Autonomous Robots
https://doi.org/10.1007/s10514-017-9692-3

Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction

Brian Reily[1] · Fei Han[1] · Lynne E. Parker[2] · Hao Zhang[1]

Received: 10 May 2016 / Accepted: 16 December 2017

Abstract
Activity prediction is an essential task in practical human-centered robotics applications, such as security, assisted living, etc., which is targeted 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. BIPOD is invariant to scales and viewpoints, runs in real-time on basic computer systems, and is able to recognize and predict activities in an online fashion. Our approach is inspired by biological research in human anatomy. 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 using Extended Kalman Filters to estimate future skeleton trajectories, we endow our BIPOD representation with the critical capabilities to reduce noisy skeleton observation data and predict the ongoing activities. Experiments on benchmark datasets have shown that our BIPOD representation significantly outperforms previous methods for real-time human activity classification and prediction from 3D skeleton trajectories. Empirical studies using TurtleBot2 and Baxter humanoid robots have also validated that our BIPOD method obtains promising performance, in terms of both accuracy and efficiency, making BIPOD a fast, simple, yet powerful representation for low-latency online activity prediction in human–robot interaction applications.

Keywords Human representation · Activity classification · Activity prediction · Real-time human–robot interaction

1 Introduction

In many human-centered robotics scenarios, including service robotics, assistive robotics, human–robot interaction, human–robot teaming, etc, automatically classifying and *predicting* human behaviors is critical to allow intelligent robots to effectively and efficiently assist and interact with people in human social environments. Although many activity recognition methods Aggarwal and Xia (2014) have been proposed in robotics applications, most of them focus on classification of finished activities (Chen et al. 2014; Pieropan et al. 2014; Zhang and Parker 2011). 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 robotic security guards to send off an alarm while someone is stealing rather than after the stealing, because early detection has significant potential to prevent the criminal activity and provide more time for police officers to react; it is helpful for semi-automated vehicles to analyze the activities of their drivers and predict ahead of time their intentions in order to determine whether to apply safety features or not; it is also desirable for an assistive robot to recognize falls as early as possible in order to reduce the incidence of delayed assistance after a fall, as illustrated by

Brian Reily and Fei Han have contributed equally to this work.

✉ Brian Reily
breily@mines.edu

Fei Han
fhan@mines.edu

Lynne E. Parker
leparker@utk.edu

Hao Zhang
hzhang@mines.edu

[1] Human-Centered Robotics Laboratory, Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA

[2] Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA

Published online: 27 December 2017

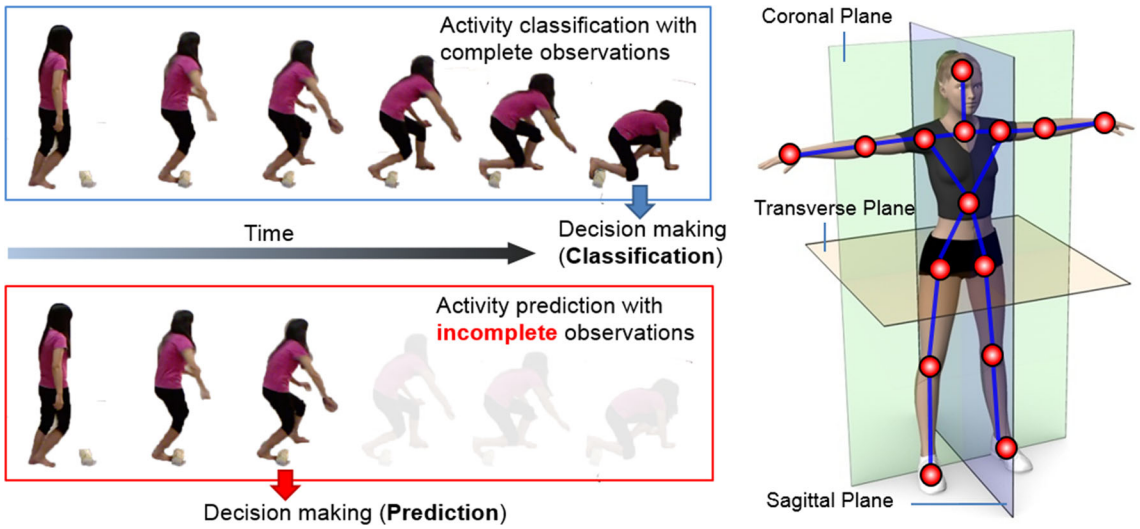


Fig. 1 A motivating example of human activity prediction: a robot needs to infer ongoing human activities and make timely decisions based on incomplete observations. We address this challenging prediction problem at the human representation level, through introducing a new skeleton-based, bio-inspired 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 and reduce noise in observed data to make accurate 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 various robotic platforms

the example in Fig. 1. It is also simpler and more efficient to interact with a robot when that robotic system can understand a human's actions and commands, even if those commands are incomplete or while the human is still expressing an intention.

The goal of activity prediction is early recognition of ongoing activities based on temporally *incomplete information*, or inferring human intention by recognizing subtasks of an intended task (Nikolaidis et al. 2017). Predicting human activities is a 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 over time. Second, prediction of human activities needs to deal with conventional activity classification difficulties, including human appearance variations (e.g., body scale, orientation, and clothing), complete or partial occlusion, etc. Third, action prediction with robotic platforms 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 (Zhang and Parker 2011) 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, such as the requirement to predict human behaviors and react to them as quickly and safely as possible (Zhang et al. 2013).

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 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 PrimeSense) and state-of-the-art skeleton estimation methods (Shotton et al. 2011), 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 applications.

The main contribution of this paper is the novel skeleton-based 3D representation of people based on the bio-inspired predictive orientation decomposition,[1] 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 effec-

[1] The code and data are publicly available at: http://hcr.mines.edu/code/bipod.html.

tively encode all human movements. (2) We introduce a novel spatio-temporal method to create 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 is able to reduce the impact of noisy observation data and offers a promising solution at the representation level to address the challenging real-time activity prediction problem. These three contributions constitute an approach that is real-time, independent of skeleton acquisition source, and able to recognize an activity before it finishes occuring, making it ideal for human–robot interaction in a wide variety of scenarios.

The rest of this paper is structured as follows. In Sect. 2, we overview related techniques on 3D robotic vision, skeleton-based representations, and human activity prediction. Section 3 discusses the proposed BIPOD method in detail. Results of empirical studies are presented in Sect. 4. After discussing the properties of our approach in Sect. 5, we conclude the paper in Sect. 6.

2 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.

2.1 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 (Müller et al. 2007) contains 24 joints, as depicted in Fig. 2c. Although a MoCap system provides very accurate and clean skeleton data, the infrastructure required makes it a better fit for applications such as immersive virtual reality software or building accurate digital character models, but not applicable on mobile robotic platforms in unstructured environments.

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 3D skeletal data with 20 joints (Shotton et al. 2011), as illustrated in Fig. 2b; and (2) OpenNI, which is adopted by the 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 (Zhang and Parker 2011).

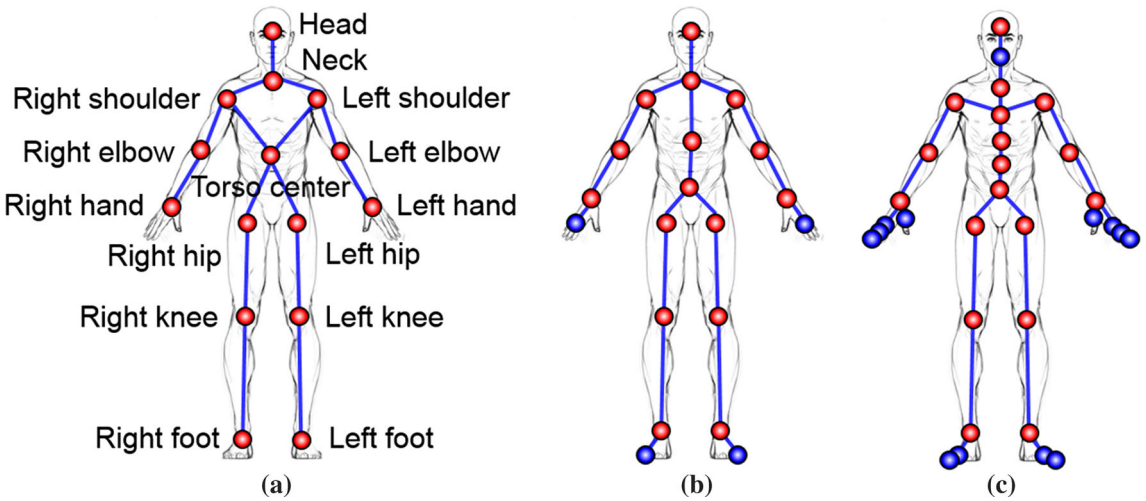


Fig. 2 Examples of skeletal kinematic human body models obtained from different 3D perception technologies. Skeleton data acquired from OpenNI contains 15 joints as depicted in Fig. 2a, 20 joints from Microsoft Kinect SDK as shown in Fig. 2b, and a varied number of joints from a MoCap system such as 31 joints in Fig. 2c. By only using the joints in red color, the proposed BIPOD representation is able to consistently process skeleton data obtained from different sensing techniques. This ability makes BIPOD available for a variety of different use cases, from human–robot interaction using simple RGB-D sensors to biomechanical motion analysis using larger scale MoCap systems. **a** OpenNi, **b** Kinect SDK and **c** MoCap

Skeletal acquisition from sensors such as the Kinect is an evolving research area, with approaches primarily based either on recognizing body parts or by matching 3D data to known skeletons. Body part recognition is the method utilized by Microsoft in the Kinect SDK (Shotton et al. 2011; Girshick et al. 2011), which analyzes single depth images and classifies individual pixels. Many other approaches have also been presented that operate on single depth images, such as Jung et al. (2015), Sung et al. (2012) and Charles and Everingham (2011), that utilize a variety of image processing methods to recognize body parts. Some methods rely on streams of depth images to take advantage of temporal information—Plagemann et al. (2010) uses Haar features to recognize body parts but builds in a Bayesian prior (Ganapathi et al. 2010) to improve accuracy. Schwarz et al. (2012) identifies joints based on geodesic distance, and uses optical flow between frames to resolve ambiguities caused by occlusions.

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.

2.2 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 [(summarized in reviews such as Han et al. (2017), Aggarwal and Xia (2014) and Han et al. (2016)]. The recognition of human activities is critical to robotics applications, as it is a necessary step for human–robot interactions. A 3D representation was introduced in Sung et al. (2012) that is based on the joint rotation matrix with respect to body torso. Another representation based on skeletal joint positions was implemented in Wang et al. (2012), 2014a to construct actionlet ensembles for activity recognition. A moving pose descriptor was introduced in Zanfir et al. (2013), 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 Rahmani et al. (2014) is used to efficiently recognize activities from skeleton data. Their work constructs a histogram of joint position differences, based on the distances in 3D space from each body joint to a reference joint, typically the torso/hip point. Similarly, Gowayyed et al. (2013) introduces histograms of oriented displacements. This representation is based on joint trajectories over time-angles are calculated from each joint to it's position in the previous time step, and a histogram is build from these angles to create a fixed length descriptor. Boubou and Suzuki (2015) presents an approach also based on orientation of joint movements, but include additional information present in the velocity vector of each movement. Many other skeleton based 3D human representations have also been developed. Some are also inspired by biology; Chaudhry et al. (2013) models humans as shapes obtained by tracking the neural processing of primates. Others are based on positions of joints For example, Hussein et al. (2013) represents a skeleton as a covariance of 3D joints, and Vantigodi and Babu (2013) is based on variance between joints. Some representations have been created based on the temporal relation of motion. Zhao et al. (2013) created a concept called motion templates, and Koppula et al. (2013) is based on relation of joints over time. Luo et al. (2013) uses sparse coding to represent human motion over incomplete observation data. More involved representations for actions such as EigenJoints (Yang et al. 2012; Yang and Tian 2014; Wu and Shao 2014) have also been descibed.

Some approaches have been developed that do perform better than our approach strictly in terms of activity recognition accuracy on benchmark datasets. Typically they are based either on custom-designed features (Wang et al. 2012; Xia and Aggarwal 2013; Yu et al. 2016) or deep learning networks (Du et al. 2015; Zhu et al. 2016), but have also been based on sparse coding (Harandi et al. 2012). These approaches sacrifice performance for accuracy and do not claim to work in real-time.

Different from previous skeleton-based human representations, our BIPOD representation is bio-inspired with a clear interpretation in human anatomy research (Gray 1973; Yokochi and Rohen 2006). Another significant difference is that our predictive representation is developed for activity prediction, instead of activity classification as in most previous works.

2.3 Activity prediction

Different from conventional action classification (Zhang and Parker 2011; Aggarwal and Xia 2014), 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 (Mori et al. 2006). A max-margin early event detector was implemented in Hoai and De la Torre (2014), which modifies structured output SVM to detect early events. Logistic regression models (Ellis et al. 2013) were employed to detect starting point of human activities. An online Conditional Random Field method was introduced in Hoare and Parker (2010) to predict human intents in human–robot collaboration applications. Perez-D'Arpino and Shah (2015) also focused on predicting human intents using a library of motion demonstrations coded as Gaussian distributions. In Yang et al. (2012), an activity classification approach based on a Naïve–Bayes-Nearest-Neighbor classifier was shown to produce similar levels of accuracy after

seeing only 15–20 frames of an action as opposed to the full activity; in essence, predicting the activity. Similarly, Niebles and Fei-Fei (2007) demonstrated a system for recognizing actions based on a single frame, but only showed successful results for very simple gestures. (Ryoo et al. 2015) designed an approach similar to our end-goal, aimed at early recognition of human activities in streaming video, from the robot's perspective. However, this approach does not run in real-time. Kim et al. (2015) presented an approach that could be applied to activity prediction; their work was focused on temporal segmentation; or dividing a sequence by activity as it occurs. To do this efficiently they developed 'event transition segments' and 'event transition probabilities'. Being able to classify these correctly makes temporal segmentation possible but also requires the ability to recognize these featues (or the absence of these features) during an activity—in essence, early activity recognition. In general, 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. Significantly different from these techniques, we focus on developing an accurate, efficient fundamental representation of humans that can be directly used by learning approaches (such as learning from demonstration approaches like Akgun et al. (2012)).

Several approaches were implemented at the representation level to predict human behaviors. For example, Pentland and Liu (1999) represented human actions as a series of Kalman filters linked together in a sequence, as a Markov chain. From this they were able to create a system that predicted the actions of automobile drivers by observing preparatory movements. Our work builds on this by incorporating Extended Kalman Filters—building on their predictive power but recognizing that joint motion for an entire skeleton is inherently a non-linear problem. Other works have built on this Markov-based approach, particularly focused on predicting driver intention based on actions (either the actions of the human driver or the actions of the vehicle (Berndt et al. 2008; Dai et al. 2011; Jin et al. 2011; He et al. 2012; Bi et al. 2013; Georgiou and Demiris 2015; Bosurgi et al. 2014), with more reviewed in (Meiring and Myburgh 2015; Wang et al. 2014b)). Similar work was done in Boussemart and Cummings (2011), using Hidden Markov Models to predict future actions of a human supervisor. The action predictions were used to control the behavior of a robotic system. A Markov-based approach was also developed by Wang et al. (2014c); their method is focused on early prediction of human actions in order to facilitate human–robot interaction with a table tennis playing robot. Their use case requires early activity recognition in order for the robot to react to the shots of the human player. They formulate the problem as a Markov Decision Process, with visual observations being used to select the correct anticipatory action. An approach that also incorporates the ability of Markov functions to describe temporal dependencies is described in Li et al. (2012) and Li and Fu (2014). Their work is not based on recognizing actions from skeleton data, but instead models relationships between simple constituent actions to predict more complex activities. They present a Predictive Accumulative Function to incorporate temporal sequences, probabilities of causal relationships between actions, and context cues relating objects and actions symbolically.

A system that represents high-level activities as a series of logical predicates was developed in Ryoo et al. (2010), which was able to analyze the progress of activities based on sub-event results. They expanded on their work with a dynamic Bag-of-Words (BoW) approach in Ryoo (2011) 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 Yu et al. (2012) 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 in general is not applicable in real-time onboard robotics applications. 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.

3 Skeleton-based BIPOD 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.

3.1 Foundation in biology

In human anatomy, human motions are described in three dimensions according to a series of planes named anatomical planes (Gray 1973; McGinnis 1999; Yokochi and Rohen 2006). There are three anatomical planes of motions that pass through the human body, as demonstrated in Fig. 3:

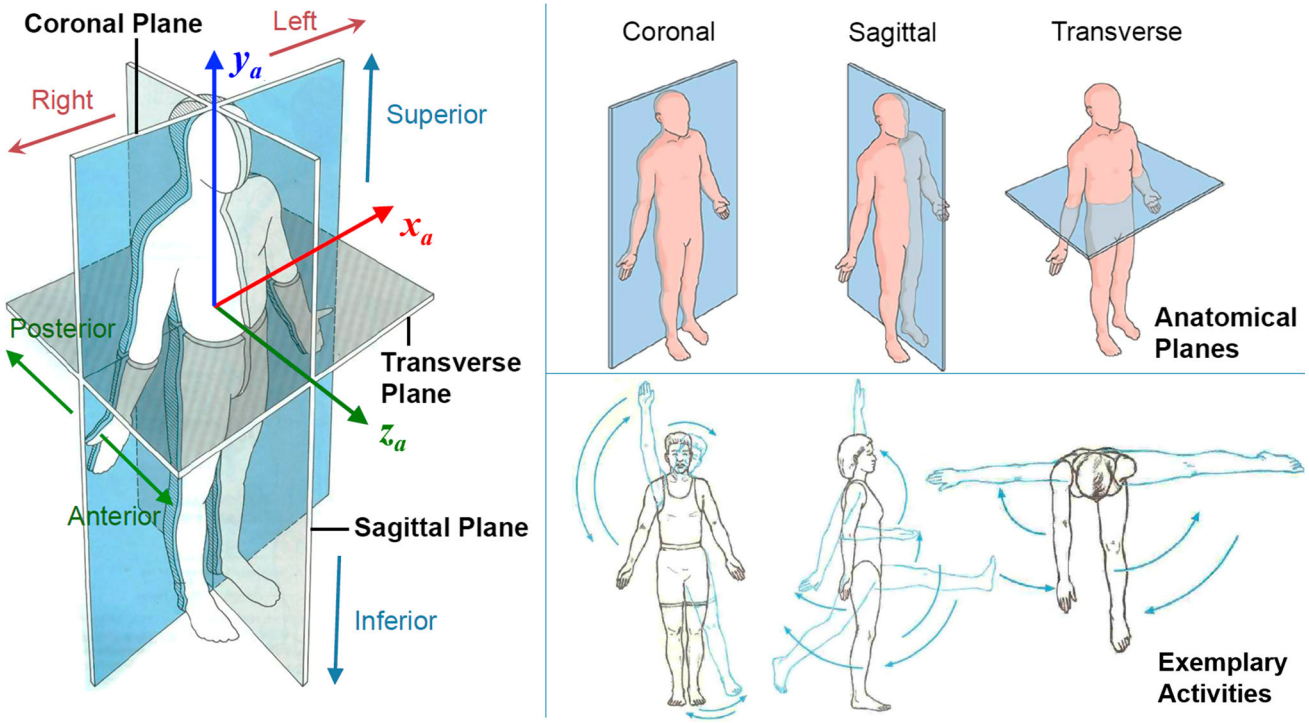


Fig. 3 Our bio-inspired representation is based on the anatomical planes in human anatomy research. This figure demonstrates how anatomical planes divide the human body into different portions and illustrates exemplary human motions performed in each anatomical plane (McGinnis 1999)

- *Sagittal plane* divides the body into right and left parts;
- *Coronal (frontal) plane* divides the human body into anterior (front) and posterior (back) portions;
- *Transverse (horizontal) plane* divides the human body into superior (top) and inferior (bottom) 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* (McGinnis 1999). For example during walking, the hip will be flexing/extending in the sagittal plane, adducting/abducting in the frontal plane and internally/externally rotating in the transverse plane. In human anatomy research (Gray 1973; Yokochi and Rohen 2006), it has been theoretically proven 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 into 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. Finally, our approach is independent of the skeleton acquisition source, making it applicable to multiple domains regardless of what source fits the problem best.

3.2 Estimation of anatomical planes

A core procedure of our bio-inspired human representation is to estimate anatomical planes, which involves three major steps: inferring (1) the *coronal axis* z_a (intersection of the sagittal and transverse planes), (2) *transverse axis* y_a (intersection of the coronal and sagittal planes), and (3) *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.

3.2.1 Estimating coronal axis z_a

Since the coronal plane is represented by human torso in anatomy (McGinnis 1999), 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, and described in Algorithm 1. Formally, given

Algorithm 1: Estimation of Human Coronal Plane

Input : ϵ_n (surface normal tolerance), ϵ (distance tolerance), I_{max} (maximum iterations), and $\boldsymbol{X}$ (torso joints in 3D space)
Output: $\boldsymbol{A}$ (parameter of coronal plane)

repeat
  Randomly select three torso joints $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\} \in \boldsymbol{X}$;
  Estimate the parameter $\boldsymbol{A}$ of the coronal plane $\boldsymbol{\Pi}_0$ using $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$;
  Compute angles between the coronal plane's surface normal $\boldsymbol{n} = (a, b, c)$ and the standard basis $\boldsymbol{e}_z = (0, 0, 1)$ along the z-axis in the real world coordinate: $\cos\theta_z = \boldsymbol{n} \cdot \boldsymbol{e}_z / \|\boldsymbol{n}\|$;
until $(1 - |\cos\theta_z|) \leq \epsilon_n$;
for $i \leftarrow 1$ **to** I_{max} **do**
  Randomly select three non-collinear torso joints from $\boldsymbol{\Pi}_{i-1}$, i.e., $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\} \in \boldsymbol{\Pi}_{i-1}$;
  Estimate the parameter $\boldsymbol{A}_i$ with $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$;
  Extract a set of points belonging to the plane: $\boldsymbol{\Pi}_i = \{\boldsymbol{x} \in \boldsymbol{X}^0 : \text{dis}(\boldsymbol{x}, \boldsymbol{A}_i) \leq \epsilon\}$;
  if $|\boldsymbol{\Pi}_i| < |\boldsymbol{C}_{i-1}|$ **then** Set $\boldsymbol{\Pi}_i = \boldsymbol{\Pi}_{i-1}$
end
Re-estimate $\boldsymbol{A}$ that best fits all points in $\boldsymbol{\Pi}_{I_{max}}$;
return $\boldsymbol{A}$

M torso joints $P_i = (x_i, y_i, z_i)$, $i = 1, \cdots, M$, the objective is to estimate the parameters A, B, C and D, so that the plane $Ax + By + Cz + D = 0$ can best fit the human torso joints in the sense that the sum of distance from all the torso joints to the coronal plane $Ax + By + Cz + D = 0$ is minimized.

Each torso joint p_i that lies on the coronal plane satisfies the plane equation, which means $Ax_i + By_i + Cz_i + D = 0$, and the plane can be represented by $A(x - x_c) + B(y - y_c) + C(z - z_c) = 0$, where (x_c, y_c, z_c) is the coordinates of the joint that lies on the plane. In this paper, $((x_c, y_c, z_c))$ is estimated by the center of all the torso joints. Then, only (A, B, C) is needed to confirm the coronal plane, since D can be obtained by $D = -(Ax_c + By_c + Cz_c)$.

In reality however, only very few joints lie exactly on the coronal plane, hence the value ϵ is introduced to stand for the fitting error. The joints p_j lie outside of the coronal plane satisfy the following equation

$$A(x_j - x_c) + B(y_j - y_c) + C(z_j - z_c) = \epsilon_j \tag{1}$$

The estimation of parameters (A, B, C) of the human coronal plane yields the regression problem as follows

$$R(A, B, C) = \sum_{i=1}^{M} \epsilon_i^2, \tag{2}$$

which can be easily solved by the SVD-based method (Mandel 1982). It is noteworthy that the estimated coronal plane's surface normal (A, B, C) 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 (McGinnis 1999), 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 the human head (as demonstrated in Fig. 2a) in the color image. Then, a highly efficient, off-the-shelf human face detector, based on Haar cascades (Viola and Jones 2001), 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 standard basis of the coronal axis z_a is pointing in the general direction of the sensor, though still orthogonal to the coronal plane (as seen in Fig. 3).

3.2.2 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 shown in Fig. 3.

3.3 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.

3.3.1 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

$$\boldsymbol{P} = \{\boldsymbol{p}_t^{(x_a y_a)}, \quad \boldsymbol{p}_t^{(y_a z_a)}, \quad \boldsymbol{p}_t^{(z_a x_a)}\}_{t=1}^{T} \tag{3}$$

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 human anatomy research (McGinnis 1999).

3.3.2 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:

$$\theta_t = \arccos \frac{\overrightarrow{\boldsymbol{p}_{t-1}\boldsymbol{p}_t} \cdot \overrightarrow{\boldsymbol{p}_t\boldsymbol{p}_{t+1}}}{\|\overrightarrow{\boldsymbol{p}_{t-1}\boldsymbol{p}_t}\| \|\overrightarrow{\boldsymbol{p}_t\boldsymbol{p}_{t+1}}\|}, \quad t = 2, \ldots, T-1 \tag{4}$$

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.

It is noted that the oriented angles computed based on Eq. (4) 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. (4). 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. Figure 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. This allows the pyramid to represent short term movements (e.g. the slight

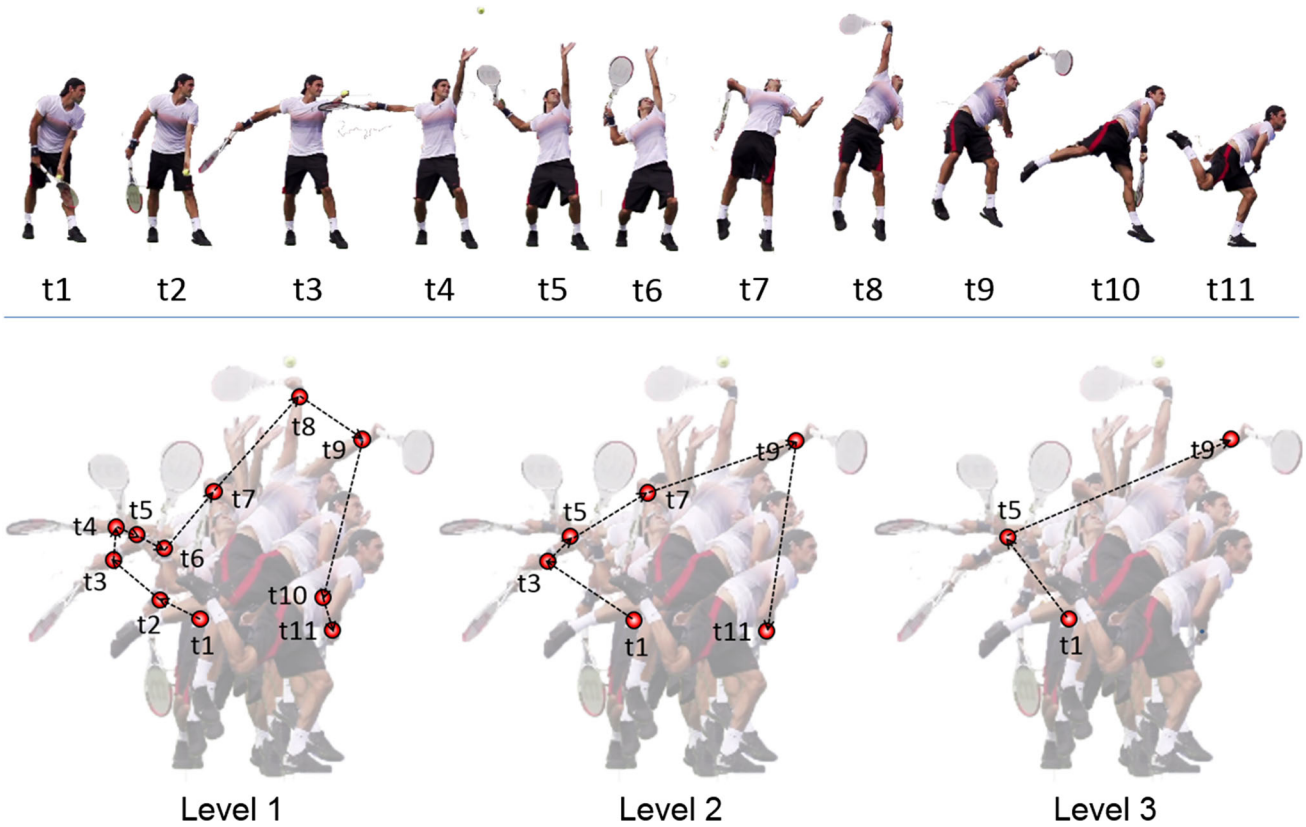


Fig. 4 An example of the temporal pyramid applied in our approach to capture long-term independencies. In this example, a temporal sequence of eleven frames is used to represent a tennis-serve activity, and the joint we are interested in is the right wrist, as denoted by the red dots. When three levels 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

movement of a wrist) as well as long term movements (the entire tennis swing).

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.

3.4 Joint trajectory refinement and prediction

Because skeletal data acquired from 3D robot perception systems can be noisy or body parts can be occluded, 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) (Einicke and White 1999) are used, which are a non-linear extension of Kalman filters and have been successfully applied in many robotics applications. 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). Relevant movement patterns of the body joints in different subspaces are typically assumed to be independent—while a knee's movement is related to the corresponding foot, there is no reason to assume a knee is related to an elbow.. For example, in many scenarios, the hands move independently of the legs, such as using a computer. 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 sensing 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, as EKFs assume state changes are consistent. 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. Since our BIPOD approach can process skeleton data much faster than typical RGB-D sensors are capable of providing it, this means our representation can be used at any time, making it relevant to a large variety of use cases that typical representations cannot accomodate.

The EKF portion of our approach is based on a non-linear Kalman Filter implemented for each of the state subspaces described above. This has the advantage of not only reducing computation (as calculating state changes for multiple low-dimensional spaces is more efficient than calculating state changes for a single high-dimension space), but also makes our implementation flexible, as applications that require only portions of the body (e.g., just arms for controlling an entertainment system from a couch, or just legs for a biomechanist interested in analyzing gait) are able to compute a representation for only the portions of the skeleton that are relevant. This allows a given implementation to select a more discriminative set of skeleton joints based on the use case, or simply reduce computational overhead by disregarding joints which do not impact the application. Each filter maintains as a state x_k the previous (x, y, z) coordinates of a joint and the current (x, y, z) coordinates of a joint—so the 'left-arm' Kalman filter tracks a 12-dimensional state (6 dimensions for the left hand and 6 dimensions for the left elbow). A measurement update z_k consists of the measured (x, y, z) of a joint. Finally, each filter has a transition model A and a measurement model H. The state x_k is updated according to the transition model:

$$x_k = A_k x_{k-1} \tag{5}$$

Then, corrections are made according to the following measurement model:

$$z_k = H_k x_k \tag{6}$$

The models in Eqs. (5) and (6) are nonlinear since the transition and measurement matrices A_k and H_k are obtained using first order approximation and are time-variant. It is different from the Linear Kalman Filter, where both the matrices are constant. The output of the EKF filter is used as the 'actual' position. This process is able to smooth out noise and deal with occluded joints, and can provide a predicted future position for the joints.

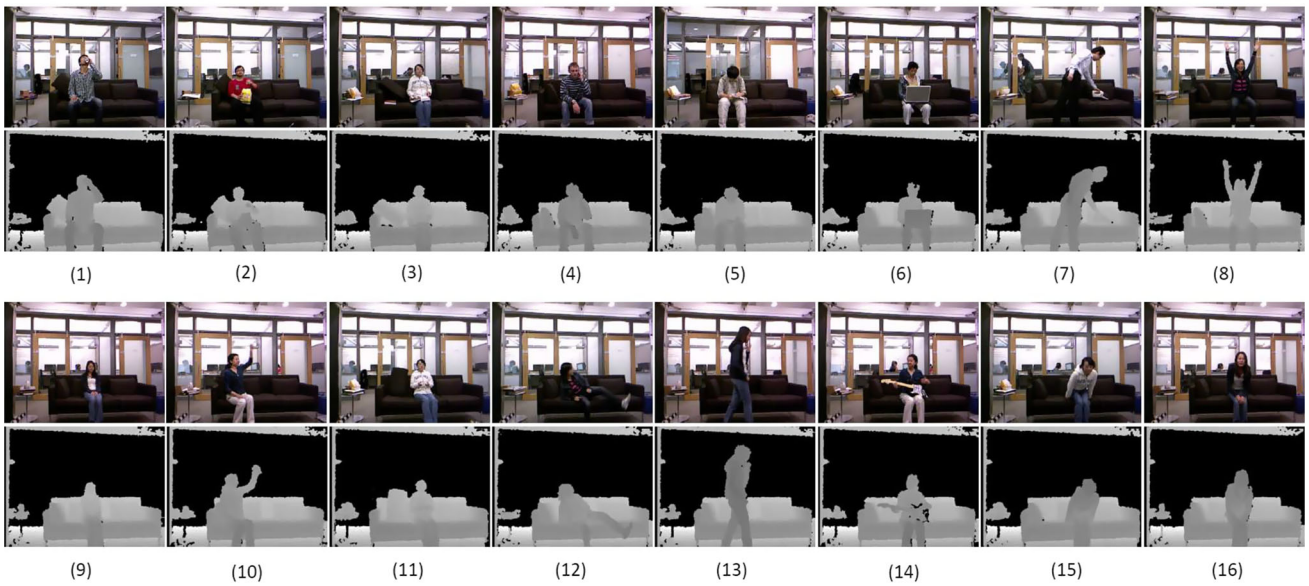


Fig. 5 The MSR Daily Activity 3D dataset is applied in the experiment to evaluate our BIPOD representation, which contains 16 activity categories: (1) drink, (2) eat, (3) read book, (4) call cellphone, (5) write on a paper, (6) use laptop, (7) use vacuum cleaner, (8) cheer up, (9) sit still, (10) toss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

4 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 both a TurtleBot2 robot and a Baxter humanoid robot to perform real-time online activity recognition.

4.1 Implementation

Our skeleton-based BIPOD representation is implemented using a mixture of the Python and C++ programming languages on various typical Linux machines. In the case of the comparison with benchmark datasets and the test on the Baxter robot, a desktop workstation with an i7 3.0 Ghz CPU with 16 GB of memory was used; the test on the TurtleBot2 was performed on an i3 1.7 Ghz CPU with 4 GB of memory. Each of the three histograms, computed from the trajectories on the coronal, transverse and sagittal planes, contains 12 bins. The histograms are concatenated to form a final feature vector. The learner employed in this paper is the non-linear Support Vector Machine (SVM) with χ^2-kernels (Chang and Lin 2011), which has demonstrated superior performance on the histogram-based feature (Vedaldi and Zisserman 2012). To address multi-class classification and prediction, the standard one-against-one methodology is applied (Chang and Lin 2011).

4.2 Evaluation on MSR activity daily 3D dataset

The MSR Daily Activity 3D dataset (Wang et al. 2012)[2] is a 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. 2b. In our experiments, we follow the experimental setups used in Xia and Aggarwal (2013); 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 are presented in Table 1. 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 other real-time state-of-the-art skeleton-based representations in human activity recognition tasks, as presented in Table 1. It is observed that our approach outperforms previous works and obtains the best recognition accuracy over this dataset. We note that approaches exist, as described in Sect. 2, that do obtain higher recognition accuracy at the expense of runtime. In addition, we evaluate the

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

Table 1 Comparison of average recognition accuracy with previous skeleton-based representations on MSR Daily Activity 3D

Skeleton-based representations	Accuracy (%)
Dynamic temporal warping (Wang et al. 2012)	54.0
Distinctive canonical poses (Ellis et al. 2013)	65.7
Actionlet ensemble (3D pose only) (Wang et al. 2012)	68.0
Relative position of joints (Seidenari et al. 2013)	70.0
Moving pose (Zanfir et al. 2013)	73.8
Fourier temporal pyramid (Wang et al. 2012)	78.0
Our BIPOD representation	79.7

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.

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. For each successive experiment, 15% future unobserved data are predicted by the component procedure of joint trajectory refinement and prediction by the EKFs, as discussed in Sect. 3.4. 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. 6a. 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.

4.3 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 (Müller et al. 2007)[3] 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. 2c. 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 Müller et al. (2007), we simply assume subjects face toward the view point.

The experimental setup applied in Gowayyed et al. (2013) and Ofli et al. (2014) 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 2 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 real-time skeleton-based human representations over the same dataset, which is reported in Table 2. A similar phenomenon is observed that our BIPOD representation obtains a superior human activity recognition accuracy and outperforms existing skeleton-based representations. Again, we note that approaches exist that do obtain higher recognition accuracy at the expense of runtime. 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. Figure 6b 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.

4.4 Real-world validation on TurtleBot2 using cross-training

Besides evaluating our representation's performance using benchmark datasets, we also implemented our BIPOD approach on physical robots to evaluate how well it performs in real-world robotics applications. The robotic platform used

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

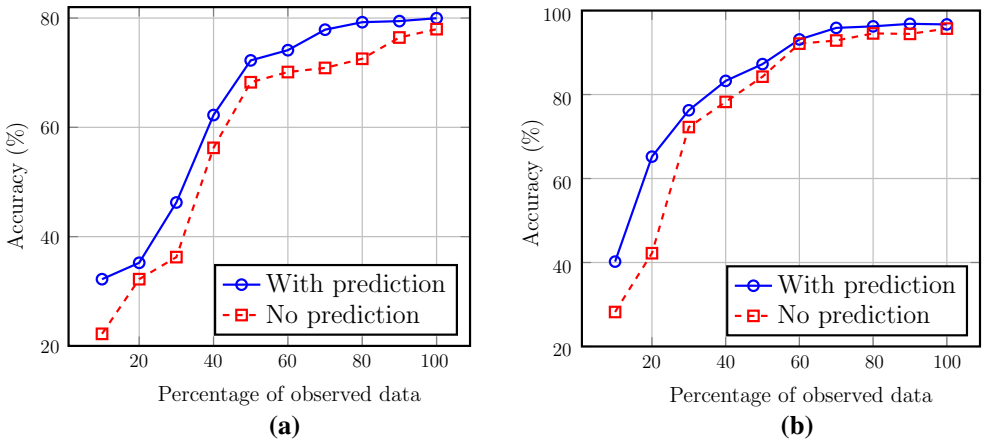


Fig. 6 Experimental results of using our BIPOD representation to predict human activities given incomplete observations. When the procedure of joint trajectory refinement and prediction is used, 15% future data are predicted at each test point. Generally, the predictive representation greatly outperforms representations without prediction capabilities. **a** MSR Daily Activity 3D, **b** HDM05 MoCap

Table 2 Comparison of average recognition accuracy with previous skeleton-based representations on HDM05 MoCap

Skeleton-based representations	Accuracy (%)
Trifocal tensor of joint positions (Liu and Cao 2012)	80.86
Sequence of most informative joints (Ofli et al. 2014)	84.40
Subtensor of joint positions (Liu and Cao 2012)	85.71
Relevant joint positions (López-Mendez et al. 2012)	92.20
Cov3DJ (Hussein et al. 2013)	95.41
Our BIPOD representation	96.70

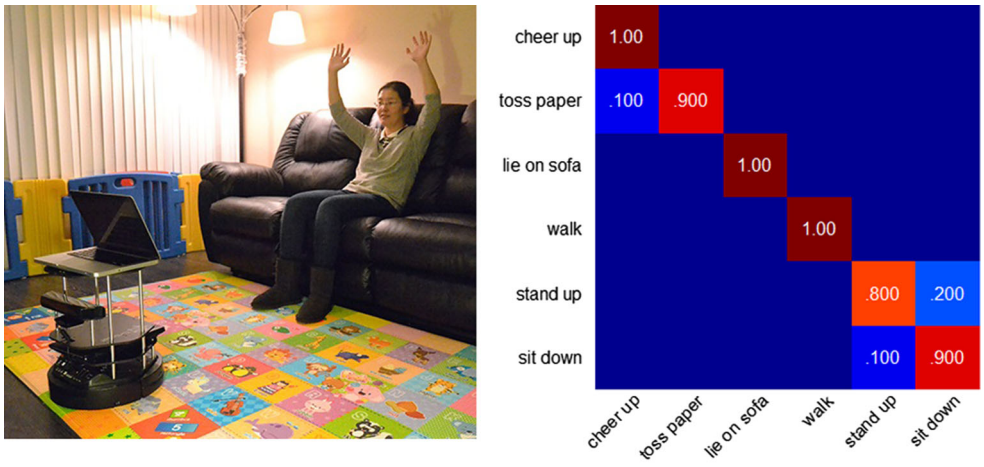
in our first experiment is the TurtleBot2 robot built upon the Kobuki mobile base. The robot employs a Kinect sensor for on-board 3D perception and an ASUS netbook (with 1.6 Ghz dual core CPU and 2 GB memory) for on-board control. To compute our representation and perform activity recognition in real time, another laptop, as described in Sect. 4.1, is placed on top of the robot. The hardware configuration of our TurtleBot2 robot is illustrated in Fig. 7a.

In order to validate our approach's generalization ability to process data from different skeleton estimation techniques, we apply a cross-training methodology in this experiment. Specifically, our BIPOD representation and the SVM classifier are trained using skeleton data obtained from the Microsoft Kinect SDK, which provides information of 20 joints, as shown in Fig. 2b. 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. 2a. 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 possible to avoid collecting a new dataset and therefore can significantly save human labor. Additionally, it demonstrates that our BIPOD representation has practical real-world uses. By using training data from an existing dataset recorded in a separate environment to recognize the same classes of activities in this new environment, we show that BIPOD is not just tuned to a particular dataset or a specific lab environment.

In this experiment, the MSR Daily Activity 3D dataset, as discussed in Sect. 4.2, is used to compute our representation and estimate the SVM's parameters. Six activity classes were used in this validation exercise: 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 human 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 s and the overlap of temporally adjacent windows is 1 s. Then, human 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. 7a. Ground truth is manually recorded and used to compare with recognition results from the TurtleBot2 robot for quantitative evaluation.

Figure 7b 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



(a) (b)

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. 7a. The confusion matrix obtained in this experiment is presented in Fig. 7b. **a** Experiment setups, **b** confusion matrix

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.

4.5 Real-world validation on a Baxter robot

Finally, we tested BIPOD on a Baxter humanoid robot. The Baxter robot, seen interacting in Fig. 8a, is ideally suited for human–robot interaction as it can capably mimic human activities. Additionally, it is designed for safe operation around humans, unlike typical industrial robotic arms. Our Baxter robot was equipped with a Kinect sensor that is mounted on it's 'chest', and all processing was done on a networked Linux desktop. Skeleton data was obtained through OpenNI running on ROS, providing 15 joints.

We created four new activites that would be used to interact with Baxter in order to control the robot through the process of making and serving a drink. Each is able to be performed on either side of the body, for a total of eight new activities: (1) pick up, (2) pour, (3) serve, and (4) put down. Ideally, a user could command Baxter to pick up a glass and a beverage, pour a drink, and serve it, with either arm.

These human activities were designed to be bilateral for two important reasons. First, BIPOD divides the body into three planes, one being the sagittal plane which divides the body into left and right halves. This means BIPOD explicitly encodes left/right information into it's representation—something often lacking in other representations and datasets (e.g., the MSR dataset classifies waving with either the left or right arm as a single activity class). Second, creating separate activity classes for actions based on the side of the body they are performed with makes human–robot interaction easier and more intuitive. The human can use this ability to control a specific arm or side of the robot he or she is interacting with, and it provides options to those who may have one arm missing or disabled.

Two human subjects were recorded performing 20 executions per subject of each action, for a total of 40 instances of each action and 320 instances overall. The actions were recorded in an open-room lab environment, seen in Fig. 8a. While the dataset available online was recorded as discrete executions of these actions, our approach allows these actions to be conducted without separation to actively direct and interact with Baxter. Because these were new activities not present in an existing dataset, half of the data was used for training and half for testing. In order to validate our BIPOD representation, we compared it with two current popular skeletal representations: histograms of oriented dis-

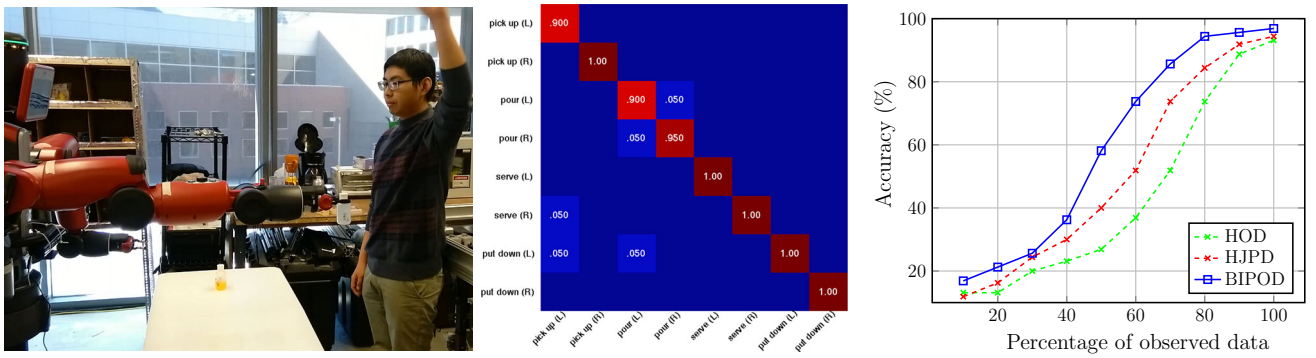


Fig. 8 Real-world validation on a Baxter robot in a human–robot interaction application. Figure 8a illustrates the experimental setup, with a human interacting with Baxter. Figure 8b shows the confusion matrix for activity classification, and Fig. 8c displays the accuracy rates for early prediction of human activities for HOD (Gowayyed et al. 2013), HJPD (Rahmani et al. 2014), and the proposed BIPOD. **a** Environment Setup, **b** confusion matrix and **c** prediction ability

Table 3 Comparison of BIPOD with HOD (Gowayyed et al. 2013) and HJPD (Rahmani et al. 2014) on classification accuracy over the eight activities created for interaction with Baxter

Representations	Pick up (L)	Pick up (R)	Pour (L)	Pour (R)	Serve (L)	Serve (R)	Put down (L)	Put down (R)	Overall
HOD	95	90	100	75	90	95	100	100	93.13
HJPD	100	95	100	95	95	90	90	90	94.38
BIPOD	95	100	90	95	100	100	100	100	**96.88**

Bold value indicates the superior performance of our representation

placements (HOD) (Gowayyed et al. 2013), and histograms of joint position differences (HJPD) (Rahmani et al. 2014).

Figure 8b illustrates the confusion matrix produced by our described BIPOD approach, where each column represents the predicted class and each row represents the actual class. All eight activities are shown to include both their left and right versions, labeled 'pick up (L)', 'pick up (R)', etc. As it illustrates, BIPOD is able to capably recognize the described actions, with no inaccuracies coming due to the bilateral nature of the human activities. Overall, our approach classifies only four instances incorrectly, which makes it 96.88% accurate. A comparison with HOD and HJPD is shown in Table 3. While HOD and HJPD both perform well (93.13 and 94.38%, respectively), BIPOD does outperform both of them.

Additionally, BIPOD outperforms both of these representations in the early prediction of activities. BIPOD's activity prediction capabilities are quantified in Fig. 8c, compared to HOD and HJPD. As these figures demonstrate, BIPOD is more accurate at predicting an activity class early at every point in time. It reaches over 50% accuracy only halfway through the activity, beating HOD by 31.25% and HJPD by 18.13% at that point of time.

Finally, BIPOD runs significantly faster than the 30 frames per second of data that is provided by the Kinect or a comparable RGB-D sensor and associated software (e.g., OpenNI). Because of this, the joint position interpolation provided by the EKF is extremely useful for systems built on BIPOD's capabilities. On a 2.7 Ghz laptop with 4 GB of memory, the BIPOD representation can be constructed at 3600 'frames' per second—a 'frame' being either a representation constructed from 3D skeleton data (with EKF processing to reduce noise) or a representation interpolated by the EKFs between frames of actual 3D skeleton data. Using a SVM as a classifier, these representations are able to be classified at a rate of 2800 per second, which allows for high-speed human–robot interaction.

5 Discussion

Our skeleton-based representation based upon bio-inspired predictive orientation decomposition possesses several desirable characteristics, many of which make it unique in the field and an ideal approach for the activity prediction part of human–robot interaction. Our BIPOD human representation is a bio-inspired approach, which has a clear biological interpretation in human anatomy. This makes it ideal for the increasing crossover of computer vision, robotics, and bio-mechanics. It's biology inspired roots means it builds on research about human biology and anatomy, instead of

attempting to reinvent it in a way that makes sense to computer scientists. This is apparent in it's ability to clearly distinguish bilateral actions, something often not considered in other representations and major existing datasets but a necessary feature when robots need to interact closely with humans. For example, this means it can distinguish between waving with the left hand versus waving with the right hand, while they are labeled as the same action in the MSR Daily Activity 3D dataset.

Additionally, it possesses several desirable characteristics from a computer vision and machine learning standpoint. Through spatially decomposing joint trajectories and projecting them onto anatomical planes, our human representation is invariant to view point changes—a subject viewed from any camera angle will still result in the same calculation of anatomical planes. 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. It also runs at a speed which will allow it to be applicable as RGB-D sensors improve. Currently it's ability to interpolate between frames makes it ideal for time sensitive actions, but it's speed also means it will adapt well as frame rates improve and skeleton data is available faster than 30 frames per second. Finally, the division of joint spaces into separate Kalman filters means that BIPOD is able to adapt to many applications; e.g., it can be easily altered to represent only portions of the body and therefore only predict and recognize human activities from that portion.

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. However, this inadequacy would be due to limitations in RGB-D sensor capabilities: BIPOD would not be effected by objects if skeletal data were obtained from motion capture systems. Additionally, BIPOD's use of Kalman filters as a noise reduction method means it would recover quickly from joint position changes caused by noisy sensors. 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.

An additional limitation is that the focus on real-time performance and activity prediction means that some accuracy is sacrificed. While BIPOD produces results near state-of-the-art on benchmark datasets, approaches do exist, and were covered in Sect. 2, that do have more accurate recognition.

6 Conclusion

In this paper, we introduce the novel BIPOD representation to enable intelligent robots to predict human activities in real time from 3D skeletal data in practical human-centered robotics applications. Our BIPOD approach is inspired by biological human anatomy research, which provides theoretical guarantees that the proposed representation is able to encode all human movements. To construct the BIPOD representation, we estimate human anatomical planes, decompose 3D skeleton trajectories, and project them onto the anatomical planes. 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 or occluded joints. We perform empirical studies, using both a TurtleBot2 mobile robot and a Baxter humanoid robot, to validate the performance of our BIPOD representation in an ongoing human activity recognition task, and demonstrate our representation's real-world and online capabilities on both platforms. 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, as well as a new dataset that is recorded specifically for interaction with Baxter. Experimental results demonstrate that BIPOD significantly improves human activity recognition accuracy and efficiency and successfully addresses the challenging activity prediction problem in real time.

References

Aggarwal, J., & Xia, L. (2014). Human activity recognition from 3D data: A review. *Pattern Recognition Letters*, *48*, 70–80.

Akgun, B., Cakmak, M., Jiang, K., & Thomaz, A. (2012). Keyframe-based learning from demonstration. *Internation Journal of Social Robotics*, *4*(4), 343–355.

Berndt, H., Emmert, J., & Dietmayer, K. (2008). Continuous driver intention recognition with hidden Markov models. In *Intelligent Transportation Systems* (pp. 1189–1194).

Bi, L., Yang, X., & Wang, C. (2013). Inferring driver intentions using a driver model based on queuing network. In *Intelligent Vehicles Symposium* (pp. 1387–1391).

Bosurgi, G., D'Andrea, A., & Pellegrino, O. (2014). Prediction of drivers' visual strategy using an analytical model. *Journal of Transportation Safety & Security*, *7*, 153–173.

Boubou, S., & Suzuki, E. (2015). Classifying actions based on histogram of oriented velocity vectors. *Journal of Intelligent Information Systems*, *44*(1), 49–65.

Boussemart, Y., & Cummings, M. L. (2011). Predictive models of human supervisory control behavioral patterns using hidden semi-Markov models. *Engineering Applications of Artifical Intelligence*, *24*, 1252–1262.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transaction on Intelligent Systems and Technology*, *2*, 27:1–27:27.

Charles, J., Everingham, M. (2011). Learning shape models for monocular human pose estimation from the Microsoft Xbox Kinect. In *IEEE international conference on computer vision*.

Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., & Vidal, R. (2013). Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. In *IEEE conference on computer vision and pattern recognition workshop*.

Chen, G., Giuliani, M., Clarke, D., Gaschler, A., & Knoll, A. (2014). Action recognition using ensemble weighted multi-instance learning. In *IEEE international conference on robotics and automation*.

Dai, F., Zhang, J., & Lu, T. (2011). The study of driver's starting intentions. In *Mechanic Automation and Control Engineering* (pp. 2758–2761).

Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1110–1118).

Einicke, G., & White, L. (1999). Robust extended Kalman filtering. *IEEE Transactions on Signal Processing*, *47*(9), 2596–2599.

Ellis, C., Masood, S. Z., Tappen, M. F., Laviola, J. J, Jr., & Sukthankar, R. (2013). Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, *101*(3), 420–436.

Ganapathi, V., Plagemann, C., Koller, D., & Thrun, S. (2010). Real time motion capture using a single time-of-flight camera. In *IEEE conference on computer vision and pattern recognition*.

Georgiou, T., & Demiris, Y. (2015). Predicting car states through learned models of vehicle dynamics and user behaviours. In *Intelligent vehicles symposium* (pp. 1240–1245).

Girshick, R., Shotton, J., Kohli, P., Criminisi, A., & Fitzgibbon, A. (2011). Efficient regression of general-activity human poses from depth images. In *IEEE international conference on computer vision*.

Gowayyed, M. A., Torki, M., Hussein, M. E., & El-Saban, M. (2013). Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition. In *International joint conference on artificial intelligence*.

Gray, H. (1973). *Anatomy of the human body*. Philadelphia: Lea & Febiger.

Han, F., Reily, B., Hoff, W., & Zhang, H. (2016). Space-time representation of people based on 3D skeletal data: A review. ArXiv e-prints 1601.01006.

Han, F., Reily, B., Hoff, W., & Zhang, H. (2017). Space-time representation of people based on 3d skeletal data: A review. *Computer Vision and Image Understanding*, *158*, 85–105.

Harandi, M., Sanderson, C., Hartley, R., & Lovell, B. (2012). Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. *Computer Vision-ECCV*, *2012*, 216–229.

He, L., Cf, Zong, & Wang, C. (2012). Driving intention recognition and behaviour prediction based on a double-layer hidden Markov model. *Journal of Zhejiang University*, *13*, 208–217.

Hoai, M., & De la Torre, F. (2014). Max-margin early event detectors. *International Journal of Computer Vision*, *107*(2), 191–202.

Hoare, J., & Parker, L. (2010). 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*.

Hussein, M. E., Torki, M., Gowayyed, M. A., & El-Saban, M. (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *International joint conference on artificial intelligence*.

Jin, L., Hou, H., & Jiang, Y. (2011). Driver intention recognition based on continuous hidden Markov model. In *Transportation, Mechanical, and Electrical Engineering* (pp. 739–742).

Jung, H. Y., Lee, S., Heo, Y. S., & Yun, I. D. (2015). Random tree walk toward instantaneous 3D human pose estimation. In *IEEE conference on computer vision and pattern recognition*.

Kim, Y., Chen, J., Chang, M. C., Wang, X., Provost, E. M., & Lyu, S. (2015). Modeling transition patterns between events for temporal human action segmentation and classification. In *IEEE international conference and workshops on automatic face and gesture recognition (FG), Ljubljana* (pp. 1–8).

Koppula, H. S., Rudhir, G., & Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*, *32*, 951–970.

Li, K., & Fu, Y. (2014). Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*, 1644–1657.

Li, K., Hu, J., & Fu, Y. (2012). Modeling complex temporal composition of actionlets for activity prediction. In *European conference on computer vision*.

Liu, Q., & Cao, X. (2012). Action recognition using subtensor constraint. In *European conference on computer vision*.

López-Mendez, A., Gall, J., Casas, J. R., & Gool, L. J. V. (2012). Metric learning from poses for temporal clustering of human motion. In *British machine vision conference*.

Luo, J., Wang, W., & Qi, H. (2013). Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE international conference on computer vision*.

Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *The American Statistician*, *36*(1), 15–24.

McGinnis, M. (1999). *Bioregionalism: The tug and pull of place*. London: Routledge.

Meiring, G. A. M., & Myburgh, H. C. (2015). A review of intelligent driving style analysis systems and related artificial intelligence algorithms. *Sensors*, *15*, 30653–30682.

Mori, A., Uchida, S., Kurazume, R., Taniguchi, R. I., Hasegawa, T., & Sakoe, H. (2006). Early recognition and prediction of gestures. In *International conference on pattern recognition*.

Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., & Weber, A. (2007). *Documentation mocap database HDM05*. Technical report, Universität Bonn.

Niebles, J. C., & Fei-Fei, L. (2007). A hierarchical model of shape and appearance for human action classification. In *IEEE conference on computer vision and pattern recognition*.

Nikolaidis, S., Hsu, D., & Srinivasa, S. (2017). Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, *36*(5–7), 618–634.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2014). Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, *25*(1), 24–38.

Pentland, A., & Liu, A. (1999). Modeling and prediction of human behavior. *Neural Computation*, *11*(1), 229–242.

Perez-D'Arpino, C., & Shah, J. A. (2015). Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In *2015 IEEE international conference on robotics and automation (ICRA)* (pp. 6175–6182). IEEE.

Pieropan, A., Salvi, G., Pauwels, K., & Kjellstrom, H. (2014). Audio-visual classification and detection of human manipulation actions. In *IEEE/RSJ international conference on intelligent robots and systems*.

Plagemann, C., Ganapathi, V., Koller, D., & Thrun, S. (2010). Real-time identification and localization of body parts from depth images. In *IEEE international conference on robotics and automation*.

Rahmani, H., Mahmood, A., Mian, A., & Huynh, D. (2014). Real time action recognition using histograms of depth gradients and random decision forests. In *IEEE winter conference on applications of computer vision*.

Ryoo, M. S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In *International conference on computer vision*.

Ryoo, M., Fuchs, T. J., Xia, L., Aggarwal, J. K., & Matthies, L. (2015). Robot-centric activity prediction from first-person videos: What will they do to me? In *Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction* (pp. 295–302). ACM.

Ryoo, M. S., Grauman, K., & Aggarwal, J. K. (2010). A task-driven intelligent workspace system to provide guidance feedback. *Computer Vision and Image Understanding*, *114*(5), 520–534.

Schwarz, L. A., Mkhitaryan, A., Mateus, D., & Navab, N. (2012). Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing*, *30*(3), 217–226.

Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A., & Pala, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *IEEE conference on computer vision and pattern recognition workshops*.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *IEEE conference on computer vision and pattern recognition*.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2012). Unstructured human activity detection from RGBD images. In *IEEE international conference on robotics and automation*.

Vantigodi, S., & Babu, R. V. (2013). Real-time human action recognition from motion capture data. In *National conference on computer vision, pattern recognition, image processing and graphics*.

Vedaldi, A., & Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(3), 480–492.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE conference on computer vision and pattern recognition*.

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *IEEE conference on computer vision and pattern recognition*.

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2014a). Learning actionlet ensemble for 3D human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(5), 914–927.

Wang, W., Xi, J., & Chen, H. (2014b). Modeling and recognizing driver behavior based on driving data: A survey. *Mathematical Problems in Engineering, 2014*, 245641. https://doi.org/10.1155/2014/245641.

Wang, Z., Boularias, A., Mulling, K., Scholkopf, B., & Peters, J. (2014c). Anticipatory action selection for human–robot table tennis. *Artificial Intelligence*, *247*, 399–414.

Wu, D., & Shao, L. (2014). Leveraging hierarchical parametric networks for skeletal joints action segmentation and recognition. In *IEEE conference on computer vision and pattern recognition*.

Xia, L., & Aggarwal, J. K. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *IEEE conference on computer vision and pattern recognition*.

Yang, X., Tian, Y. (2012). EigenJoints-based action recognition using Naï–Bayes-Nearest-Neighbor. In *IEEE conference on computer vision and pattern recognition workshop*.

Yang, X., & Tian, Y. (2014). Effective 3D action recognition using EigenJoints. *Journal of Visual Communication and Image Representation*, *25*(1), 2–11.

Yokochi, C., & Rohen, J. W. (2006). *Color atlas of anatomy: A photographic study of the human body*. Philadelphia: Lippincott Williams & Wilkins.

Yu, G., Yuan, J., & Liu, Z. (2012). Predicting human activities using spatio-temporal structure of interest points. In *ACM international conference on multimedia*.

Yu, M., Liu, L., & Shao, L. (2016). Structure-preserving binary representations for RGB-D action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(8), 1651–1664.

Zanfir, M., Leordeanu, M., & Sminchisescu, C. (2013). The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *IEEE international conference on computer vision*.

Zhang, H., & Parker, L. (2011). 4-dimensional local spatio-temporal features for human activity recognition. In *IEEE/RSJ international conference on intelligent robots and systems*.

Zhang, H., Reardon, C. M., & Parker, L. E. (2013). Real-time multiple human perception with color-depth cameras on a mobile robot. *IEEE Transactions on Cybernetics*, *43*(5), 1429–1441.

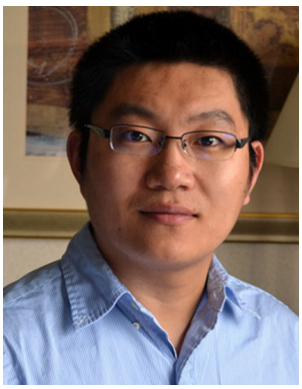
Zhao, X., Li, X., Pang, C., Zhu, X., & Sheng, Q. Z. (2013). Online human gesture recognition from motion data streams. In *ACM international conference on multimedia*.

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., & Xie, X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. arXiv preprint arXiv:160307772.

Brian Reily received the M.S. degree in computer science from the Colorado School of Mines in 2016 and the B.A. degree in computer science from the University of Virginia, Charlottesville in 2009 and served as a Transportation Officer in the US Army from 2010 to 2013. He is currently pursuing further graduate work in Computer Science at Colorado School of Mines. His research interests include computer vision, 3D perception, and machine learning for human pose estimation, human activity recognition, and human–robot interaction applications.

Fei Han received the Ph.D. degree in Automation from the University of Science and Technology of China in 2014, and the B.S. in Automation from the University of Science and Technology of China in 2009. He is currently pursuing the graduate degree in Computer Science at Colorado School of Mines. His research interests include 3D perception, computer vision, machine learning, decision making, artificial cognitive modeling and nonlinear control systems.

Lynne E. Parker received her Ph.D. in computer science from the Massachusetts Institute of Technology. She is the Division Director for the Information and Intelligent Systems Division in the Computer and Information Science and Engineering Directorate at the National Science Foundation. While at NSF, she is on leave from the Electrical Engineering and Computer Science Department at the University of Tennessee, Knoxville (UTK), where she is Professor and previously served as Associate Department Head. Prior to joining the UTK faculty, she worked for several years as a Distinguished Research and Development Staff Member at Oak Ridge National Laboratory. She is a Fellow of IEEE.



Hao Zhang received the Ph.D. degree in Computer Science from the University of Tennessee, Knoxville in 2014, the M.S. in Electrical Engineering from the Chinese Academy of Sciences in 2009, and the B.S. in Electrical Engineering from the University of Science and Technology of China in 2006. He is currently an Assistant Professor in the Department of Electrical Engineering and Computer Science at Colorado School of Mines. His research interests include human–robot teaming, 3D perception, robot learning and decision making, and artificial intelligence. He is a member of IEEE.