$$\xi_s = \sum_{k=1}^{K} \beta_k \xi_{s,k}, \quad \Sigma_s = \sum_{k=1}^{K} \beta^2 \Sigma_{s,k} \tag{9}$$

Consequently, the spatial information ξ_s and the associated covariance matrix are derived by the time information ξ_t . Note that the generalized result is not the average or part of the dataset, but rather the inherent characteristics implied in the observed activities.

To achieve the good generalization in motion trajectory learning and real time performance, an incremental approach [25] is adopted to gradually add each instance to existing models without re-computing preceding data. Note that for each activity, all instances are used in the incremental learning, to obtain motion trajectories generalized through different subjects.

Given the generated ξ vectors in the shared skeleton, we can calculate the position of each joint in the shared model, and then use the developed inverse kinematic models of each limb to map joint positions to the robot's joint angles. The mapped joint angles are then used to control the robot directly, which addresses the correspondence problem [26]. Example of the left arm and left leg inverse kinematic models are shown in Eq. 10 and 11.

$$\begin{cases} ShouderPitch = f(p(nk, shd, elb), Plane_{horiz.}) \\ ShoulderRoll = f(p(nk, shd, elb), Plane_{vert.}) \\ ElbowRoll = f(v(shd, elb), v(elb, wrist)) \\ ElbowYaw = f(p(nk, shd, elb), p(shd, elb, wst)) \end{cases} \tag{10}$$

$$\begin{cases} HipRoll = f(v(hip, knee), Plane_{vertical}) \\ HipPitch = f(v(knee, ankle), Plane_{torso}) \\ KneePitch = f(v(hip, knee), v(knee, ankle)) \\ AnklePitch = f(v(knee, ankle), v(ankle, foot)) \end{cases} \tag{11}$$

where $p(\cdot)$ and $v(\cdot)$ are the fitting plane and vector defined by given points, and $f(\cdot)$ is the function of calculating angles between vectors and planes. $Plane_{torso}$ is a plane defined by the torso joints *RShoulder, LShoulder, Neck, Root, RHip, LHip*. $Plane_{vertical}$ is a plane perpendicular to $Plane_{torso}$ and intersects in $v_{neck,root}$, as showed in Fig. 6 shows. Accordingly, $Plane_{horizontal}$ is perpendicular to $Plane_{torso}$ and $Plane_{vertical}$, and intersects them in the root joint.

IV. Experiments

To demonstrate the effectiveness of our approach, we conducted experiments using a physical humanoid robot. In this section, we first discuss the setups of our experiments. Then, we evaluate the action label learning as well as the motion trajectory learning performance of our approach on two challenging human activity datasets. Finally, generalization of learning results are quantitatively evaluated.

A. Experimental Setup

Our approach is implemented using a mixture of Matlab and Python programming languages on a Mac OS X machine with an i5 2.5G CPU and 6GB memory. The learner employed in this paper is the non-linear SVM [24] with RBF kernels, and the standard one against-one methodology is applied to address the multi-class classification problem.

The MSR Daily Activity 3D dataset [12] is a most widely used benchmark dataset in human activity recognition tasks. This dataset contains color-depth and skeleton information of 16 activity categories: *(1) drink, (2) eat, (3) read book, (4) call cellphone, (5) write on a paper, (6) use laptop, (7) use vacuum cleaner, (8) cheer up, (9) sit still, (10) toss paper, (11) play game, (12) lie down on sofa, (13) walk, (14) play guitar, (15) stand up, (16) sit down*. Each activity is performed by 10 subjects for 2 to 3 times in various office environments, leading to 320 data instances. In this paper, we use 192 instances for training and 128 instances for testing.

The second dataset we use to evaluate our approach is the HDM05 dataset [17] captured by the MoCap system. It has more than 70 activities in 10-50 realizations executed by five subjects amounting to roughly 1500 instances. In this paper we use 11 activities similar to settings in [4]: *(1) deposit floor, (2) elbow to knee, (3) grab high, (4) hop both legs, (5) jog, (6) kick forward, (7) lie down floor, (8) rotate both arms backward, (9) sneak, (10) squat, and (11) throw basketball* performed by 5 subjects. In this paper, 189 instances performed by 3 subjects are used for training and 99 instances from 2 subjects are used for testing. Since the GMM is capable of capturing the temporal information in sequences, it is unnecessary to align the testing instances. During the encoding of testing data, we only assign the same number of Gaussians as in the training data.

B. Action Label Learning

We investigate our unified representation's performance in the action label learning tasks. Experimental results in Table I and II indicate a classification accuracy of 72.8% for MSR Daily Activity 3D dataset and 93.44% for HDM05 dataset. The higher performance on HDM05 implies that the way of recording human activities has a considerable effect on the classification performance. Because the MoCap perception system (HDM05 dataset) uses the inertial and/or optical sensors attached to the human bodies, it can estimate the human skeletons more accurately than the Kinect system (MSR Daily Activity 3D dataset) using the RGB camera and the infrared depth sensor. Therefore, our proposed approach achieves better performance for the HDM05 dataset, and is yet robust for more challenging datasets like MSR Daily Activity 3D. We also tested a greater number of GMMs, and found that when the number is too large, e.g., $K > 20$,

TABLE I

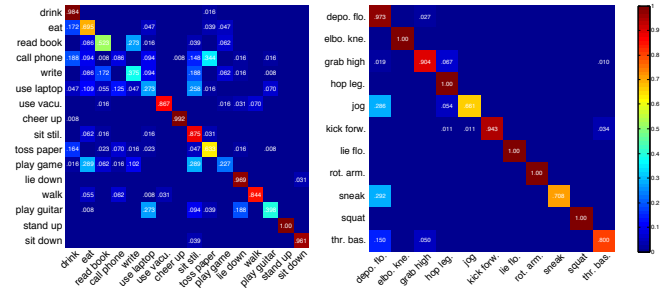
LEARNING ACCURACY OF ACTION LABEL IN MSR ACTIVITY DAILY 3D DATASET.

# of Gaussian Components	5	10	15	20
Individual mode	68.13%	66.56%	65.78%	65.00%
Batch mode	70.94%	**72.81%**	67.19%	69.69%

TABLE II

LEARNING ACCURACY OF ACTION LABEL IN HDM05 DATASET.

# of Gaussian Components	5	10	15	20
Individual mode	92.93 %	91.92 %	**93.44 %**	90.91%
Batch mode	86.36 %	84.09 %	80.68 %	84.09%



(a) MSR Daily Activity 3D (b) HDM05

Fig. 7. Confusion Matrices. Each column represents the instances in a predicted class, while each row is the instances in the ground truth category. Warmer colors denote better accuracy.

the performance decreases, since a large input vector may contain trivial details that are distracting for the SVM, and more GMMs also leads to a longer learning time. For the datasets used in this paper, $K \leq 20$ generally achieves a good performance, as shown in Table I and II.

To show the unified representation's action label learning performance, we also compare our approach with state-of-the-art skeleton-based representations in human activity recognition tasks, as presented in Table III and Table IV. These results show that our approach outperforms most works except our previous work of Bio-inspired Predictive Orientation Decomposition (BIPOD) [27]. Compared with the prior works directly using joint positions [12], [13], [14], [28], [29] to describe human activities, our proposed method employs preprocessed joint positions based on a shared skeleton model that are human scale and body position independent, so that the effect of configurations of perception system and subjects are eliminated. Compared with the histogram approach [4] that ignores the order of actions, mixture modeling adopted by our approach is capable of capturing the temporal information in human activity, leading to a better performance for reversal activity separation, such as *stand up* and *sit down* in Fig. 7(a). Moreover, we only use the discriminative joints in the shared model and abandon the other joints (e.g., fingers and feet in Fig. 3(d)) to describe the human activities, since the motions of these joints are either redundant or cannot be accurately estimated by the perception systems, which leads to a more compact and accurate representation. Generally, these advantages of our approach are desired to outperform most of the skeleton-based representations. The BIPOD approach deploying angles between temporally adjacent motion vectors has a higher classification accuracy, but it can only be used for human

TABLE III

COMPARISON OF RECOGNITION ACCURACY WITH PREVIOUS SKELETON-BASED REPRESENTATIONS ON MSR DAILY ACTIVITY 3D.

Skeleton-based representations	Accuracy
Dynamic Temporal Warping [12]	54.0%
Distinctive Canonical Poses [13]	65.7%
Actionlet Ensemble (3D pose only) [12]	68.0%
Relative Position of Joints [14]	70.0%
Bio-inspired Predictive Orientation Decomposition [27]	79.7%
Our representation	**72.8%**

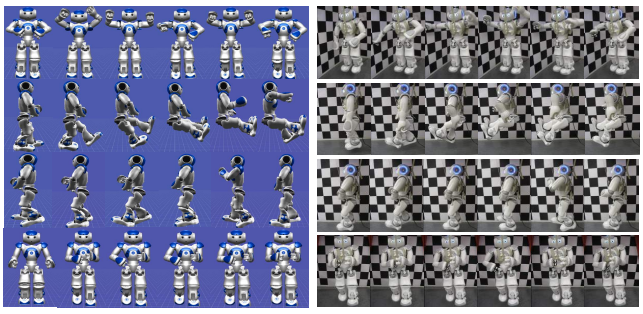
TABLE IV

COMPARISON OF RECOGNITION ACCURACY WITH PREVIOUS SKELETON-BASED REPRESENTATIONS ON HDM05.

Skeleton-based representations	Accuracy
Trifocal tensor of joint positions [28]	80.86%
Sequence of Most Informative Joints [4]	84.40%
Subtensor of joint positions [28]	85.71%
Relevant Joint Positions [29]	92.20%
Bio-inspired Predictive Orientation Decomposition [27]	96.70%
Our representation	**93.44%**

activity recognition (and not for robot imitation learning); our proposed approach is capable of learning action labels and motion trajectories simultaneously.

It is observed that activities involving lower body movements have a better recognition accuracy (e.g., *lie down, walk, stand up, sit down, hop both legs, lie down floor, squat*), while activities involving upper body movements (e.g., *read book, write on paper, use laptop*) are difficult to be distinguished, as Fig. 7(a) and Fig. 7(b) demonstrate. A possible explanation is that the upper body activities not only have similar motion trajectories, e.g., horizontal hand movements, but also a relatively small range of motion, which makes them difficult to be separated. On the other hand, the lower body activities have opposite characteristics in terms of motion trajectory and range of motion, which usually results in higher recognition accuracy.

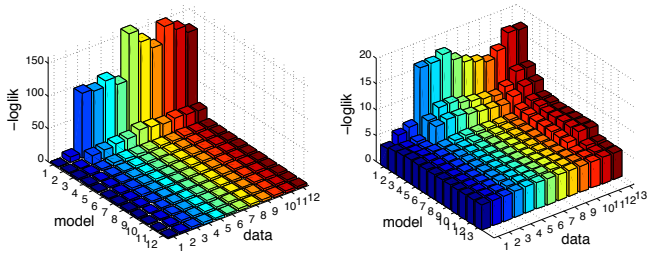


(a) Simulated robot (b) Real robot

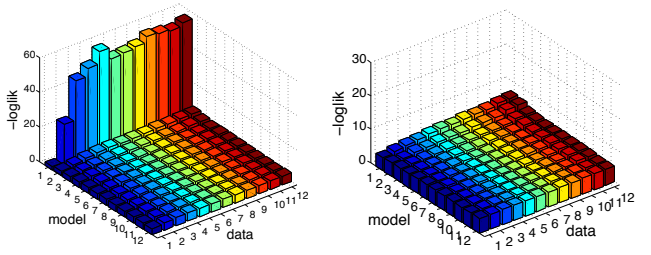
Fig. 8. Robot reproduces four activities (top to bottom: cheer up, lie to sofa, walking, writing) learned from MSR Daily Activity 3D dataset. An accompanying video is in `http://web.eecs.utk.edu/~czhang24/videos/uni.mp4`

C. Motion Trajectory Learning

To evaluate the effectiveness of learning motion trajectories, we use a 3D NAO robot simulator in Choreography software. The learned robot joint angle sequences are plugged into the simulator to enable the robot to perform all 27 activities. Fig. 8(a) shows two upper body activities and two lower body activities from the MSR Daily Activity



(a) Write on paper (b) Cheer up



(c) Lie down on sofa (d) Walk

Fig. 9. Evolution of the inverse log-likelihood averaged on all features, incrementally training with the MSR Daily Activity 3D dataset.

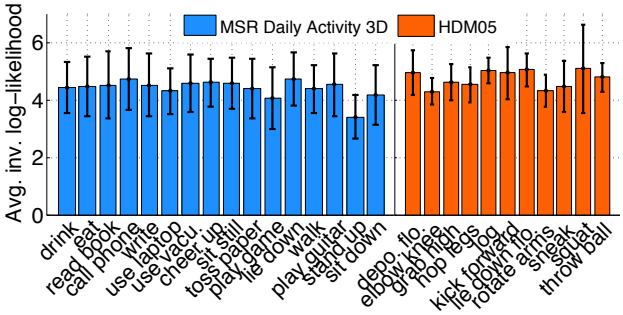


Fig. 10. Average inverse log-likelihoods.

3D dataset learned by our approach. For the *write on paper* activity containing trivial movements of the right or left hand, the robot can still reproduce the motion trajectory correctly. Experimental results with a real robot are shown in Fig. 8(b).

D. Evaluation

Besides evaluating our approach's classification accuracy on benchmark datasets and the effectiveness of reproducing motion trajectories on physical robots, we also study two important metrics that measure how well it performs.

1) Generalization: To quantitatively evaluate the motion trajectory learning performance of our united representation, we introduce the metric of generalization, which indicates the capability to generalize well on unseen data. We measure the likelihood of learned models as the indication of generalization level. For visualization purposes, inverse log-likelihood is used as shown in Fig. 9, where lower values indicate better performance.

Each activity in Fig. 9 is composed of 12 samples. The index of the model represents the corresponding size of training data, e.g., model m is trained by data composed by sample $\{1, 2, ..., m\}$. It is observed that the generalization performance can be improved by learning more training data, i.e., inverse log-likelihood value decreases as data size

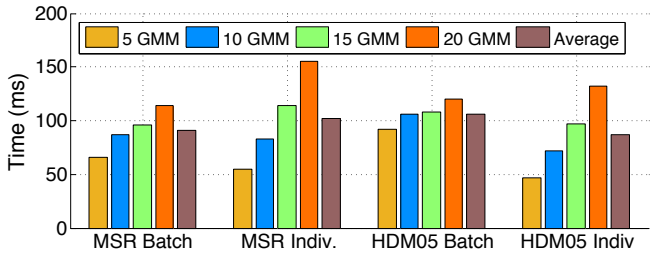
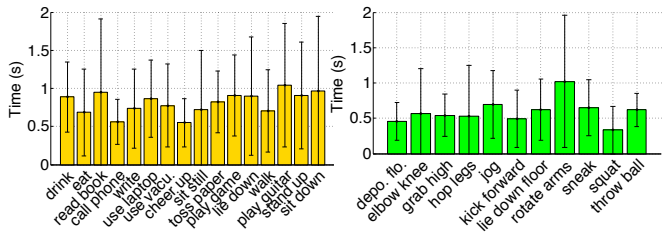


Fig. 11. Average running time for action label classification (milliseconds).



(a) MSR Daily Activity 3D (b) HDM05

Fig. 12. Average running time for incremental motion trajectory learning.

increases. When the data size is extremely limited, e.g., only sample 1 is used for training ($data = 1$), the generalization is poor (high inverse log-likelihood of $model = 1$ for unseen $data = 2, 3, ..., 12$). By incrementally learning more data, the inverse log-likelihood decreases and finally converges to a low value, which indicates that the data are well represented by our unified representation. It is also observed that the inverse log-likelihood decreases quickly. For all four activities in Fig. 9, $model = 3$ achieves a low inverse log-likelihood value not only for the training set $data = 1, 2, 3$, but also for the unseen $data = 4, 5, ..., 12$, which further validates that the proposed representation has a superior generalization capability in motion trajectory learning.

The overall generalization performance is presented in Fig. 10. Both datasets have a low inverse log-likelihood value. We see that the variations ofMSR Daily Activity 3D are smaller than HDM05, due to the fact that Kinect sensors are less accurate in estimating human skeletons.

2) Computational Cost: The real-time performance of our proposed method is evaluated by classifying action labels and regressing generalized trajectories on unseen data. Note that the representation building process is off-line, while the action label testing and incremental motion trajectory learning are online. Results in Fig. 11 indicate that for all model configurations, our proposed method can recognize an activity within 150 milliseconds, making it applicable to real-time onboard robotics applications. It is also observed that the time cost increases as the number of Gaussians increases, because a larger number of Gaussians leads to more complex representations, and accordingly a longer running time.

The real-time performance of the motion trajectory learning is presented in Fig. 12. For most activities, our method achieves less than 1 second running time. We see that *cheer up* in MSR Daily Activity 3D and *squat* in HDM05 have the best real-time performance. A plausible explanation is that when subjects are performing these activities, they follow more strict constraints, making instances of these two classes more consistent, and resulting in a lower convergence time

for the EM algorithm. The computation time of activities in both datasets has highly variable results as presented by the error bars (black lines). Indeed, the GMM encoding procedure depends on the random initialization of $k-$means, so that it may have a very different number of training iterations, resulting in the large variations of running time.

V. Discussion

This paper addresses the HAUL problem in robot imitation learning at the representational level and illustrates the learning results using data with highly varied human activity datasets. Our approach possesses several desirable characteristics. By constructing a shared model and extracting subtraction vectors between limbs as features, our representation is invariant to the variations of human body shape, scale, and position. By employing DTW and GMM modeling, our approach can generalize the spatial-temporal characteristics from the highly varied data, which is applicable to the large-scale datasets.

On the other hand, the learning performance is inherently limited by the modeling method we adopted in this paper. Because GMM explicitly encodes time in the model, it is less efficient in periodic activity learning. This limitation can be leveraged by using approaches like Hidden Markov Model or Dynamic Motion Primitives. The proposed approach learns motion trajectories of activities, but does not aim to accomplish the tasks by interacting with objects (e.g., learn the *lie on sofa* trajectory, without lying on a real sofa). In the future, we plan to introduce task constraints into our unified learning framework to enable the robot to interact with objects to accomplish tasks. Besides of the quantitative metric, i.e., log-likelihood, introducing human factors into the loop to qualitatively evaluate the robot's performance is another interesting direction we plan to explore, as in [21].

VI. Conclusion

We introduce a novel unified learning approach based on compact human representations to solve the critical HAUL problem, which allows a robot to simultaneously learn semantic meanings and motion trajectories of human actions. Our representation encodes the spatio-temporal information of human activities based on the compact parameter sets of GMMs. A SVM model is employed with our representation to recognize human actions, while the same representation is also used to reproduce the action motions based on GMR. To reduce human efforts on demonstration data collection, we propose to utilize the existing 3D human skeletal datasets that are publicly available, and relatively large-scale for robot imitation learning. The variations in the datasets are robustly addressed using a new incremental learning method. Experiments on human action recognition are performed using publicly available datasets. Recognition methods using our representation obtain an average accuracy of 72.8% and 93.44% over the MSR Daily Activity 3D and HDM05 datasets, respectively. The recognized activities are also imitated and reproduced using both simulated and real humanoid robots, through combining GMR with our representation. Empirical studies show our approach obtains promising generalization results and achieves real-time performance to reproduce the learned human activities.

References

[1] J. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recogn. Lett.*, vol. 48, no. 0, pp. 70–80, 2014.

[2] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, "Simplex-based 3d spatio-temporal feature description for action recognition," in *CVPR*, 2014.

[3] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, *Handbook of Robotics*, ch. 59: *Robot Programming by DemonstrationRobot Programming by Demonstration*. Springer, 2008.

[4] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014.

[5] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *IJCAI*, 2013.

[6] S. Calinon, F. Guenter, and A. Billard, "On learning, representing and generalizing a task in a humanoid robot," *IEEE Trans. Syst. Man, Cybern. B, Cybern*, vol. 37, no. 2, pp. 286–298, 2007.

[7] T. Inamura, H. Tanie, and Y. Nakamura, "Keyframe compression and decompression for time series data based on the continuous Hidden Markov Model," in *IROS*, 2003.

[8] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *WACV*, 2013.

[9] O. Brdiczka, J. M. M. Langet, and J. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Trans. Autom. Sci. Eng.*, vol. 6, no. 4, pp. 588–597, 2009.

[10] H. Zhang and L. Parker, "4-dimensional local spatio-temporal features for human activity recognition," in *IROS*, 2011.

[11] A. Veeraraghavan, A.Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *PAMI*, vol. 27, no. 12, pp. 1896–1909, 2005.

[12] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.

[13] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola, Jr., and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 420–436, 2013.

[14] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *CVPR Workshops*, 2013.

[15] J. Koenemann, F. Burget, and M. Bennewitz, "Real-time imitation of human whole-body motions by humanoids," in *ICRA*, 2014.

[16] J. Kober and J. Peters, "Imitation and reinforcement learning: Practical learning algorithms for motor primitives in robotics," *IEEE Robot. Autom. Mag.*, vol. 17, no. 2, pp. 55–62, 2010.

[17] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," tech. rep., Universität Bonn, 2007.

[18] M. Lopes, F. S. Melo, and L. Montesano, "Affordance-based imitation learning in robots," in *IROS*, 2007.

[19] N. Pollard and J. Hodgins, "Generalizing demonstrated manipulation tasks," in *WAFR*, 2004.

[20] A. J. Ijspeert, J. Nakanishi, and S. Schaal, "Movement imitation with nonlinear dynamical systems in humanoid robots," in *ICRA*, 2002.

[21] M. J. Gielniak, C. K. Liu, and A. L. Thomaz, "Generating human-like motion for robots," *IJRR*, vol. 32, no. 11, pp. 1275–1301, 2013.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc, 2006.

[23] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 239–472, 1978.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[25] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *HRI*, 2007.

[26] C. L. Nehaniv and K. Dautenhahn, *The Correspondence Problem*. No. 21, Cambridge, MA, USA: MIT Press, 2002.

[27] H. Zhang and L. E. Parker, "Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction," in *ICRA*, 2015.

[28] Q. Liu and X. Cao, "Action recognition using subtensor constraint," in *ECCV*, 2012.

[29] A. López-Mendez, J. Gall, J. R. Casas, and L. J. V. Gool, "Metric learning from poses for temporal clustering of human motion," in *BMVC*, 2012.