2015 IEEE International Conference on Robotics and Automation (ICRA)
Washington State Convention Center
Seattle, Washington, May 26-30, 2015

Bio-Inspired Predictive Orientation Decomposition of Skeleton Trajectories for Real-Time Human Activity Prediction

Hao Zhang[1] and Lynne E. Parker[2]

***Abstract*— Activity prediction is an essential task in practical human-centered robotics applications, such as security, assisted living, etc., which targets at inferring ongoing human activities based on incomplete observations. To address this challenging problem, we introduce a novel bio-inspired predictive orientation decomposition (BIPOD) approach to construct representations of people from 3D skeleton trajectories. Our approach is inspired by biological research in human anatomy. In order to capture spatio-temporal information of human motions, we spatially decompose 3D human skeleton trajectories and project them onto three anatomical planes (i.e., coronal, transverse and sagittal planes); then, we describe short-term time information of joint motions and encode high-order temporal dependencies. By estimating future skeleton trajectories that are not currently observed, we endow our BIPOD representation with the critical predictive capability. Empirical studies validate that our BIPOD approach obtains promising performance, in terms of accuracy and efficiency, using a physical TurtleBot2 robotic platform to recognize ongoing human activities. Experiments on benchmark datasets further demonstrate that our new BIPOD representation significantly outperforms previous approaches for real-time activity classification and prediction from 3D human skeleton trajectories.**

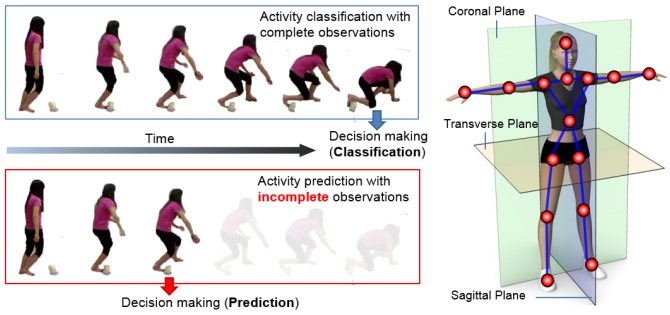


Fig. 1. A motivating example of human activity prediction: a robot needs to infer ongoing human activities and make a decision based on incomplete observations. We address this challenging prediction problem at the human representation level, through introducing a new skeleton-based, bio-inspired predictive orientation decomposition approach. Our human representation is constructed based upon biological research in human anatomy, which is able to (1) encode spatio-temporal information of 3D human joint trajectories, (2) estimate unobserved future data to make predictions of human activities, (3) deal with human rotations, body scale variations, and different formats of skeletal data obtained from a variety of 3D sensing devices, and (4) run in real time on physical robotic platforms.

I. Introduction

In human-centered robotics applications, including service robotics, assistive robotics, human-robot interaction, human-robot teaming, etc, automatically classifying and *predicting* human behaviors is essentially important to allow intelligent robots to effectively and efficiently assist and interact with people in human social environments. Although many activity recognition methods [1] have been proposed in robotics applications, most of them focus on classification of finished activities [2], [3], [4]. However, in a large number of practical human-centered robotics tasks, it is desirable for autonomous robotic systems to recognize human behaviors even before the entire motion is completed. For example, it is necessary for robot security guards to send off an alarm while someone is stealing rather than after the stealing, since early detection has significant potential to prevent the criminal activity and provide more time for police officers to react; it is desirable for an assistive robot to recognize falls as early as possible to reduce the incidence of delayed assistance after a fall, as illustrated by the example in Fig. 1.

The goal of activity prediction is to infer ongoing activities given temporally *incomplete information*. Predicting human activities is an extremely challenging problem in robot perception. First, a robot has to perform reasoning and decision making based on incomplete observations, which in general contain significant uncertainties and can change dramatically over time. Second, prediction of human activities must deal with conventional activity classification challenges, including significant variations of human appearance (e.g., body scale, clothes, etc.), complete or partial occlusion, etc. Third, action prediction with a mobile robot introduces additional, unique challenges to robot perception:

- A moving robotic platform typically results in frequent changes in viewing angles of humans (e.g., front, lateral or rear views).
- A moving robot leads to a dynamic background. In this situation, human representations based on local features [4] are no longer appropriate, since a significant amount of irrelevant features can be extracted from the dynamic background.
- Prediction performed under computational constraints by a robot introduces new temporal constraints, including the need to predict human behaviors and react to them as quickly and safely as possible [5].

To address the aforementioned challenges, we introduce a novel 3D human representation called *Bio-Inspired Predictive Orientation Decomposition* (BIPOD) of skeleton trajectories. Our BIPOD representation models the human body as

[1]Hao Zhang is with the Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA. hzhang@mines.edu
[2]Lynne E. Parker is with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA. leparker@utk.edu

an articulated system of rigid segments that are connected by joints in 3D (xyz) space. Then, human body motions can be modeled as a temporal evolution of spatial joint configurations in 3D space. Taking advantage of modern technologies of 3D visual perception (e.g., structured-light sensors, such as Kinect and PrimSense) and state-of-the-art skeleton estimation methods [6], we can reliably extract and track human skeletons in real time. Given the skeleton trajectory, our representation is able to encode spatio-temporal information of joint motions in an efficient and compact fashion that is highly descriptive for classification and prediction of ongoing human activities in real-world environments.

The main contribution of this work is the skeleton-based 3D representation of people, based on our novel bio-inspired predictive orientation decomposition, which includes several novelties: (1) We construct our representation based upon biological human anatomy research, which provides theoretical guarantees that our approach is able to effectively encode all human movements. (2) We introduce a novel spatio-temporal method to build human representations in 4D ($xyzt$) space, which spatially decomposes and projects 3D joint trajectories onto 2D anatomical planes and encodes temporal information of joint movements including high-order time dependencies. (3) We implement a simple, yet effective procedure to endow our human representation with critical predictive capabilities, which offers a satisfactory solution at the representation level to address the challenging activity prediction problem.

The rest of the paper is structured as follows. In Section II, we overview related work on 3D robotic vision and activity prediction. Section III discusses our new approach in detail. Results of empirical study are presented in Section IV. After discussing the characteristics of our approach in Section V, we conclude the paper in Section VI.

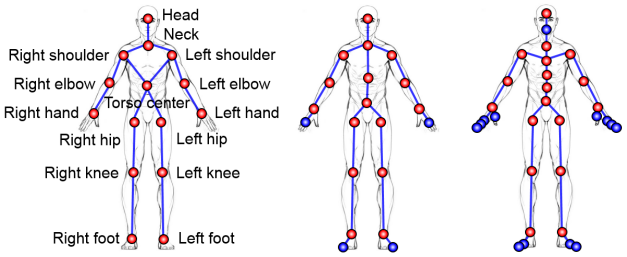
II. Related Work

We first overview perception systems that can be applied to acquire skeleton data in 3D space. Then, we review existing skeleton-based human representations applied for the activity recognition task. Finally, we discuss previous approaches for activity prediction.

A. Skeleton Acquisition from 3D Perception

The skeleton is a natural representation of the human body structure, which assumes that the human body is an articulated system of rigid segments that are connected by joints. Acquisition of 3D human skeleton sequences has been a desirable goal for a long time. An approach to obtain 3D human skeleton data is using a motion capture (MoCap) system, which typically uses multiple cameras to track reflective markers attached to the human body. For example, 3D skeleton data in the HDM05 Mocap dataset [7] contains 24 joints, as depicted in Fig. 2(c). Although a MoCap system provides very accurate and clean skeleton data, it cannot be used on mobile robotic platforms.

Recently, structured-light sensors or color-depth cameras have attracted significant attention, especially from robotics researchers. These sensors have become a standard device to construct 3D perception systems on intelligent mobile robots. Two sophisticated, off-the-shelf approaches are available to acquire 3D human skeletons from a structured-light sensor: (1) Microsoft provides a SDK for Kinect sensors, which can provide skeletal data with 20 joints [6], as illustrated in Fig. 2(b); and (2) OpenNI, which is adopted by Robot Operating System (ROS), estimate human body skeletons with 15 joints. These affordable structured-light sensors generally obtain satisfactory skeleton data, and can be easily installed on mobile robotic platforms [4].



(a) OpenNi (b) Kinect SDK (c) MoCap

Fig. 2. Examples of skeletal kinematic human body models obtained from different 3D perception technologies. Skeleton data acquired from OpenNI contains 15 joints as depicted in Fig. 2(a), 20 joints from Microsoft Kinect SDK as shown in Fig. 2(b), and a varied number of joints from a MoCap system such as 31 joints in Fig. 2(c). By only using the joints in red color, the proposed BIPOD representation is able to consistently process skeleton data obtained from different sensing techniques.

Our approach directly works on the skeletal data that are estimated using different technologies (i.e., OpenNI, Kinect SDK, and MoCap). In addition, a representation trained using one type of skeletal data can be directly applied to recognize human activities contained in other types of skeletal data.

B. Skeleton-Based Activity Classification

After the recent release of affordable structured-light sensors, we have witnessed a growth of studies using 3D skeletal data to interpret human behaviors. A 3D representation was introduced in [8] that is based on the joint rotation matrix with respect to body torso. Another representation based on skeletal joint positions was implemented in [9] to construct actionlet ensembles for activity recognition. A moving pose descriptor was introduced in [10], which uses joint positions in a set of key frames and encodes kinematic information as differential 3D quantities. By computing joint displacement vectors and joint movement volume, the representation in [11] is used to efficiently recognize activities from skeleton data. Other skeleton based 3D human representations were also implemented based on histograms of oriented displacements [12], covariance of 3D joints [13], etc.

Different from previous skeleton-based human representations, our BIPOD representation is bio-inspired with a clear interpretation in human anatomy research [14], [15]. Another significant difference is that our predictive representation is developed for activity prediction, instead of activity classification as in previous works.

C. Activity Prediction

Different from conventional action classification [4], [1], several approaches exist in the literature that focus on activity prediction, i.e., inferring ongoing activities before they are finished. An early approach applied dynamic programming to do early recognition of human gestures [16]. A max-margin early event detector was implemented in [17], which modifies structured output SVM to detect early events. Logistic regression models [18] were employed to detect starting point of human activities. An online Conditional Random Field method was introduced in [19] to predict human intents in human-robot collaboration applications. Prediction in the aforementioned methods is performed at the classifier level, through extending conventional machine learning methods to deal with time in an online fashion, in general. Significantly different from these techniques, we focus on developing an accurate, efficient fundamental representation of humans that can be directly used by learning approaches.

Only a few approaches were implemented at the representation level. For example, a dynamic Bag-of-Words (BoW) approach was introduced in [20] to enable activity prediction, which divides the entire BoW sequence into subsegments to find the structural similarity between them. To capture the spatio-temporal structure of local features, a spatial-temporal implicit shape model was implemented in [21] based on BoW models. Despite certain successes of the BoW representation for human behavior prediction, it suffers from critical limits. BoW-based representations cannot explicitly deal with view angle variations, and therefore typically cannot perform well on moving robotic platforms. In addition, computing BoW-based representations is computationally expensive, which is not applicable in real-time onboard robotics applications, in general. Moreover, the aforementioned BoW representations do not make use of depth information that is available from structured-light sensors. Different from previous studies on BoW representations, our work focuses on developing a new skeleton-based 3D human representation that is accurate and efficient to predict human activities and is able to deal with the aforementioned limitations.

III. Proposed Skeletal Representation

This section introduces our novel BIPOD representation. First, we discuss our bio-inspired representation's foundation in human anatomy. Then, we introduce our approaches to estimate anatomical planes and human facing direction and to decompose spatio-temporal joint orientations on anatomical planes. Finally, we discuss our approach's predicative ability to address activity prediction.

A. Foundation in Biology

In human anatomy, human motions are described in three dimensions according to a series of planes named anatomical planes [14], [22], [15]. There are three anatomical planes of motions that pass through the human body, as demonstrated in Fig. 3:

- *Sagittal plane* divides the body into right and left parts;

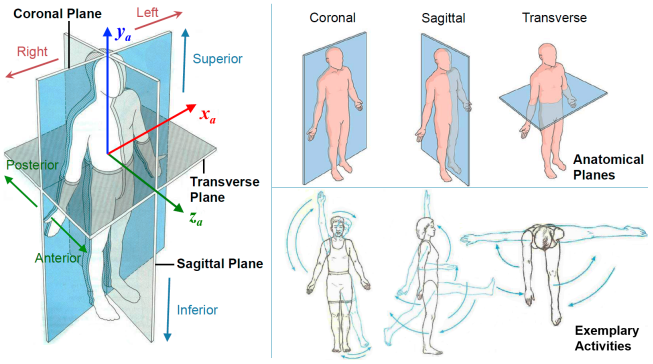


Fig. 3. Our bio-inspired representation is based on anatomical planes in human anatomy research. This figure demonstrates how anatomical planes divide the human body into different portions and illustrates exemplary human motions performed in each anatomical plane [22].

- *Coronal (frontal) plane* divides the human body into anterior and posterior portions;
- *Transverse (horizontal) plane* divides the human body into superior and inferior parts.

When describing a human movement in anatomy, there is a tendency to refer to it in a particular anatomical plane that dominates the movement. Examples of human movements in each anatomical plane are demonstrated in Fig. 3. When human movement occurs in several planes, this simultaneous motion can be seen as one movement with three planes, which is referred to as *tri-planar motion* [22]. In human anatomy research [14], [15], it has been theoretically proved and clinically validated that all human motions can be encoded by the tri-planar motion model.

The proposed BIPOD representation is inspired by the tri-planar movement model in human anatomy research: human skeletal trajectories are decomposed and projected onto three anatomical planes, and spatio-temporal orientations of joint trajectories are computed in anatomical planes. Based on the tri-planar motion model in anatomy research, it is guaranteed that our bio-inspired BIPOD representation is able to represent all human motions and thus activities. In addition, since we use the same standard terminology, it is biomechanically understood by biomedical researchers.

B. Estimation of Anatomical Planes

A core procedure of our bio-inspired human representation is to estimate anatomical planes, which involves three major steps: inferring the *coronal axis* z_a (intersection of the sagittal and transverse planes), *transverse axis* y_a (intersection of the coronal and sagittal planes), and *sagittal axis* x_a (intersection of the coronal and transverse planes). The anatomical axes x_a, y_a, z_a are illustrated in Fig. 3.

1) Estimating coronal axis z_a: Since the coronal plane is represented by human torso in anatomy [22], we can adopt joints of the human torso to estimate the coronal plane. Toward this goal, an efficient planar fitting approach based on least squares minimization is implemented to fit a plane to human torso joints in 3D space. Formally, given a set of M torso joints $\boldsymbol{P} = \{(x_i, y_i, z_i)\}_{i=1}^{M}$, the objective is

to estimate the parameters A, B and C, so that the plane $z = Ax + By + C$ can best fit the human torso joints in the sense that the sum of the squared errors $err(A, B, C)$ is minimized. Given the definition of squared errors:

$$err(A, B, C) = \sum_{i=1}^{M} \|(Ax_i + By_i + C) - z_i\|^2, \quad (1)$$

the parameters (A_c, B_c, C_c) of the human coronal plane are estimated using least squares minimization as follows:

$$(A_c, B_c, C_c) = \underset{A,B,C}{\operatorname{argmin}} \; err(A, B, C) \quad (2)$$

which can be solved by computing its derivative as:

$$\nabla err = 2\sum_{i=1}^{M} ((Ax_i + By_i + C) - z_i)\,(x_i, y_i, 1) = 0 \quad (3)$$

It is noteworthy that the estimated coronal plane's surface normal $(A, B, 1)$ lies along the z_a-axis as shown in Fig. 3.

After the coronal plane is estimated, we need to determine the coronal axis z_a, which is defined to point to the anterior direction (i.e., the same as human facing direction) in human anatomy [22], as shown in Fig. 3. Based upon this definition, we estimate the human facing direction in order to initialize the direction of the coronal axis z_a (performed only once). To this end, a detection window is placed around the joint representing human head (as demonstrated in Fig. 2(a)) in the color image. Then, a highly efficient, off-the-shelf human face detector, based on Haar cascades [23], is employed to detect whether a face exists in the detection window. If a positive is obtained, which means the human subject is facing to the sensor, then we define the coronal axis z_a is pointing to the sensor.

2) Estimating the sagittal axis x_a and transverse axis y_a: The origin of the estimated anatomy coordinate is place at the human torso center, as shown in Fig. 3. Then, the transverse axis y_a points from the torso center to the neck joint within the coronal plane, and the sagittal axis x_a is defined to point to the left side of the human body, which lies within the coronal plane and is perpendicular to y_a and z_a as illustrated in Fig. 3.

C. Anatomy-Based Orientation Decomposition

To construct a discriminative and compact representation, our novel bio-inspired approach decomposes 3D trajectories of each joint of interest, and describes them separately within the 2D anatomical planes in a spatio-temporal fashion.

1) Anatomy-based spatial decomposition: Given the estimated human anatomical coordinate $x_a y_a z_a$, the trajectory of each joint of interest in 3D space is spatially decomposed into three 2D joint trajectories, through projecting the original 3D trajectory onto anatomical planes. Formally, for each joint of interest $\boldsymbol{p} = (x, y, z)$, its 3D trajectory $\boldsymbol{P} = \{\boldsymbol{p}_t\}_{t=1}^{T}$ can be spatially decomposed as

$$\boldsymbol{P} = \{\boldsymbol{p}_t^{(x_a y_a)},\ \boldsymbol{p}_t^{(y_a z_a)},\ \boldsymbol{p}_t^{(z_a x_a)}\}_{t=1}^{T} \quad (4)$$

where $(x_a y_a)$ denotes the coronal plane, $(y_a z_a)$ denotes the sagittal plane, $(z_a x_a)$ denotes the transverse plane, and $\boldsymbol{p}_t^{(\cdot)}$ represents the 2D location of the joint $\boldsymbol{p}$ on the $(\cdot)$ anatomical plane at time t. Due to this bio-inspired spatial decomposition, our novel 3D human representation is invariant to view point variations and global human movements, as proved in the human anatomy research [22].

2) Temporal orientation description: After each 3D joint trajectory is decomposed and projected onto 2D anatomical planes, we represent the 2D trajectories on each plane using a histogram of the angles between temporally adjacent motion vectors. Specifically, given the decomposed 2D human joint trajectory $\boldsymbol{P}^{(\cdot)} = \{\boldsymbol{p}_t^{(\cdot)}\}_{t=1}^{T}$ on an anatomical plane, i.e., the coronal $(x_a y_a)$, transverse $(z_a x_a)$, or sagittal $(y_a z_a)$ plane, our approach computes the following angles:

$$\theta_t = \arccos \frac{\overrightarrow{\boldsymbol{p}_{t-1}\boldsymbol{p}_t} \cdot \overrightarrow{\boldsymbol{p}_t\boldsymbol{p}_{t+1}}}{\|\overrightarrow{\boldsymbol{p}_{t-1}\boldsymbol{p}_t}\|\|\overrightarrow{\boldsymbol{p}_t\boldsymbol{p}_{t+1}}\|}, \quad t = 2, \ldots, T-1 \quad (5)$$

where $\theta \in (-180°, 180°]$. Then, a histogram of the angles is computed to encode statistical characteristics of the temporal motions of the joint on the anatomical plane. Intuitively, the histogram represents how many degrees a body joint changes its direction at each time point. Because the direction change of a joint is independent of its moving distance, the proposed representation, based on orientation changes, is invariant to variations of human body scales.

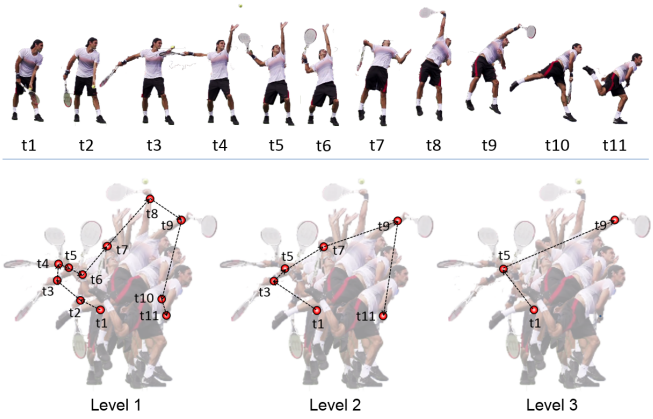


Fig. 4. An example of the temporal pyramid applied in our approach to capture long-term independencies. In this example, a temporal sequence of eleven frames is used to represent a tennis-serve activity, and the joint we are interested in is the right wrist, as denoted by the red dots. When three levels are used in the temporal pyramid, level 1 uses human skeleton data at all time points $(t_1, t_2, \ldots, t_{11})$; level 2 selects the joint positions at odd time points $(t_1, t_3, \ldots, t_{11})$; and level 3 continues this selection process and keeps half of the temporal data points (t_1, t_5, t_9) to compute long-term orientation changes.

It is noted that the oriented angles computed based on Eq. (5) can only capture temporal information within a short time interval. In order to encode long-term temporal relationships, a temporal pyramid framework is applied, which temporally decomposes the entire trajectory into different levels. In level 1, the entire trajectory of a joint of interest is used to compute the orientation changes on each anatomical plane, which is exactly the same as Eq. (5). In level 2 of the pyramid, only half of the temporal joint positions are adopted, for example, $t = 1, \ldots, 2n-1$ where $n \in \mathbb{R}$. If a temporal pyramid has

three levels, then in level 3, only the joint data that satisfy $t = 1, \ldots, 4n-1$ where $n \in \mathbb{R}$ are applied to compute the orientation changes. Fig. 4 illustrates an intuitive example of using a 3-level temporal pyramid to capture long-term time dependencies in a tennis-serve activity. Temporal orientation changes that are calculated in different levels of the pyramid are accumulated in the same histogram.

To construct a final representation based on the orientation changes, three histograms computed from the 2D trajectories on the coronal, transverse and sagittal anatomical planes are concatenated into a single feature vector. Through capturing both space (anatomy-based spatial decomposition) and time (temporal orientation description) information, our novel bio-inspired approach provides a spatio-temporal representation of humans and their movements.

D. Joint Trajectory Refinement and Prediction

Because skeletal data acquired from 3D robot perception systems can be noisy, it is important to estimate true positions of human joints given the observed skeleton data. In addition, in order to solve the activity prediction task, our representation requires the capability of predicting future human joint positions. To solve these problems, Extended Kalman Filters (EKFs) [24] are used, which are a non-linear extension of Kalman filters. Estimating and predicting body joint positions using observable skeleton data is essentially a non-linear tracking problem that can be solved by EKFs, in which the true joint position is the state and the position from acquired skeleton data is the observation.

To reduce the computational cost of large state space (i.e., all body joints), we divide the state space into five subspaces: left-arm space (left elbow and hand, 2 states), right-arm space (2 states), left-leg space (left knee and foot, 2 states), right-leg space (2 states), and torso space (number of states may vary when different types of skeleton data are used, as shown in Fig. 2). When redundant joints are provided (such as the skeletal data from MoCap systems), our approach only uses the aforementioned joints (as illustrated by the red-colored joints in Fig. 2), which guarantees the direct applicability of our representation on skeletal data obtained from different technologies such as using OpenNI or MS Kinect SDK.

Our simple yet effective solution of applying EKFs to track true human joint positions provides two advantages. First, the procedure endows our bio-inspired representation approach with the capability of encoding human motions in the near future, which is essential to human activity prediction using incomplete observations. This is achieved by using past and current states to predict future states in an iterative fashion. Second, besides filtering out the noise in observed skeleton data, this procedure makes our representation available all the time to a robotic system, even during time intervals between frames when skeletal data are acquired. In this situation, by treating the non-existing observation (between frames) as a missing value, the estimated state can be applied to substitute the observation at that time point.

IV. Experiments

To evaluate the performance of our BIPOD representation on human activity classification and prediction, we perform comprehensive experiments using publicly available benchmark datasets. Also, to demonstrate the impact of our BIPOD representation in real-world robotics applications, we test our approach on a TurtleBot2 robot to perform real-time online activity recognition.

A. Implementation

Our skeleton-based BIPOD representation is implemented using a mixture of Matlab and C++ programming languages on a Linux machine with an i7 3.0G CPU and 16Gb memory. Each of the three histograms, computed from trajectories on the coronal, transverse and sagittal planes, contains 12 bins. The learner employed in this paper is the non-linear Support Vector Machine (SVM) [25] with χ^2-kernels [26], which has demonstrated superior performance on the histogram-based input (e.g., our BIPOD representation). In order to address multi-class classification and prediction, the standard one-against-one methodology is applied [25].

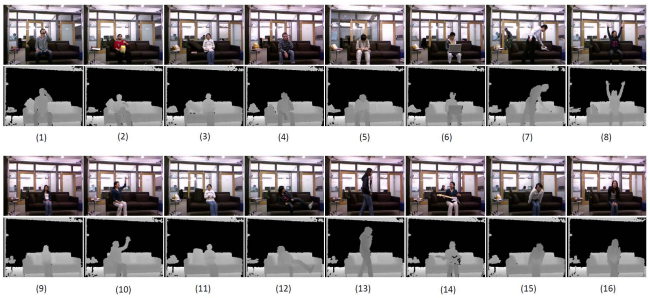


Fig. 5. The MSR Daily Activity 3D dataset is applied in the experiment to evaluate our BIPOD representation, which contains 16 activity categories: (1) drink, (2) eat, (3) read book, (4) call cellphone, (5) write on a paper, (6) use laptop, (7) use vacuum cleaner, (8) cheer up, (9) sit still, (10) toss paper, (11) play game, (12) lie down on sofa, (13) walk, (14) play guitar, (15) stand up, (16) sit down.

B. Evaluation on MSR Activity Daily 3D dataset

The MSR Daily Activity 3D dataset [9][1] is a most widely used benchmark dataset in human activity recognition tasks. This dataset contains color-depth and skeleton information of 16 activity categories, as illustrated in Fig. 5. Each activity is performed by 10 subjects twice, once in a standing position and once in a sitting position in typical office environments, which results in a number of 320 data instances. The skeleton data in each frame contains 20 joints, as shown in Fig. 2(b). In our experiments, we follow the experimental setups used in [27]; accuracy is applied as the performance metric.

We investigate our BIPOD representation's performance in the activity recognition task, i.e., classifying human activities using complete observations. Experimental results obtained by our approach over the MSR Daily Activity 3D dataset

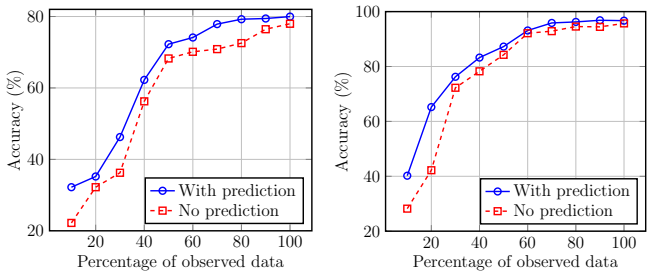
[1]MSR Daily Activity 3D dataset: `http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc`.

TABLE I

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

Skeleton-based representations	Accuracy
Dynamic Temporal Warping [9]	54.0%
Distinctive Canonical Poses [18]	65.7%
Actionlet Ensemble (3D pose only) [9]	68.0%
Relative Position of Joints [28]	70.0%
Moving Pose [10]	73.8%
Fourier Temporal Pyramid [9]	78.0%
Our BIPOD representation	**79.7%**

are presented in Table I. When a human activity is complete and all frames are observed, our approach obtains an average recognition accuracy of 79.7%. In order to show the proposed representation's superior performance, we also compare our approach with state-of-the-art skeleton-based representations in human activity recognition tasks, as presented in Table I. It is observed that our approach outperforms previous works and obtains the best recognition accuracy over this dataset. In addition, we evaluate the efficiency of our approach in the activity classification task. An average processing speed of 53.3 frames-per-second is obtained, which demonstrates the high efficiency of our representation. Because this processing speed is faster than the frame rate of structured-light cameras, real-time performance can be achieved on a robotic platform equipped with such 3D sensors.



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

Fig. 6. Experimental results of using our BIPOD representation to predict human activities given incomplete observations. When the procedure of joint trajectory refinement and prediction is used, 15% future data are predicted. Generally, the predictive representation greatly outperforms representations without prediction capabilities.

To demonstrate that our BIPOD representation is capable of predicting ongoing activities based on incomplete observations, we conduct a series of experiments by feeding different percentages of observations to our method. Then, 15% future unobserved data are predicted by the component procedure of joint trajectory refinement and prediction, as discussed in Section III-D. After combining the predicted data with the observed trajectories of joints, the robot can make a decision to respond to the ongoing activity before it is complete. The quantitative experimental results on the MSR Daily Activity 3D dataset are illustrated in Fig. 6(a). It can be observed that, comparing with the representation without feature prediction, the BIPOD version obtains much better recognition accuracy. This highlights the fact that the predicted data do contribute to improving recognition accuracy, which also demonstrates the importance of endowing human representations with the critical prediction capability.

C. Evaluation on HDM05 MoCap Dataset

To validate the generalizability and applicability of our BIPOD representation on skeleton data collected from different sensing technologies, we conduct another set of experiments using skeletal data obtained using motion capture systems. The HDM05 MoCap dataset [7][2] is used in our experiments. Comparing with skeleton datasets collected using structured-light sensors, this MoCap dataset has several unique characteristics. First, the skeleton data are much less noisy than the data acquired by a color-depth sensor. Second, the human skeleton obtained by a MoCap system contains 31 joints, as shown in Fig. 2(c). Third, the frame rate of a MoCap system is 120 FPS, which is much higher than maximum 30 FPS of a structured-light sensor. Fourth, only skeleton trajectories are provided by the HDM05 dataset, which does not provide color images. In this case, face recognition is not performed. Since all motion sequences begin with a T-pose, as explained in [7], we simply assume subjects face toward the view point.

The experimental setup applied in [12], [29] is adopted in our empirical study: Eleven categories of activities are used, which are performed by five human subjects, resulting in a total number of 249 data instances. Skeleton data from three subjects are used for training, and two subjects for testing. The activities used in our experiment include: deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sneak, squat, and throw basketball.

Table II presents the experimental results obtained using our BIPOD representation over the HDM05 MoCap dataset. The proposed method obtains an average accuracy of 96.70% in the human activity classification task using fully observed skeleton sequences. In addition, we compare our bio-inspired method with state-of-the-art skeleton-based human representations over the same dataset, which is reported in Table II. A similar phenomenon is observed that our BIPOD representation obtains a superior human activity recognition accuracy and outperforms existing skeleton-based representations. In terms of computational efficiency, a processing speed of 48.6 FPS is obtained, which is a little slower than processing the skeleton data from structured-light sensors, since more torso joints are used.

Additional experiments are also conducted to evaluate our BIPOD representation's ability to predict ongoing activities, based on incomplete observations from the HDM05 MoCap dataset. In this experiment, 15% future data is predicted by the process of joint trajectory refinement and prediction. Fig. 6(b) shows the experimental results obtained by our BIPOD representation. Comparison with the non-predictive version is also illustrated in the figure, which shows that the activity recognition accuracy can be significantly improved if human representations are predictive.

[2]HDM05 motion capture dataset: `http://resources.mpi-inf.mpg.de/HDM05`.

TABLE II

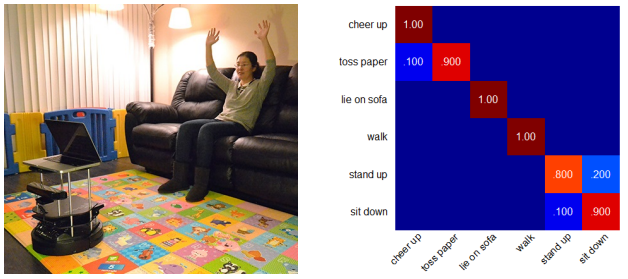
COMPARISON OF AVERAGE RECOGNITION ACCURACY WITH PREVIOUS SKELETON-BASED REPRESENTATIONS ON HDM05 MOCAP

Skeleton-based representations	Accuracy
Trifocal tensor of joint positions [30]	80.86%
Sequence of Most Informative Joints [29]	84.40%
Subtensor of joint positions [30]	85.71%
Relevant Joint Positions [31]	92.20%
Cov3DJ [13]	95.41%
Our BIPOD representation	**96.70%**

D. Case Study on TurtleBot2 Using Cross-Training

Besides evaluating our representation's performance using benchmark datasets, we also implement our BIPOD approach on physical robots to evaluate how well it performs in real-world robotics applications. The robotic platform used in our experiments is the TurtleBot2 robot built upon the Kobuki mobile base. The robot employs a Kinect sensor for on-board 3D perception and an ASUS netbook (with 1.6G dual core CPU and 2G memory) for on-board control. To compute our representation and perform activity recognition in real time, another laptop, as described in Section IV-A, is placed on top of the robot. The hardware configuration of our TurtleBot2 robot is illustrated in Fig. 7(a).

In order to validate our approach's generalization ability to process data from different skeleton estimation techniques, we apply the cross-training methodology in this experiment. Specifically, our BIPOD representation and the SVM classifier are trained using skeleton data obtained from MS Kinect SDK, which provides information of 20 joints, as shown in Fig. 2(b). Then, the learned models (i.e., human representation plus classifier) are directly applied to recognize activities from skeleton data that are obtained using OpenNI in ROS, which provide 15 body joints as depicted in Fig. 2(a). The essential advantage of cross-training is that, through applying similar datasets that are available on the internet to train a learning system, it is able to avoid collecting a new dataset and therefore can significantly save human labor.



(a) Experiment setups (b) Confusion matrix

Fig. 7. Our BIPOD representation is evaluated using a TurtleBot2 robotic platform to recognize ongoing activities in an online fashion in a standard living room environment, as illustrated in Fig. 7(a). The confusion matrix obtained in this experiment is presented in Fig. 7(b).

In this experiment, the MSR Daily Activity 3D dataset, as discussed in Section IV-B, is used to compute our representation and estimate the SVM's parameter. Six activity classes are adopted, including cheer up, toss paper, lie on sofa, walk, stand up, and sit down. Cross-training is performed using a five-fold cross-validation over all instances of each activity. Then, the learned reasoning system is directly applied by the robot to recognize the six activities in an online fashion. To deal with the online streaming skeleton data, a temporal sliding window technique is applied, where the window size is 2 seconds and the overlap of temporally adjacent windows is 1 second. Then, activity recognition is performed using the skeleton data falling in each window.

Two human subjects, a male and a female with different body scales and motion patterns, are involved in the online testing process to evaluate the performance of our BIPOD representation. Each subject performs each activity five times in a random order in a standard living room environment, as illustrated in Fig. 7(a). Ground truth is manually recorded and used to compare with recognition results from the TurtleBot2 robot for quantitative evaluation.

Fig. 7(b) shows the confusion matrix produced by the online activity recognition system based on our novel skeleton-based BIPOD human representation, where each column corresponds to the predicted category and each row corresponds to the ground truth category. It is observed that, when cross-training is used, our algorithm is able to accurately recognize continuous human activities from streaming skeleton data. This observation validates our method's capability of encoding skeleton data from different resources, which is achieved by only using discriminative joints and removing redundant joints on human limbs, as demonstrated in Fig 2. In addition, it is observed that our representation is able to encode time information, which is indicated by the successful separation between "stand up" and "sit down" activities. Since the used SVM classifier is not capable of modeling time, we can infer that the separation between these reversal activities results from our spatio-temporal representation. Finally, we observe that a small portion of "toss paper" activities are misclassified as "cheer up", since these activities share similar arm-moving motions.

In summary, our representation obtains an average online recognition accuracy of 93.33%, with a processing speed of 30 FPS (which is the maximum frame rate of the onboard Kinect sensor on TurtleBot2). The experimental results show that our skeleton-based bio-inspired algorithm is a promising human representation that is able to accurately and efficiently address online activity recognition.

V. DISCUSSION

Our skeleton-based representation based upon bio-inspired predictive orientation decomposition possesses several desirable characteristics. Our BIPOD human representation is a bio-inspired approach, which has clear biological interpretation in human anatomy. Through spatially decomposing joint trajectories and projecting them onto anatomical planes, our human representation is invariant to view point changes. By computing the temporal orientation, instead of using the joint moving distance, our representation is invariant to variations of human body scales. Through selecting the discriminative human joints that are available from all skeleton estimation

techniques, our BIPOD representation can be directly applied on different categories of skeleton data, which makes cross-training possible.

On the other hand, similar to other skeleton-based human representations, our approach cannot encode object information, and may not be able to effectively distinguish activities involving human-object interactions. In addition, the same as all skeleton-based methods, our representation heavily relies on the accuracy of global human skeleton estimation, which may suffer from severe occlusions. These limitations can be leveraged by combining 3D human skeleton data with color depth information.

VI. Conclusion

In this paper, we introduce the novel BIPOD representation to enable intelligent robots to predict human activities from skeletal data in real-world human-centered robotics applications. Our approach is inspired by biological human anatomy research, which provides theoretical and clinical guarantees that our representation can encode all human movements. To construct our the BIPOD representation, we estimate human anatomical planes, decompose 3D skeleton trajectories, and project them onto the anatomical planes. Then, we describe time information through computing motion orientations on each plane and encoding high-order time dependency using temporal pyramids. In addition, to endow our representation with the predictive capability, we use the simple yet effective EKF technique to estimate future skeleton trajectories, which can also reduce noise and deal with missing observations. We perform empirical studies, using a TurtleBot2 mobile robot, to evaluate the performance of our BIPOD representation in an ongoing human activity recognition task. In addition, our BIPOD representation is compared with methods in previous studies on activity classification and prediction, using MSR Daily Activity 3D and HDM05 MoCap benchmark datasets. Experimental results demonstrate that BIPOD significantly improves human activity recognition accuracy and efficiency and successfully addresses the challenging activity prediction problem.

References

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

[2] G. Chen, M. Giuliani, D. Clarke, A. Gaschler, and A. Knoll, "Action recognition using ensemble weighted multi-instance learning," in *IEEE International Conference on Robotics and Automation*, 2014.

[3] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellstrom, "Audio-visual classification and detection of human manipulation actions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014.

[4] H. Zhang and L. Parker, "4-dimensional local spatio-temporal features for human activity recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.

[5] H. Zhang, C. M. Reardon, and L. E. Parker, "Real-time multiple human perception with color-depth cameras on a mobile robot," *IEEE Trans. Cybernetics*, vol. 43, no. 5, pp. 1429–1441, 2013.

[6] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

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

[8] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *IEEE International Conference on Robotics and Automation*, 2012.

[9] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[10] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *IEEE International Conference on Computer Vision*, 2013.

[11] H. Rahmani, A. Mahmood, A. Mian, and D. Huynh, "Real time action recognition using histograms of depth gradients and random decision forests," in *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[12] 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 *International Joint Conference on Artificial Intelligence*, 2013.

[13] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *International Joint Conference on Artificial Intelligence*, 2013.

[14] H. Gray, *Anatomy of the Human Body*. Lea & Febiger, 1973.

[15] C. Yokochi and J. W. Rohen, *Color Atlas of Anatomy: A Photographic Study of the Human Body*. Lippincott Williams & Wilkins, 2006.

[16] A. Mori, S. Uchida, R. Kurazume, R.-I. Taniguchi, T. Hasegawa, and H. Sakoe, "Early recognition and prediction of gestures," in *International Conference on Pattern Recognition*, 2006.

[17] M. Hoai and F. De la Torre, "Max-margin early event detectors," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 191–202, 2014.

[18] 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," *International Journal of Computer Vision*, vol. 101, pp. 420–436, Feb. 2013.

[19] J. Hoare and L. Parker, "Using on-line conditional random fields to determine human intent for peer-to-peer human robot teaming," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[20] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *International Conference on Computer Vision*, pp. 1036–1043, 2011.

[21] G. Yu, J. Yuan, and Z. Liu, "Predicting human activities using spatio-temporal structure of interest points," in *ACM International Conference on Multimedia*, 2012.

[22] M. McGinnis, *Bioregionalism: The Tug and Pull of Place*. Routledge, 1999.

[23] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[24] G. Einicke and L. White, "Robust extended Kalman filtering," *IEEE Trans. Signal Processing*, vol. 47, pp. 2596–2599, Sept. 1999.

[25] 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.

[26] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 480–492, 2012.

[27] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[28] 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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[29] 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, pp. 24–38, Jan. 2014.

[30] Q. Liu and X. Cao, "Action recognition using subtensor constraint," in *European Conference on Computer Vision*, 2012.

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