HUMAN ACTIVITY RECOGNITION AND GYMNASTICS ANALYSIS THROUGH DEPTH IMAGERY

by

Brian J. Reily

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Computer Science).

Golden, Colorado

Date ______________________

Signed: ______________________
Brian J. Reily

Signed: ______________________
Dr. William Hoff
Thesis Advisor

Signed: ______________________
Dr. Hao Zhang
Thesis Advisor

Golden, Colorado

Date ______________________

Signed: ______________________
Dr. Atef Elsherbeni
Professor and Interim Head
Department of Electrical Engineering and Computer Science

ABSTRACT

Depth imagery is transforming many areas of computer vision, such as object recognition, human detection, human activity recognition, and sports analysis. The goal of my work is twofold: (1) use depth imagery to effectively analyze the pommel horse event in men's gymnastics, and (2) explore and build upon the use of depth imagery to recognize human activities through skeleton representation. I show that my gymnastics analysis system can accurately segment a scene based on depth to identify a 'depth of interest', ably recognize activities on the pommel horse using only the gymnast's silhouette, and provide an informative analysis of the gymnast's performance. This system runs in real-time on an inexpensive laptop, and has been built into an application in use by elite gymnastics coaches. Furthermore, I present my work expanding on a bio-inspired skeleton representation obtained through depth data. This representation outperforms existing methods in classification accuracy on benchmark datasets. I then show that it can be used to interact in real-time with a Baxter humanoid robot, and is more accurate at recognizing both complete and ongoing interactions than current state-of-the-art methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisors, William Hoff and Hao Zhang, for all of their guidance, instruction, and patience through weekly meetings over the course of this masters degree, as well as giving me the freedom to explore this field. I also owe them and a local gymnastics facility a debt of gratitute for the opportunity to work on this very unique project. I'd like to thank the members of my thesis committee. I'd like to thank Evan Balogh, Travis Johnson, Austin Kauffman, and Zac McClain for their work in building my methods into a user-friendly application. I'd like to thank Fei Han for his work on skeleton representations. Lastly, I'd like to thank my parents and Erin for all of the encouragement.

CHAPTER 1
INTRODUCTION

The availability of depth imagery is having a transformative effect on many areas of computer vision. The introduction of color-depth, or RGB-D, cameras such as Kinect [1] have made depth imagery cheap and easy to obtain, providing researchers with information about the 3D nature of a scene. In addition, the development of fast and accurate methods to identify joint locations using depth data [2] has popularized the use of skeleton-based representations for humans and actions. These two advancements have impacted a number of fields; in this thesis I will focus on their impact and use in the area of gymnastics analysis and human activity recognition.

Sports analysis methods have been developed that use a variety of data sources; computer vision in particular has been used extensively but is limited by trying to analyze 3D activities through 2D data (RGB imagery). Depth imagery introduces opportunities that traditional intensity imagery does not offer; methods can be more accurate, more efficient, or analyze new areas of a sport. Sports analysis is also limited by the need for expert input on what aspects of a sport are important to analyze; this limitation is magnified in sports such as gymnastics, where evaluation is based on human judges, and is thus subjective instead of objectively based on statistics. My work to analyze the pommel horse event in men's gymnastics was developed with the assistance of professional gymnastics coaches. The pommel horse event, seen in Figure 3.1, consists of a gymnast performing various moves while supporting themselves with only their hands on the apparatus. The most common and important move is spinning or performing circles, and key to being judged well is performing these circles with very consistent timing.

Depth imagery has also revolutionized the field of human activity recognition; current research is based almost exclusively on representations of human skeletons obtained through devices like the Kinect. Activity recognition is important in many fields, such as security, assisted living, sports analysis, and human-robot interaction. Efficient and accurate activity classification is important and many methods have been proposed. The problem of activity prediction - classifying an action based on incomplete or ongoing data - is both more critical and less reearched. Predicting activities accurately can enable security systems to identify crimes as they are committed, or simply allow a robot to discern a human's intention before they finish giving a command.

This thesis makes four important contributions. First, I provide a comprehensive review of the current state-of-the-art work in human skeleton data acquisition and representation. This review, a condensed version of [3], focuses on the most recent and advanced methods, and provides an effective categorization based on the approaches. Second, I introduce a dataset of gymnasts performing routines on the pommel horse apparatus, recorded as depth data with a portable Kinect 2 camera, with the hope that other researchers may find it useful for developing further sports analysis techniques. Third, I introduce a novel system that integrates multiple highly efficient methods to analyze gymnastics performance in real-time, which assists professional gymnastics coaches. Finally, I present my contributions to BIPOD, a skeleton-based 3D representation of humans based on human anatomy research, that effectively encodes spatio-temporal information of human actions. BIPOD is able to reduce noise in observed data, predict activities in real-time, and recognize incomplete activities more accurately than existing methods. I demonstrate that BIPOD is an effective solution for human-robot interactions.

Work in this thesis has been presented in three publications that I am the lead or co-lead author on:

- 'Space-Time Representation of People Based on 3D Skeletal Data: A Review' by Han, Reily, Hoff, and Zhang [3], currently submitted for publication to *ACM Computing Surveys*.

- ‘Skeleton-Based Bio-Inspired Human Activity Prediction for Real-time Human-Robot Interaction’ by Reily, Han, Parker, and Zhang [4], currently submitted for publication to *IEEE Transactions on Cybernetics*.

- ‘Real-time Gymnast Detection and Performance Analysis with a Portable 3D Camera’ by Reily, Zhang, and Hoff [5], currently submitted for publication to *Computer Vision and Image Understanding*, special issue on Computer Vision in Sports.

The rest of the paper is structured as follows: In Chapter 2, I discuss the current state of research in the areas of human detection, RGB-D sensors and skeleton acquisition methods, skeleton representation approaches, activity recognition and classification, and finally sports analysis, specifically gymnastics. In Chapter 3, I introduce my system to automatically analyze the performance of gymnasts using depth imagery. Specifically, I discuss the dataset I created, describe my approach to the problem, discuss my experimental results, and finally introduce the real-world application created to validate my methods. In Chapter 4, I discuss my work towards the BIPOD skeleton representation. Specifically, I introduce the motivation and approach, and present experimental results. Finally in Chapter 5, I conclude the paper.

CHAPTER 2
RELATED WORK

Human activity recognition is a broad field, involving many areas of computer vision. Since this thesis covered topics along the entire path from the depth image source to the resulting output, it was important to review current work in a multitude of areas. I cover detecting humans in a scene, acquiring skeleton data from RGB-D cameras, representing those skeletons effectively, recognizing human activities, and analyzing sports and gymnastics.

2.1 Human Detection and Regions of Interest

A variety of methods have been developed to automatically detect humans and regions of interest in a scene. In 2D images, gradient based features like SIFT [6] resulted in the use of histograms of oriented gradients (HOG) to detect pedestrians developed by Dalal et al. [7, 8]. Bourdev et al. [9, 10] developed this further into a concept termed ‘poselets’ to detect humans and body parts. The pictorial structures model often used in modeling objects or poses has also been applied to people detection by a number of researchers [11–13]. A key drawback of many of these methods is their use on upright humans in RGB images, whereas our problem area is humans in non-standard poses captured with a depth camera. While human detection has been done using depth data [14–17], these have all been aimed at detecting people from mobile robots and are targeted at that task.

Human detection methods have been aided by work in automating the identification of likely regions to search for people. [18] proposes a method to learn likely segmentations of 3D data using Markov Random Fields. Automated segmentation has also been used extensively in neural network based approaches, most notably the Multiscale Combinatorial Grouping (MCG) technique described in [19] (with background in [20–22]). MCG applies a fast normalized cuts algorithm to an image at different scales, and then effectively merges the cuts to describe likely segmentation candidates. An advantage of this approach is that it provides an exact boundary for an object, unlike the commonly used bounding boxes or the 3D regions described here [23]. Most approaches similar to this are unable to run in real-time however, making them difficult to use in applications where performance is important or computing resources are limited. Zhang et al. proposes a novel 'depth of interest' approach [24], extending regions of interest to 3D space. This approach relies on building a probability distribution of depth values in an image and identifying peaks, corresponding to objects or people in the foreground. A key advantage is that this method is efficient and reduces the amount of data necessary to process afterwards.

2.2 RGB-D Sensors and Skeleton Acquisition

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. Structured-light RGB-D sensors are a type of camera that uses infrared light to capture depth information about a scene, such as the Microsoft Kinect [1], ASUS Xtion PRO LIVE [25], and PrimeSense [26], among others. A structured-light sensor consists of an infrared-light source and a receiver that can detect infrared light. The light projector emits a known pattern, and the way that this pattern distorts on the scene allows the camera to decide the depth. A color camera is also available on the sensor to acquire color frames that can be registered to depth frames, thereby providing color-depth information at each pixel of a frame or 3D color point clouds. Structured-light sensors are inexpensive and can provide 3D skeleton information in real-time. However, since structured-light cameras are based on infrared light, they can only work in an indoor environment. The frame rate (30 Hz) and resolution of depth images (320×240) are also relatively low.

Several drivers are available to provide the access to the color-depth data acquired by the sensor, including the Microsoft Kinect SDK [1], the OpenNI library [27], and the OpenKinect library [28]. The Kinect SDK also provides 3D human skeletal data using the method described by Shotton et al. [2], localizing 20 joints (seen in Figure 2.1(b)). OpenNI uses NITE

[29] – a skeleton generation framework developed as proprietary software by PrimeSense, to generate a similar 3D human skeleton model. This method, available through the Robot Operating System (ROS), provides the locations of 15 joints (seen in Figure 2.1(a)). An alternative 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 [30] contains 24 joints, as depicted in Figure 2.1(c). 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 - it cannot be used on mobile robotic platforms or in real-world sports analysis situations where the markers would inhibit athletic performance.

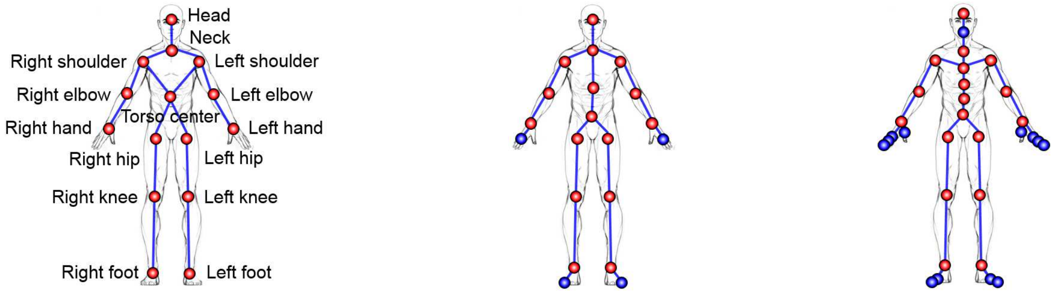
The skeleton is a natural representation of the human body structure, which assumes that the human body is an articulated system of rigid segments that are connected by joints. Acquisition of 3D human skeleton sequences has been a desirable goal for a long time. In addition to NITE and the Kinect SDK, many other skeleton acquisition approaches have been developed that are based on depth imagery. While some are based on manual joint annotation [31–33], the majority of the current methods are based on body part recognition, and then fit a flexible model to the now 'known' body part locations. An alternate main methodology is starting with a 'known' prior, and fitting the silhouette or point cloud to this prior after the humans are localized [14, 15, 34]. A summary of the reviewed skeleton construction techniques is presented in Table 2.1.

Human joint estimation via body part recognition is one popular approach to construct the skeleton model [32, 35, 36, 38–40, 46, 48]. A seminal paper by Shotton et al. [35] in 2011 provided an extremely effective skeleton construction algorithm based on body part recognition, that was able to work in real-time – now available in the Kinect SDK. A single depth image (independent of previous frames) is classified on a per-pixel basis, using a randomized decision forest classifier. Each branch in the forest is determined by a simple

Table 2.1: Summary of skeleton construction techniques based on depth images.

Reference	Approach	Performance Notes
[35],[36]	Pixel-by-pixel classification	16 joints, real-time, 200 fps
[37]	Motion exemplars	38mm accuracy
[38]	Random tree walks	real-time, 1000fps
[39]	Conditional regression forests	over 80% average precision
[40]	Limb-based shape models	robust to occlusions
[32]	Decision tree poselets, pictorial structures	little training data needed
[41]	ICP using optimized Jacobian	over 10 fps
[42]	Matching previous joint positions	20 joints, real-time, 100 fps
[43]	Regression to predict correspondences	19 joints, real-time, 120fps
[44]	ICP on individual parts	10fps, robust to occlusion
[45]	ICP with physical constraints	real-time, 125fps
[46],[47]	Haar features and Bayesian prior	real-time

relation between the target pixel and various others. The pixels that are classified into the same category form the body part, and the joint is inferred by the mean-shift method from a certain body part, using the depth data to 'push' them into the silhouette. While training the decision forests takes a large number of images (around 1 million) as well as a considerable amount of computing power, the fact that the branches in the forest are very simple allows this algorithm to generate 3D human skeleton models within about 5ms per frame. An extended work was published in [36], with both accuracy and speed improved. Plagemann et al. [46] introduced an approach to recognize body parts using Haar features [49] and construct a skeleton model on these parts. Using data over time, they construct a Bayesian network, which produces the estimated pose using body part locations and starts with the previous pose as a prior [47]. Holt et al. [32] proposed Connected Poselets to estimate 3D human pose from depth data. The approach utilizes the idea of poselets [9], which are widely applied for pose estimation from RGB images. For each depth image, a multi-scale sliding window is applied, and a decision forest is used to detect poselets and estimate joint locations. Using a skeleton prior inspired by pictorial structures [11, 50], the method begins with a torso point and connects outwards to body parts. By applying kinematic inference



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

Figure 2.1: Skeleton diagrams from OpenNI, Kinect, and MoCap. Skeleton data acquired from OpenNI contains 15 joints as depicted in Figure 2.1(a), 20 joints from Microsoft Kinect SDK as shown in Figure 2.1(b), and a varied number of joints from a MoCap system such as 31 joints in Figure 2.1(c).

to eliminate impossible poses, they are able to reject incorrect body part classifications and improve their accuracy.

Another widely investigated methodology to construct 3D human skeleton models from depth imagery is based on nearest-neighbor matching [37, 41–44, 51]. Several approaches for whole-skeleton matching are based on the Iterative Closest Point (ICP) method [52], which can iteratively decide a rigid transformation such that the input query points fit to the points in the given model under this transformation. Using point clouds of a person with known poses as a model, several approaches [41, 44] apply ICP to fit the unknown poses by estimating the translation and rotation to fit the unknown body parts to the known model. While these approaches are relatively accurate, they suffer from several drawbacks. ICP is computationally expensive for a model with as many degrees of freedom as a human body. Additionally, it can be difficult to recover from tracking loss. Typically the previous pose is used as the known pose to fit to; if tracking loss occurs and this pose becomes inaccurate, then further fitting can be difficult or impossible. Finally, skeleton construction methods based on the ICP algorithm generally require an initial T-pose to start the iterative process.

2.3 Skeleton Representation Modalities

Skeleton-based human representations are constructed from various features computed from raw 3D skeletal data, where each feature source is called a *modality*, a standard term used in multi-view learning [53]. From the perspective of information modality, 3D skeleton-based human representations can be classified into four categories, in terms of whether they are based on joint displacement, orientation, raw position, or multiple modalities.

2.3.1 Representations Based on Displacement

Features extracted from displacements of skeletal joints are widely applied in many skeleton-based representations due to their simple structure and easy implementation. They can be based either on the displacement between different human joints within the same frame or the displacement of the same joint across different time periods.

2.3.1.1 Spatial Displacement Between Joints

Representations based on relative joint displacements compute spatial displacements of coordinates of human skeletal joints in 3D space, which are acquired from the same frame at a time point.

The pairwise relative position of human skeleton joints is the most widely studied displacement feature for human representation [54–59]. Within a skeleton model obtained at a single time point, for each joint $\boldsymbol{p} = (x, y, z)$ in 3D space, the difference between the location of joint i and joint j is calculated by $\boldsymbol{p}_{ij} = \boldsymbol{p}_i - \boldsymbol{p}_j, i \neq j$. The joint locations $\boldsymbol{p}$ are often normalized, so that the feature is invariant to the absolute body position, initial body orientation and body size [54–56]. Chen and Koskela [57] implemented a method based on pairwise relative position of skeleton joints with normalization calculated by $\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|}{\sum_{i \neq j} \|\boldsymbol{p}_i - \boldsymbol{p}_j\|}, i \neq j$, which is illustrated in Figure 2.2(a).

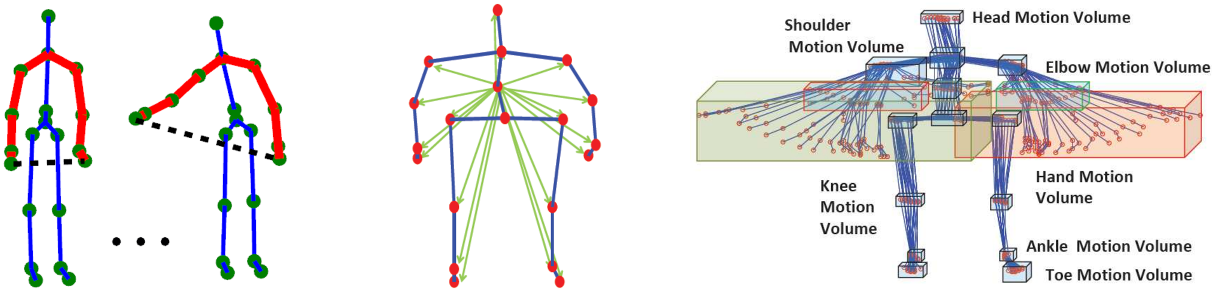
Another group of joint displacement features extracted from the same frame are based on the difference to a reference joint. In these features, the displacements are obtained by calculating the coordinate difference of all joints with respect to a single joint, usually manually selected. Given the location of a joint (x, y, z) and a given reference joint (x_c, y_c, z_c) in the world coordinate system, Rahmani et al. [60] defined the spatial joint displacement as $(\Delta x, \Delta y, \Delta z) = (x, y, z) - (x_c, y_c, z_c)$, where the reference joint can be the skeleton centroid or a manually selected, fixed joint (seen in Figure 2.2(b)). For each sequence of human skeletons representing an activity, the computed displacements along each dimension (e.g., Δx, Δy or Δz) are used as features to represent humans. Luo et al. [61] performed similar calculations but selected the hip center as the reference joint since it has relatively small motions for most actions.

2.3.1.2 Temporal Joint Displacement

3D human representations based on temporal joint displacements compute the location difference across a sequence of frames acquired at different time points. Usually, they employ both spatial and temporal information to represent people in space and time.

A widely used temporal displacement feature is implemented by comparing the joint coordinates at different time steps. Yang and Tian [56, 62] introduced a novel feature based on the position difference of joints, called EigenJoints, which combines static posture, motion, and joint offsets. In particular, the joint displacement of the current frame with respect to the previous frame and initial frame is calculated. Ellis et al. [63] introduced an algorithm to reduce latency for action recognition using a 3D skeleton-based representation that depends on spatio-temporal features computed from the information in three frames: the current frame, the frame collected 10 time steps ago, and the frame collected 30 steps ago. Then, the features are computed as the temporal displacement among those three frames. Another approach to construct temporal displacement representations incorporates the object being interacted with in each pose [64]. This approach constructs a hierarchical graph to represent positions in 3D space and motion through 1D time. The differences of joint coordinates in two successive frames are defined as the features.

The joint movement volume is another feature construction approach for human representation that also uses joint displacement information for feature extraction, especially when a joint exhibits a large movement [60]. For a given joint, extreme positions during the full joint motion are computed along x, y, and z axes. The maximum moving range of each joint along each dimension is then computed by $L_a = \max(a_j) - \min(a_j)$, where $a = x, y, z$; and the joint volume is defined as $V_j = L_x L_y L_z$, as demonstrated in Figure 2.2(c). For each joint, L_x, L_y, L_z and V_j are flattened into a feature vector. The approach also incorporates relative joint displacements with respect to the torso joint into the feature.



(a) Pairwise Displacement (b) Relative Displacement (c) Joint Motion Volume

Figure 2.2: Examples of 3D human representations based on joint displacements. Figure 2.2(a) illustrates the displacement of pairwise joints [57], Figure 2.2(b) shows relative joint displacement and Figure 2.2(c) illustrates joint motion volume features [60].

2.3.2 Representations Based on Orientation

Another widely used information modality for human representation construction is joint orientations, since in general orientation-based features are invariant to human position, body size, and orientation to the camera.

2.3.2.1 Spatial Orientation of Pairwise Joints

Approaches based on spatial orientations of pairwise joints compute the orientation of displacement vectors for pairs of human skeletal joints acquired at the same time step.

A number of popular orientation-based representations compute the orientation of each joint to the human centroid in 3D space. For example, Gu et al. [65] collected skeleton data with fifteen joints and extracted features representing joint angles with respect to the person's torso. Sung et al. [66] computed the orientation matrix of each human joint with respect to the camera, and then transformed the joint rotation matrix to obtain the joint orientation with respect to the person's torso. A similar approach was also introduced in [67] based on the orientation matrix. Xia et al. [68] introduced Histograms of 3D Joint Locations (HOJ3D) features by assigning 3D joint positions into cone bins in 3D space. Twelve key joints are selected and their orientation are computed with respect to the center torso point. Using linear discriminant analysis (LDA), the features are reprojected to extract the dominant ones. Since the spherical coordinate system used in [68] is oriented with the x axis aligned with the direction a person is facing, their approach is view invariant.

Another approach is to calculate the orientation of two joints, called relative joint orientations. Jin and Choi [69] utilized vector orientations from one joint to another joint, named the first order orientation vector, to construct 3D human representations. The approach also proposed a second order neighborhood that connects adjacent vectors. The authors used a uniform quantization method to convert the continuous orientations into eight discrete symbols to guarantee robustness to noise. Zhang and Tian [70] used a two mode 3D skeleton representation, combining structural data with motion data. The structural data is represented by pairwise features, relating the positions of each pair of joints relative to each other. The orientation between two joints i and j was also used, which is given by $\theta(i,j) = \arcsin\left(\frac{i_x - j_x}{dist(i,j)}\right)/2\pi$, where $dist(i,j)$ denotes the geometric distance between two joints i and j in 3D space.

2.3.2.2 Temporal Joint Orientation

Human representations based on temporal joint orientations usually compute the difference between orientations of the same joint across a temporal sequence of frames. Boubou and Suzuki [71] describe a representation called Histogram of Oriented Velocity Vectors (HOVV), which is a histogram of the velocity orientations computed from 19 human joints in a skeleton kinematic model acquired from a Kinect camera. Each temporal displacement vector is described by its orientation in 3D space as the joint moves from the previous position to the current location. By using a static skeleton prior to deal with static poses with little or no movement, this method is able to effectively represent humans with still poses in 3D space in human action recognition applications.

2.3.3 Representations Based on Raw Joint Positions

Besides joint displacements and orientations, raw joint positions directly obtained from sensors are also used by many methods to construct space-time 3D human representations.

A category of approaches flatten joint positions acquired in the same frame into a column vector. Given a sequence of skeleton frames, a matrix can be formed to naïvely encode the sequence with each column containing the flattened joint coordinates obtained at a specific time point. Following this direction, Hussein et al. [72] computed the statistical Covariance of 3D Joints (Cov3DJ) as their features, as illustrated in Figure 2.3. Specifically, given K human joints with each joint denoted by $\boldsymbol{p}_i = (x_i, y_i, z_i), i = 1, \dots, K$, a feature vector is formed to encode the skeleton acquired at time t: $\boldsymbol{S}^{(t)} = [x_1^{(t)}, \dots, x_K^{(t)}, y_1^{(t)}, \dots, y_K^{(t)}, z_1^{(t)}, \dots, z_K^{(t)}]^\top$. Given a temporal sequence of T skeleton frames, the Cov3DJ feature is computed by $C(\boldsymbol{S}) = \frac{1}{T-1} \sum_{t=1}^{T} (\boldsymbol{S}^{(t)} - \bar{\boldsymbol{S}}^{(t)})(\boldsymbol{S}^{(t)} - \bar{\boldsymbol{S}}^{(t)})^\top$, where $\bar{\boldsymbol{S}}$ is the mean of all $\boldsymbol{S}$.

Some representations on basic features consisting of just joint positions, but focused on intelligently selecting which joints to use. Since not all the joints are equally informative, several methods were proposed to select key joints that are more descriptive [73–76]. Chaaraoui et al. [73] introduced an evolutionary algorithm to select a subset of skeleton joints to form features. Then a normalizing process was used to achieve position, scale and rotation invariance. Similarly, Reyes et al. [74] selected 14 joints in 3D human skeleton models without normalization for feature extraction in gesture recognition applications.

Similar to the application of deep learning techniques to extract features from images where raw pixels are typically used as input, skeleton-based human representations built by deep learning methods generally rely on raw joint position information. For example, Du et al. [77] proposed an end-to-end hierarchical recurrent neural network (RNN) for the skeleton-based representation construction, in which the raw positions of human joints are directly used as the input to the RNN. Zhu et al. [78] used raw 3D joint coordinates as

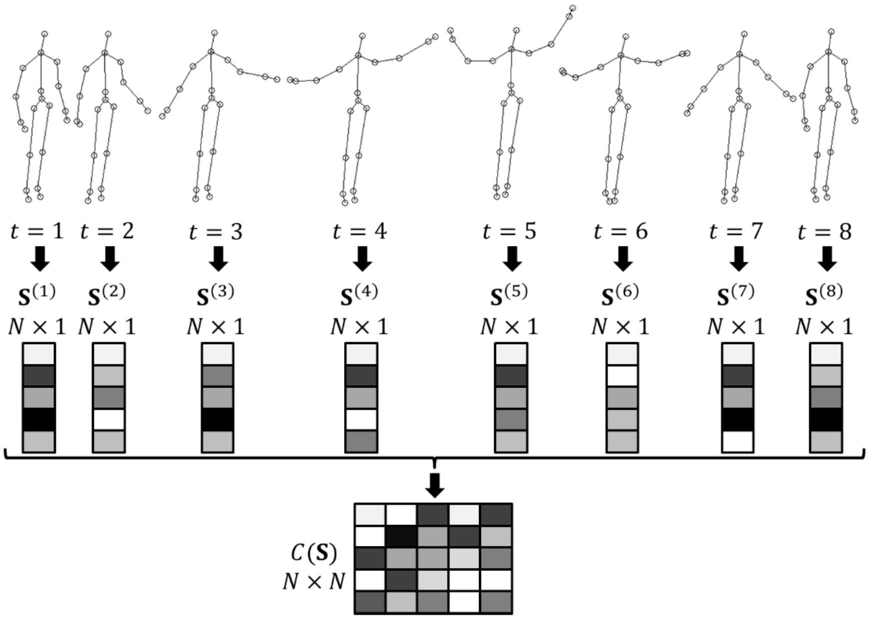

$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 5$
$t = 6$
$t = 7$
$t = 8$
$\mathbf{S}^{(1)}$ $N \times 1$
$\mathbf{S}^{(2)}$ $N \times 1$
$\mathbf{S}^{(3)}$ $N \times 1$
$\mathbf{S}^{(4)}$ $N \times 1$
$\mathbf{S}^{(5)}$ $N \times 1$
$\mathbf{S}^{(6)}$ $N \times 1$
$\mathbf{S}^{(7)}$ $N \times 1$
$\mathbf{S}^{(8)}$ $N \times 1$
$C(\mathbf{S})$
$N \times N$

Figure 2.3: 3D human representation based on the Cov3DJ descriptor [72].

the input to a RNN with Long Short-Term Memory (LSTM) to automatically learn human representations.

2.3.4 Representations Based on Multiple Modalities

Since multiple information modalities are available, an intuitive way to improve the descriptive power of a human representation is to integrate multiple information sources and build a multi-modal representation to encode humans in 3D space. For example, the spatial joint displacement and orientation can be integrated together to build human representations. Guerra-Filho and Aloimonos [79] proposed a method that maps 3D skeletal joints to 2D points in the projection plane of the camera and computes joint displacements and orientations of the 2D joints in the projected plane. Gowayyed et al. [80] developed the histogram of oriented displacements (HOD) representation that computes the orientation of temporal joint displacement vectors and uses their magnitude as the weight to update the histogram in order to make the representation speed-invariant. Yu et al. [81] integrated three types of features to construct a spatio-temporal representation, including pairwise joint distances, spatial joint coordinates, and temporal variations of joint locations. Masood et al. [82] implemented a similar representation by incorporating both pairwise joint distances and temporal joint location variations. Zanfir et al. [83] introduced a feature that integrates raw 3D joint positions as well as first and second derivatives of the joint trajectories, based on the assumption that the speed and acceleration of human joint motions can be described accurately by quadratic functions.

2.3.5 Summary

Through computing the difference of skeletal joint positions in 3D real-world space, displacement-based representations are invariant to absolute locations and orientations of people with respect to the camera, which can provide the benefit of forming view-invariant spatio-temporal human representations. Similarly, orientation-based human representations can provide the same view-invariance because they are also based on the relative information between human joints. In addition, since orientation-based representations do not rely on the displacement magnitude, they are usually invariant to human scale variations. Representations based directly on raw joint positions are widely used due to the simple acquisition from sensors. Although normalization procedures can make human representations partially invariant to view and scale variations, more sophisticated construction techniques (e.g., deep learning) are typically needed to develop robust human representations.

Representations that do not involve temporal information are suitable to address problems such as pose estimation and gesture recognition. However, if we want the representations to be capable of encoding dynamic human motions, integrating temporal information is helpful. Applications such as activity recognition can benefit from spatio-temporal representations that incorporate time and space information simultaneously.

2.4 Skeleton Representation Encodings

Feature encoding is a necessary and important component in representation construction [84], which aims at integrating all extracted features together into a final feature vector that can be used as the input to classifiers or other reasoning systems. In the scenario of 3D skeleton-based representation construction, the encoding methods can be broadly grouped into three classes: concatenation-based encoding, statistics-based encoding, and bag-of-words encoding.

2.4.1 Concatenation-Based Encoding

Many methods directly use extracted skeleton-based features, such as displacements and orientations of 3D human joints, and concatenate them into a 1D feature vector to build a human representation [55–57, 62, 64, 66, 74, 75, 81, 85–91]. For example, Fothergill et al. [90] encoded a feature vector by concatenating 35 skeletal joint angles, 35 joint angle velocities, and 60 joint velocities into a 130-dimensional vector at each frame. Then, feature vectors from a sequence of frames are further concatenated into a large final feature vector that is fed into a classifier for reasoning. Similarly, Gong et al. [88] directly concatenated 3D joint positions into a 1D vector as a representation at each frame to address the time series segmentation problem.

2.4.2 Statistics-Based Encoding

Statistics-based encoding is a common and effective method to incorporate all features into a final feature vector, without applying any feature quantization procedure. This encoding methodology processes and organizes features through simple statistics. For example, the Cov3DJ representation [72], as illustrated in Figure 2.3, computes the covariance of a set of 3D joint position vectors collected across a sequence of skeleton frames. Since a covariance matrix is symmetric, only upper triangle values are utilized to form the final feature. An advantage of this statistics-based encoding approach is that the size of the final feature vector is independent of the number of frames.

The most widely used statistics-based encoding methodology is histogram encoding, which uses a 1D histogram to estimate the distribution of extracted skeleton-based features. For example, Xia et al. [68] partitioned the 3D space into a number of bins using a modified spherical coordinate system and counted the number of joints falling in each bin to form a 1D histogram, which is called the Histogram of 3D Joint Positions (HOJ3D). A large number of skeleton-based human representations using similar histogram encoding methods have also been introduced, including Histogram of Joint Position Differences (HJPD)[60], Histogram of Oriented Velocity Vectors (HOVV)[71], and Histogram of Oriented Displacements (HOD)[80], among others [70, 76, 92–95]. When multi-modal skeleton-based features are involved, concatenation-based encoding is usually employed to incorporate multiple histograms into a single final feature vector [95].

2.4.3 Bag-of-Words Encoding

Unlike concatenation and statistics-based encoding methodologies, bag-of-words encoding applies a coding operator to project each high-dimensional feature vector into a single code (or word) using a learned codebook (or dictionary) that contains all possible codes. This procedure is also referred to as feature quantization. Given a new instance, this encoding methodology uses the normalized frequency of code occurrence as the final feature vector. Bag-of-words encoding is widely employed by a large number of skeleton-based human representations [59, 61, 65, 73, 83, 96–112]. According to how the dictionary is learned, the encoding methods can be broadly categorized into two groups, based on clustering or sparse coding.

The k-means clustering algorithm is a popular unsupervised learning method that is commonly used to construct a dictionary. Wang et al. [109] grouped human joints into five body parts, and used the k-means algorithm to cluster the training data. The indices of the cluster centroids are utilized as codes to form a dictionary. During testing, query body part poses are quantized using the learned dictionary. Similarly, Kapsouras and Nikolaidis [110] used the k-means clustering method on skeleton-based features consisting of joint orientations and orientation differences in multiple temporal scales, in order to select representative patterns to build a dictionary.

Sparse coding is another common approach to construct efficient representations of data as a (often linear) combination of a set of distinctive patterns (i.e., codes) learned from the data itself. Zhao et al. [59] introduced a sparse coding approach regularized by the $l_{2,1}$ norm to construct a dictionary of templates from the so-called Structured Streaming Skeletons (SSS) features in a gesture recognition application. Luo et al. [61] proposed another sparse coding method to learn a dictionary based on pairwise joint displacement features. This approach uses a combination of group sparsity and geometric constraints to select sparse and more representative patterns as codes. An illustration of the dictionary learning method to encode skeleton-based human representations is presented in Figure 2.4.

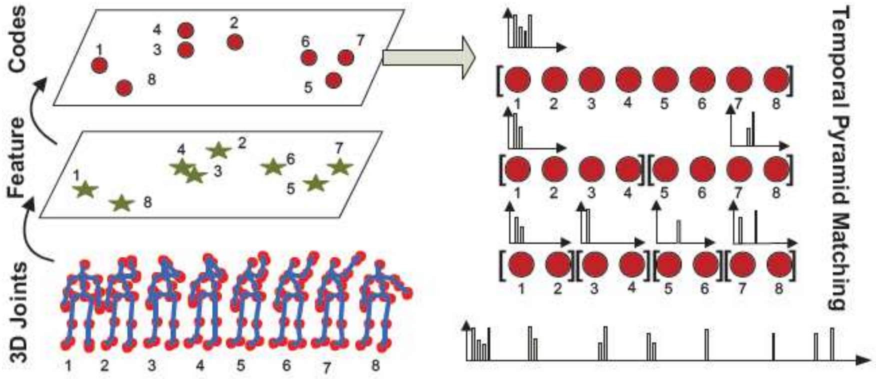


Figure 2.4: Dictionary learning based on sparse coding for human representation [61].

2.4.4 Summary

Due to its simplicity and high efficiency, the concatenation-based feature vector construction method is widely applied in real-time online applications to reduce processing latency. The method is also used to integrate features from multiple sources into a single vector for further encoding/processing. By not requiring a feature quantization process, statistics-based encoding, especially based on histograms, is efficient and relatively robust to noise. However, this method is incapable of identifying the representative patterns and modeling the structure of the data, thus making it lacking in discriminative power. Bag-of-words encoding can encode a feature vector using a sparse solution to minimize approximation error, and is also validated to be robust to data noise. However, dictionary construction and feature quantization require additional computation.

2.5 Activity Classification and Prediction

Activity recognition in most research consists of classification of single activity sequences, as opposed to temporal segmentation of mixed sequences. Many of these approaches [113–118] are based on histograms of skeletal features, manually crafted in a variety of ways. These feature histograms are built over the length of an action sequence and used to train multi-class Support Vector Machines (SVM). Future action sequences can then be classified. Other approaches train SVMs using other representations, such as building models of joint trajectories for specific actions [72, 107, 119–122], and similar approaches model joint angles involved in an action [123]. Approaches have also been developed to learn a representation of an action using parametric networks [124] or random forests [125, 126], instead of defining a manual representation.

A more difficult problem, and what I address in my gymnastics analysis system, is segmenting a video sequence that consists of multiple actions, and determining which frames depict a particular action. A number of approaches have been applied to this problem: Markov chains [127], dividing sequences into blocks with 'fuzzy' boundaries [128], k-means clustering of segments [129], and classifying frames individually [130]. A similar line of research is modeling sub-sequences within a video, without event annotations [131–133]. These works divide a video into activities, without having knowledge of which specific activities are being performed.

Different from conventional action classification [134, 135], 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 [136]. A max-margin early event detector was implemented in [13] to detect early events. Logistic regression models [137] were employed to detect starting points of human activities. An online Conditional Random Field method was introduced in [138] to predict human intentions in human-robot collaboration applications. In [120], 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, [139] demonstrated a system for recognizing actions based on a single frame, but only showed successful results for very simple gestures. [140] 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 features (or the absence of these features) during an activity; essentially, 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. Only a few approaches have been implemented at the representation level. For example, [141] represented 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 [142–148], with more reviewed in [149, 150]). Similar work was done in [151], 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 [152]; 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 [153, 154]. 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 [155], 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 [156] 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 [157] 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.

2.6 Sports Analysis and Gymnastics

Analysis of coaches in various sports show that they value data-based analysis done by sports scientists [158, 159] but that currently gaps exist in transferring this analytical knowledge to those actually involved in coaching. Current vision based systems also require too much operator input, something that can be difficult for non-technical users [160]. Additionally, research often focuses on biomechanical principles at the expense of technical analysis with the aim of improving performance [161].

In gymnastics, some work has been done analyzing the biomechanics involved in the sport; [162] reviews biomechanical research into artistic gymnastics, aiming to identify the variables that contribute to success in the sport. Research has been done to analyze the vault event [163] and the parallel bar event [164]. For artistic gymnastics, camera based systems were shown to significantly increase the accuracy of judging [165]. Pommel horse circles have been focused on in a few studies, but nearly exclusively with the goal being biomechanical analysis and not for performance evaluation or improvement. [166] studied the pattern of wrist impacts during pommel horse routines, noting that smoother routines are less likely to cause impact forces leading to injury. [167] conducted an analysis using force sensors built beneath a pommel horse, focusing specifically on the velocity differences between single-hand and double-hand circles. Their work continued in [168, 169], where they analyzed pommel horse circles using a suspended aid. While this work doesn't attempt to provide a qualitative assessment as ours does, they did show that the use of an aid caused gymnasts to perform circles with greater diameter and smoother hand movement. One notable study that was aimed at performance improvement showed that feedback in the form of video review and quantitative analysis (much like the visual feedback and automated analysis our system produces) results in improved performance versus that of a control group [170].

CHAPTER 3
AUTOMATED EVALUATION OF GYMNAST PERFORMANCE

Sports analysis is a useful application of technology, providing value to athletes, coaches, and sports fans by producing quantitative evaluations of performance. To address this field in the context of men's gymnastics, I introduce a system that utilizes a Microsoft Kinect 2 camera to automatically evaluate the performance of a gymnast on the pommel horse apparatus, specifically in regards to the consistency of the gymnast's timing. The Kinect's ability to determine the depth at each pixel provides information not available to typical sports analysis approaches based solely on RGB data. My approach consists of a three stage pipeline that automatically identifies a depth of interest, localizes the gymnast, detects when the gymnast is performing a certain routine, and finally provides an analysis of that routine. I demonstrate that each stage of the pipeline produces effective results: my depth of interest approach identifies the gymnast 97.8% of the time and removes over 60% of extraneous data; my activity recognition approach is highly efficient and identifies 'spinning' by the gymnast with 93.8% accuracy; and my performance analysis method evaluates the gymnast's timing with accuracy only limited by the frame rate of the Kinect. Additionally, I validate my system and the proposed methods with a real-world online application, used by actual gymnastics coaches and viewed as a highly effective training tool.

3.1 Motivation

Sports are a phenomenon common to all cultures, and the popular interest in them naturally causes people to wonder how athletes perform at high levels, and more importantly, how those athletes can perform even better. Many sports, such as skiing and cycling, have benefited from performance analysis. Automated analysis in alpine skiing increased athletes' performance by identifying techniques to extend their glide, reducing the energy necessary to maintain a competitive speed [171]. Similarly, in cycling analysis showed which movements were necessary and which were wasted on extended rides [172]. In addition, sports analysis provides an interesting perspective to fans and viewers of a sport. Television shows such as ESPN's 'Sports Science' [173] capitalize on this, using video to highlight the impressive abilities of popular athletes.

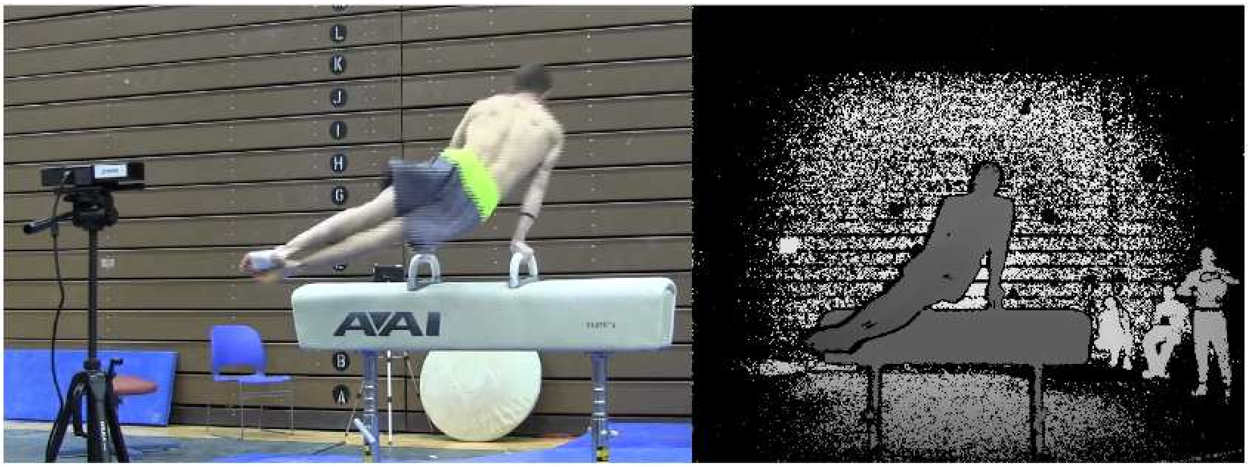


Figure 3.1: Pommel horse with Kinect camera. The pommel horse is a men's gymnastics event where spin and body angle consistency is important for scoring. I used a Kinect 2 [174] to record performances and create our analysis system. An example image from the Kinect can be seen on the right.

One of the main challenges in the analysis of gymnastics, and sports in general, is the limited amount of data available. Because it's difficult to collect this data in a lab environment, and because access to high-level athletes is limited, sports datasets are rare. Limited examples can be found in datasets such as the Carnegie Mellon motion capture dataset [175] and the Multi-View TJU [176] depth-based dataset, but these are often more general physical activities as opposed to the actions of high-level athletes. In addition, in order to perform a useful analysis of a sports dataset, researchers must have knowledge of what an ideal result would look like. With objective sports such as football and basketball, this can be relatively simple - statisticians routinely calculate the expected values of decisions, e.g. in football whether it is better to 'go for it' on 4th and short versus conventionally punting the ball to the other team [177]. However, subjective sports such as gymnastics, figure skating, and many others that are scored by judges make this much more difficult. Researchers need knowledge of what these judges look for and what differentiates good performances from bad ones. Additionally, analyzing many gymnastics events is difficult with a 2D camera. Much of a performance happens in three dimensions, unlike a sport like football that can be effectively analyzed with an overhead camera.

3.2 Gymnastics Dataset

Working with an elite gymnastics facility, I constructed a dataset consisting of gymnasts performing on the pommel horse apparatus.

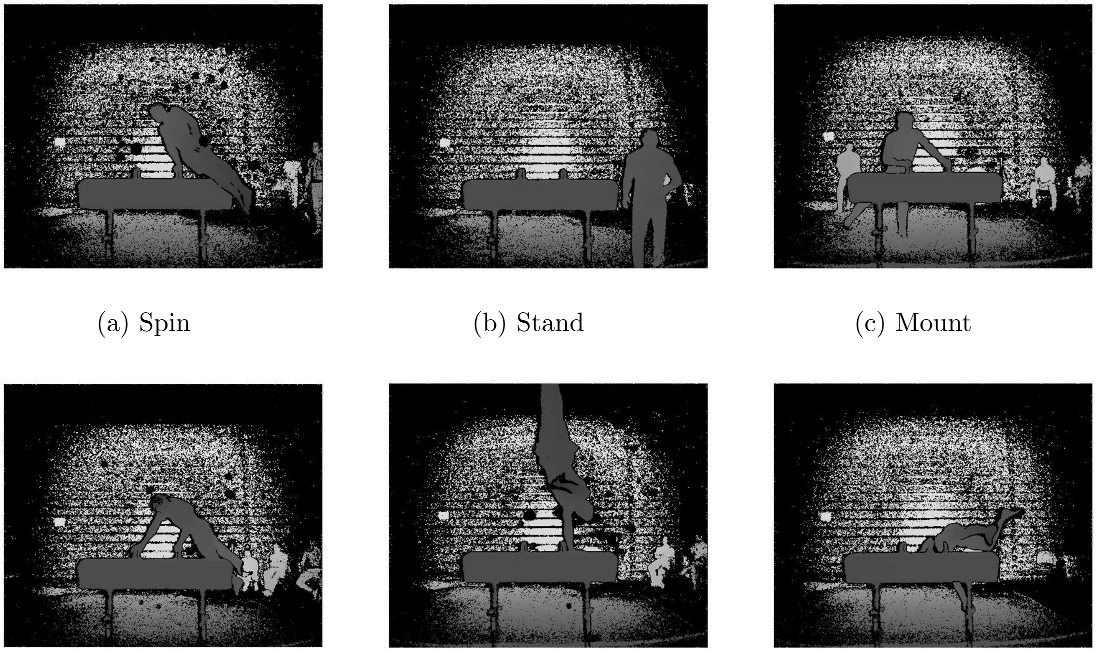
3.2.1 Recording the Dataset

As I noted in Section 3.1, no extensive dataset exists that consists of depth imagery of high level athletics, much less gymnastics or the pommel horse event specifically. Working with a gymnastics training facility, I collected a dataset consisting of male gymnasts performing on the pommel horse, seen in Figure 3.2. I believe that this dataset is a unique contribution and will be of great to use to researchers working on vision-based analysis of many sports.

The dataset was recorded with the setup seen in Figure 3.1. The Kinect 2 camera was placed in front of the pommel horse, and recordings were made at a variety of different distances.

3.2.2 Details and Annotation

The dataset consists of 10115 frames of gymnasts performing on the pommel horse, recorded as 16-bit PNG images. These images are organized into 39 routines, each of which begins with the gymnast mounting the pommel horse and ends with the gymnast dismounting the pommel horse. Some routines involve falls or dismounts during the routine, but in each case the gymnast re-mounts the apparatus. Overall, the dataset contains a large variety of situations: spinning (Figure 3.2(a)), standing (Figure 3.2(b)), mounting (Figure 3.2(c)), dismounting (Figure 3.2(d)), miscellaneous moves such as handstands (Figure 3.2(e)), and falls (Figure 3.2(f)).



(a) Spin

(b) Stand

(c) Mount

(d) Dismount

(e) Handstand

(f) Fall

Figure 3.2: A variety of dataset image examples. Figure 3.2(a) depicts spinning; Figure 3.2(b) is a gymnast standing by the pommel horse; Figure 3.2(c) is a gymnast mounting the pommel horse while Figure 3.2(d) is a gymnast dismounting; Figure 3.2(e) shows an example of the variety of movements performed (in this case a handstand); and Figure 3.2(f) is a gymnast falling from the pommel horse. Only Figure 3.2(a) is labeled 'spinning'; all others are labeled 'not spinning'.

The dataset was annotated to assign each frame an activity of 'spinning' (performing circles) or 'not spinning' (e.g. mounting the pommel horse, dismounting, and other moves such as scissors). In total, there are 6405 frames annotated as spinning and 3710 frames as not spinning. Of the frames in which the gymnast is spinning, 2231 frames were annotated with the locations of the head and feet of the gymnast. These positions at the left and right extrema of a spin were interpolated using a cubic spline (described in Section 3.3.3) to determine an exact ground truth timestamp and frame number for each extrema.

3.3 Approach

My approach to this problem is a three stage pipeline, illustrated in Figure 3.3, beginning with a depth image obtained from a Kinect 2. This depth image is processed to identify likely depths of interest where the gymnast and pommel horse would be located in the scene. Identified depths of interest are the input to a HOG based detector trained to identify silhouettes of gymnasts. The localized silhouette is then used as input to the activity recognition classifier, which determines whether the gymnast is spinning or not. As current skeleton construction approaches such as the algorithm used in the Kinect SDK [2] do not provide accurate joint locations for unusual poses such as a gymnast in the middle of a performance, I define a new feature representation for activity representation based on the 3D information present in a depth silhouette. If the gymnast is spinning, his performance is analyzed in order to determine the speed of his spins and the angle of his legs. These values are used to qualitatively evaluate the performance, and are also available as training data for gymnastics coaches.

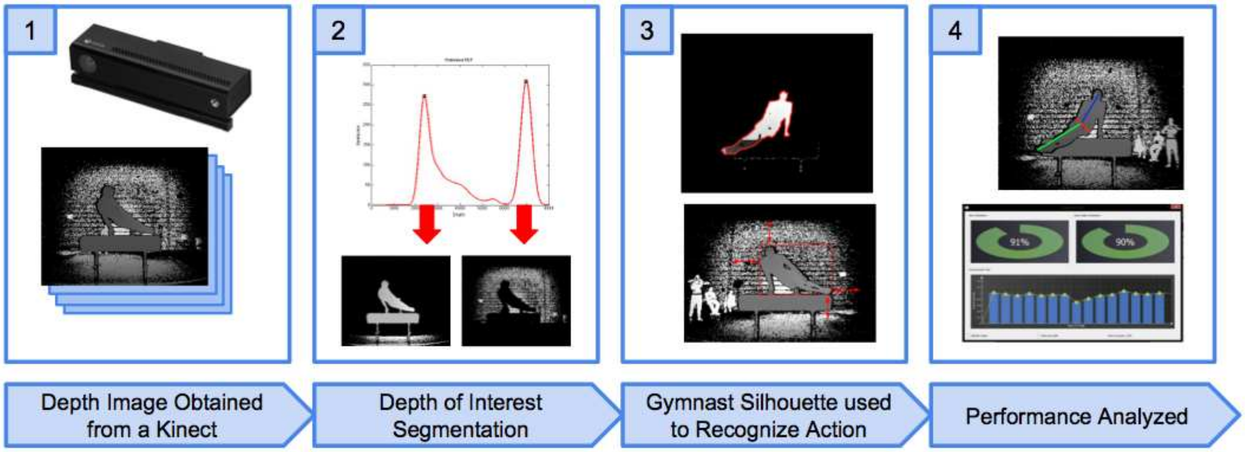

1
2
3
4
91%
90%
Depth Image Obtained from a Kinect
Depth of Interest Segmentation
Gymnast Silhouette used to Recognize Action
Performance Analyzed

Figure 3.3: Visualization of the gymnastics analysis pipeline. The system 1) begins with a depth image stream from a camera such as a Kinect, 2) identifies depths of interest likely to contain the gymnast, 3) uses this identified gymnast silhouette to recognize when the gymnast is performing circles, and 4) produces an analysis of the gymnast's performance.

3.3.1 Depth of Interest Segmentation and Gymnast Identification

My approach to segmentation is based heavily on the method described by [24], which aims to locate depth ranges in the image which are likely to contain objects or people of interest. I start with the depth image from the Kinect, instead of a point cloud, and randomly select a sample of non-noise pixels n (e.g., 1000). Using a Parzen window [178] based approach, we can estimate a probability distribution function (PDF) for the depth values in the entire frame. Each of the n sampled pixels is described with a Gaussian kernel, whose width is based on the maximum depth in the frame, where x is the range of possible depth values and $D(i)$ returns the depth at index i.

$$PDF = \sum_{i=1}^{n} \exp \frac{-(x - D(i))^2}{2 \times MAXDEPTH} \tag{3.1}$$

I then identify the three highest peaks in this distribution (see Figure 3.4). These peaks identify depths where we should focus our segmentation. Each peak is padded with a neighborhood of 10% of the maximum depth (i.e. 800mm for the 8m maximum depth of the Kinect 2) on either side of the peak. A sample of these segmented neighborhoods can be seen in Figure 3.5.

The aim of the proposed 'depth of interest' (DOI) segmentation is to remove areas of the image which are not likely to be relevant, in order to reduce the load on the human detection process, and provide a cleaner image for later processing. Since the overall approach operates on depth imagery, if the target depth is known it is trivial to remove the background and foreground noise.

The DOI provides highly efficient segmentation proposals. Each DOI (e.g., 'DOI 2' in Figure 3.5) is processed by a human detector based on HOG, which operates only on the depth range identified. This detector is trained by randomly hand sampling a large variety of gymnast poses and using these to train a SVM that will recognize a silhouette.

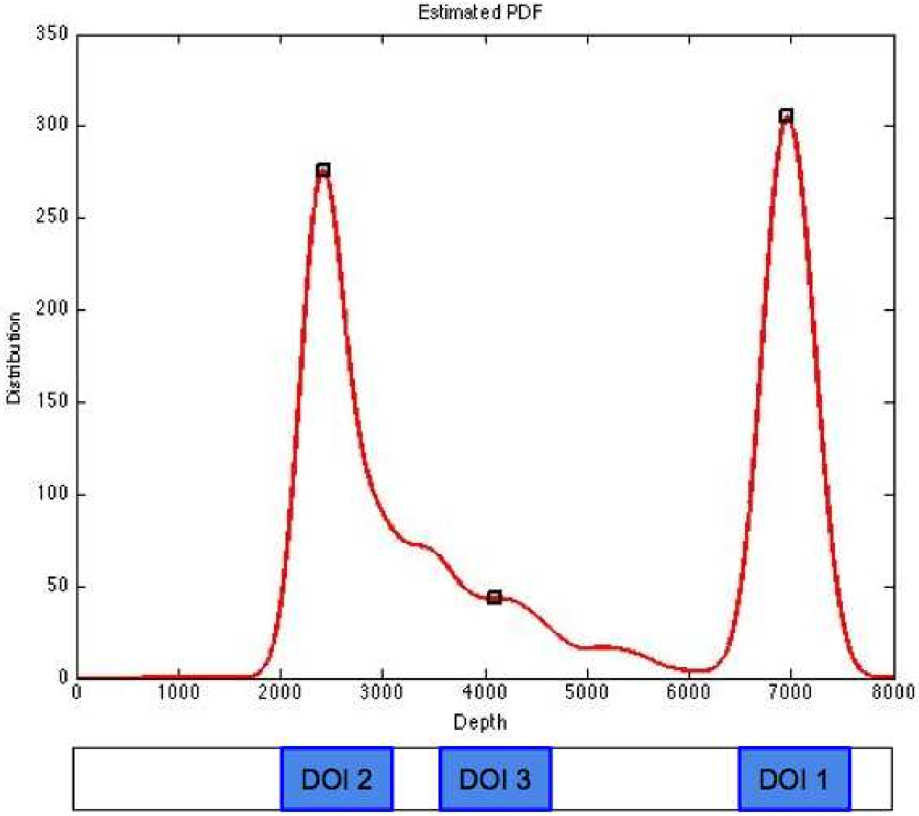


Figure 3.4: Estimated probability distribution for the depth values in a frame. The three identified depths of interest are marked at the bottom, corresponding to the three largest peaks of the distribution.

3.3.2 Activity Classification

The next stage of the approach is recognizing the activity the gymnast is performing and segmenting the video based on this. Since the performance analysis method is focused on a specific portion of the gymnast's routine (spinning / performing circles), it will generate noisy data if applied to a different situation, such as the gymnast mounting the pommel horse. In order to ensure an accurate analysis is generated, it is vital to know when the gymnast is spinning.

Many activity recognition approaches are based on skeleton joint data. Since these approaches are inadequate to analyze situations like a gymnastics performance (see Figure 3.6), I have defined a new activity recognition approach based on silhouette data. Additionally, as

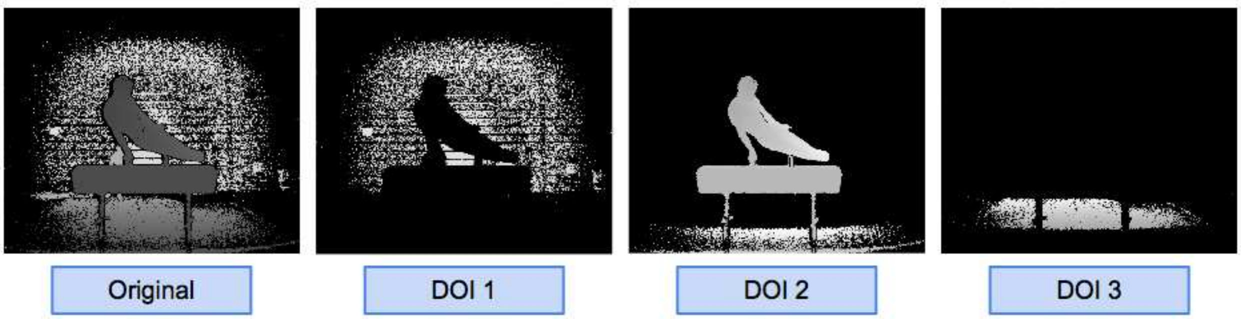


Figure 3.5: Example depth of interest segmentations for the probability distribution in Figure Figure 3.4.

we do not know how long an action sequence will take or when a particular action will occur, methods assuming that the start and end points of an event are known, such as histogram-based methods commonly seen in activity recognition works, are difficult to apply. Instead, each frame will be classified individually, similar to [88]. We describe a gymnast's action with a Silhouette Activity Descriptor in 3 Dimensions (SAD3D). The gymnast in each frame is described by these features: 1) the width of their silhouette, 2) the height of their silhouette, 3-4) the depth values at the leftmost and rightmost ends of the silhouette, and 5-8) the shift in the left-most x, right-most x, upper y, and lower y coordinates compared to the previous frame. A graphical description of SAD3D can be seen in Figure 3.7. As described in the first stage of the pipeline, each frame in the dataset was automatically segmented to identify the gymnast, and the identified silhouette was processed to obtain the described feature vector. Additionally, each performance segment in the dataset was hand segmented to identify whether the gymnast was spinning or not in each frame. This combination of data was used to train a SVM that will identify whether a given frame depicted a spinning gymnast or not. For two SAD3D features, x_i and x_j, we used a radial basis function kernel:

$$K(x_i, x_j) = \exp -\gamma \|x_i - x_j\|^2 \tag{3.2}$$

After classification by an SVM, we apply a smoothing technique that takes advantage of the fact that activities are longer than a single frame in nearly every application. Where c_i is

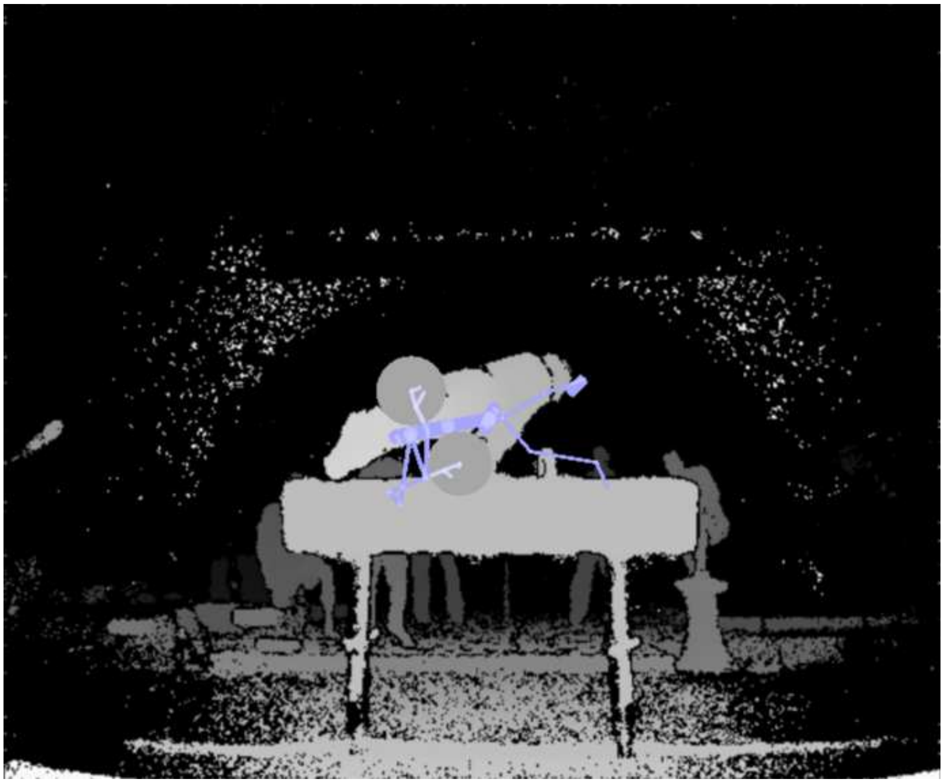


Figure 3.6: Kinect skeleton construction error. The Kinect SDK attempts to construct a skeleton for human forms. However, since it is trained on upright poses, it generates noisy and inaccurate data when applied to gymnasts, such as this wildly unrealistic skeleton.

a binary class label, a frame being considered in a neighborhood of 5 frames is adjusted as such:

$$c_i = \lfloor \frac{1}{5} \sum_{j=-2}^{2} c_{i+j} \rceil \tag{3.3}$$

This allows a single frame to be considered in the context of the surrounding frames, and class labels to be adjusted accordingly.

3.3.3 Performance Analysis

The final stage of the pipeline is only performed if the gymnast was determined to be spinning. The aim of this stage is to track these spins in order to analyze their consistency. Using the identified silhouette from the human detection stage, vectors pointing from the

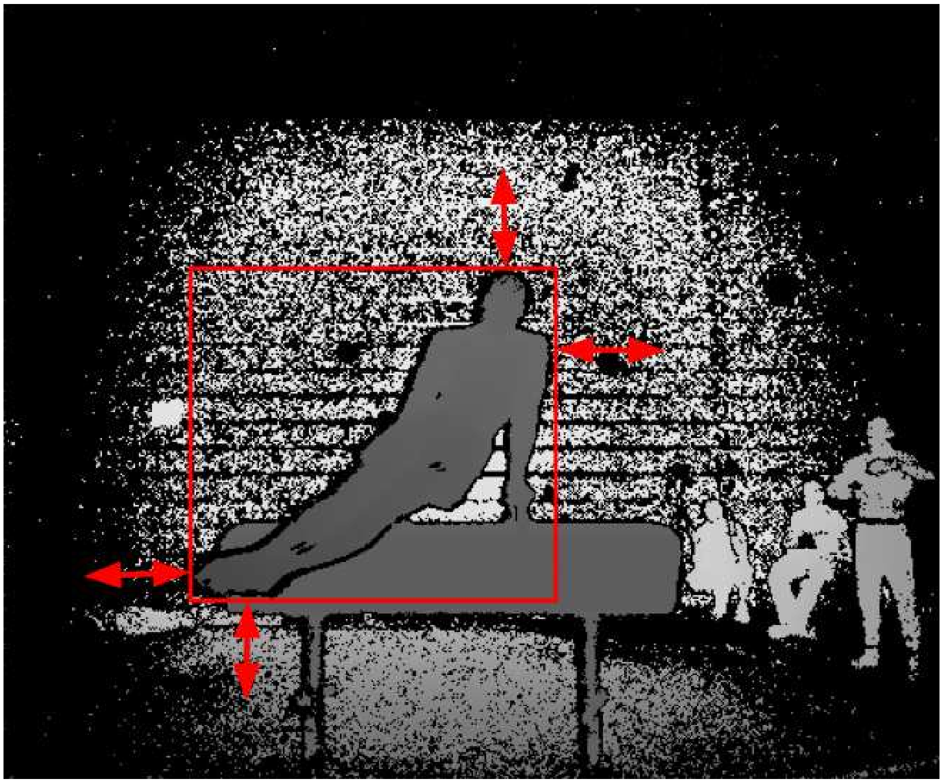


Figure 3.7: Illustration of the Silhouette Activity Descriptor in 3 Dimensions (SAD3D). The feature describes the width and height of a silhouette, the depth values at the left and right extremes, and the shift of the silhouette in both the x and y dimensions (indicated by the red arrows).

center of the gymnast's body to their head and feet were identified. Once the gymnast's head and feet are identified, accurately timing the spin around the pommel horse is a solveable problem.

This approach is based on defining a major axis for the gymnast's body, allowing for a bend at the waist to connect the head and feet. The first step is defining the longest vector from the gymnast's center to his contour - either the axis to his head or feet. Then, the 'waist' is defined by identifying the shortest vector from the center to the contour, and the corresponding 180 degree opposite. With this waist, we can find the other major axis of the body - the vector that approximates the opposite of the original vector. If the original vector points to the head, this will point to the feet. Otherwise, if the original vector points to the feet, this vector will point to the head. These vectors can be seen in Figure 3.8.

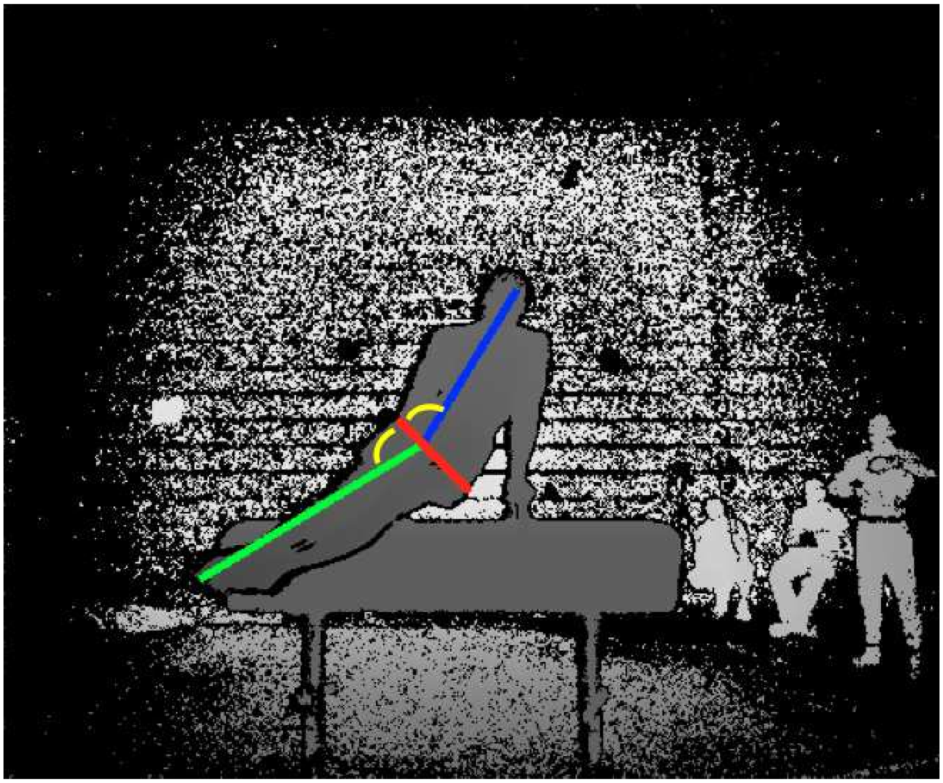


Figure 3.8: Illustration of the vectors involved in feet tracking. The blue vector identifies the location of the head, the red vector identifies the waist, and the green vector identifies the feet. The yellow angles are equal, corresponding to a gymnast bent at the waist.

This information provides reliable information about when the gymnast's feet are at their left and right extrema, relative to the camera. To find an extrema on either side of the pommel horse, the position of the gymnast's feet is compared to the two points before and the two points after the foot position. If the feet position is further to the left or further to the right of the other four points, then it is considered the extrema of a gymnast's spin.

These detected positions and timestamps are fitted to a cubic spline, to interpolate the exact timestamp of an extrema, even if it is between the frames recorded by the Kinect (Figure 3.9). Using this data, we compute the length of time it took for the gymnast to complete the spin by tracking the amount of time between consecutive left extrema or consecutive right extrema.

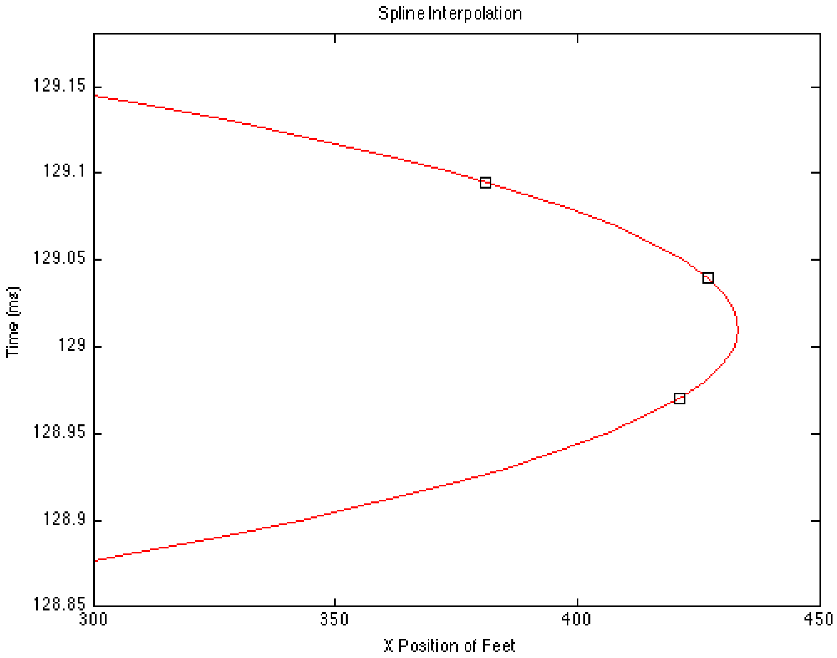


Figure 3.9: Cubic spline fit to observed foot positions. The three black squares indicate the positions of the gymnast's feet near a far right extrema. Because of the limited frame rate of the Kinect, fitting a spline to these points allows us to identify the specific timestamp of the extrema of the spin.

In addition, the foot position is used to record the angle of the gymnast's legs. As spins with level legs (in relation to the pommel horse) are judged as better, our approach measures the angle of the legs relative to the pommel horse. These angles are computed at both the left and right extremes.

3.4 Experimental Results

In this section I present results for each stage of the proposed pipeline. The system was tested on a setup consisting of a 1.7GHz i5 processor with 8GB of memory. This setup was used to validate the entire system, from recording gymnasts with the Kinect all the way through to analyzing the performances. The system was implemented in C++ and Matlab, with LibSVM [179] used as the activity classifier.

3.4.1 Depth of Interest Segmentation

The depth of interest segmentation approach was tested to determine how often the gymnast's silhouette is located in one of the three proposed segmentations, and evaluated by it's ability to reduce the amount of data necessary for later stages to analyze. In addition to the recorded gymnastics dataset, I also tested our approach on the CAD60 dataset [180].

Over a randomly selected sample of 500 frames, the described segmentation method was performed and the proposals were presented to a human evaluator. The evaluator recorded which proposal the gymnast was present in - if the gymnast was not completely present in the segmentation (e.g., a foot was outside of the segmentation neighborhood), then the proposal was not marked as correct. In our tests, 53.2% of proposals corresponding to the highest peak in the previously described PDF contained the gymnast's entire silhouette; the second highest peak contained the silhouette 36.6% of the time; and the third highest peak contained the silhouette only 8% of the time. Overall, our depth of interest approach contained the gymnast's complete silhouette in the top three proposals 97.8% of the time; and contained it in the top two proposals 89.8% of the time.

We also show that our approach drastically reduces the amount of data that later stages of the pipeline need to process. On average, original frames obtained from the Kinect contained 71255.71 non-zero pixels (out of a total possible of 217088 pixels at the 424x512 resolution). After segmentation, frames containing the gymnast had an average of 26948.17 non-zero pixels; this means that our method allows later stages of the pipeline (and further processing in other applications) to perform calculations on only 37.8% of the original data.

Additionally, we evaluated the depth of interest method on the CAD60 dataset, which consists of depth imagery of a variety of human activities. We measured the percentage of frames where the subject of the activity appeared in the proposals suggested by our method, using a similar sample size. The subject appeared 18.92% of the time in the first proposal, 66.67% in the second, and 14.41% in the third - there were zero instances where none of the proposed segmentations was correct. The large amount of occurecnces in the second proposal is due to the fact that some scenes in the CAD60 dataset have furniture closer to the camera than the subject. The gymnastics dataset has no objects between the camera and the pommel horse, so the first peak contains the gymnast more often the other two peaks. We also see that there was a similar reduction in the amount of data remaining after segmentation. Originally, the CAD60 frames contained an average of 61674.22 non-zero pixels. After segmentation, correct proposals contained only 23070.82 pixels on average, only 37.4% of the original data.

The depth of interest segmentation approach runs in real-time. On average, this method can generate proposed segmentations and run a HOG silhouette detector in only 26ms per frame.

3.4.2 Activity Recognition

The SAD3D feature was first evaluated on our constructed dataset. The data was split into 5024 training frames and 5091 testing frames, and our feature was computed for each frame. Using this representation, we were able to classify 93.81% of frames correctly, effectively segmenting the data. After the smoothing process, accuracy improved to classifying 94.83% of frames correctly.

Additionally, the approach was evaluated on the MSR Daily Activity 3D [181] dataset. We constructed our SAD3D feature from the ground truth joint data present in this dataset - e.g. the width of the silhouette was judged to be from the leftmost joint to the rightmost joint. Our approach was tested with these action types: ‘cheer up’, ‘lie on sofa’, and ‘walking’. Initially, SAD3D recognized 72.1% of frames correctly. After the smoothing described previously, this accuracy improved to 74.8%, making it competitive with existing activity recognition approaches while only using basic silhouette data (see Table 4.1).

With an identified silhouette, we are able to construct our feature representation in a trivial amount of time. Additionally, we are able to classify over half a million frames per second, a benefit of our low-dimensional feature.

Table 3.1: SAD3D activity recognition accuracy on the MSR Daily Activity 3D dataset. The performance of previous representations is compared to our representation's. Note that our accuracy was computed for only a portion of the dataset.

Representation	Accuracy
Dynamic Temporal Warping [181]	54.0%
Distinctive Canonical Poses [137]	65.7%
Actionlet Ensemble (3D pose only) [181]	68.0%
Relative Position of Joints [100]	70.0%
Moving Pose [182]	73.8%
Our SAD3D Representation	**74.8%**

3.4.3 Performance Analysis

The described method to detect the position of the gymnast's head and feet was evaluated against the dataset, for frames in which the gymnast was performing circles. As our dataset has a ground truth for both the time and frame number of each spin extrema, we compared our method against both. The foot detection algorithm was used to produce a timestamp, which was also mapped to a frame number.

In evaluation, the timestamps obtained from our performance analysis method were compared to the ground truth. The described approach achieves a root mean squared error (RMSE) of 12.9942ms from ground truth timestamps, with a frame number RMSE of 0.2393 frames. It achieves an average absolute error (treating a detection of 5ms late the same as a detection 5ms early) of 7.8168ms and 0.1352 frames. Additionally, errors were extremely clustered with few outliers - the standard deviation of timestamp errors was 12.7105ms, and the standard deviation of frame errors was 0.2361 frames.

Over the dataset, an average elite gymnast spin is 973ms, with a standard deviation of 67ms. If the first and last spins of a routine are removed (i.e. corresponding to mounting and dismounting the pommel horse), the mean time drops to 960ms with a standard deviation of 25ms. A graph showing spin times over a routine can be seen in Figure 3.10. Given this, the RMSE of our approach is below the standard deviation of a spin, meaning that the described method is accurate enough to analyze even extreme consistency of elite gymnasts.

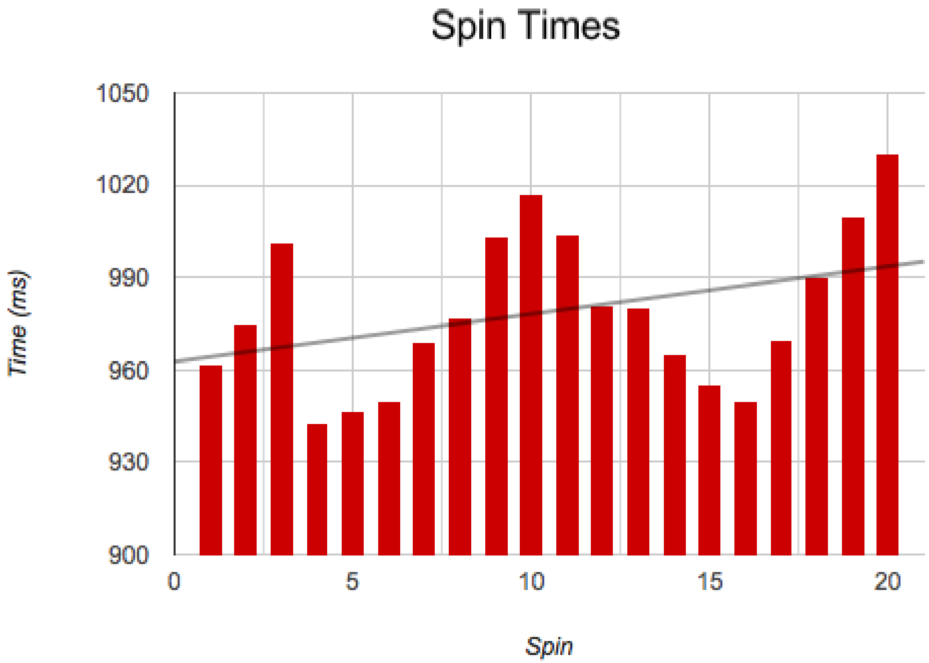


Figure 3.10: Gymnast spin times graphed over a routine. The black line indicates the trend, an example of the analysis our system makes available to coaches.

The described performance analysis approach runs in real-time on our system. It is able to process over 66 frames per second, putting it far ahead of the 30 frames per second limit of the Kinect and comparable cameras.

3.5 Case Study in a Real-world Application

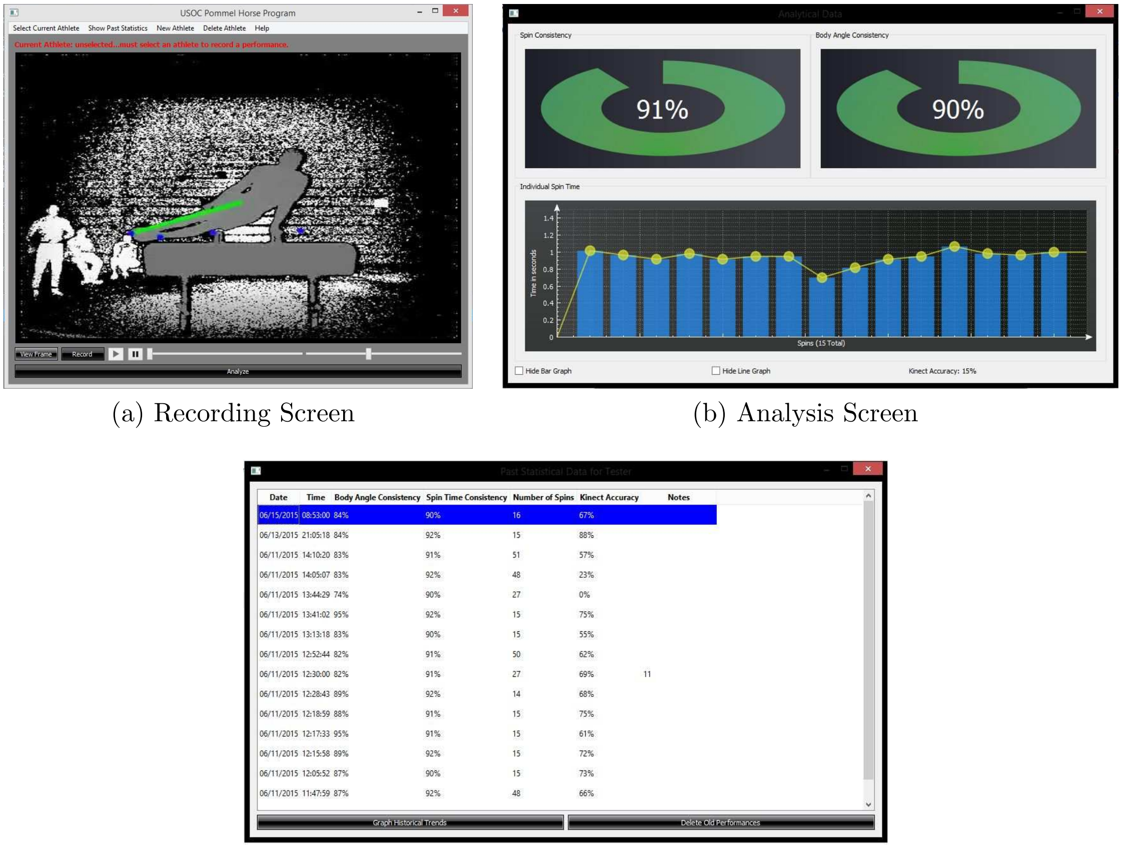
Our described method was integrated into a complete application, available for gymnastics coaches to use. The application allows coaches to record and analyze routines, and automatically scores them on spin time consistency and body angle consistency. Additionally, it allows coaches to record routines and scores for multiple gymnasts, tracking their progress and providing an average consistency score as they progressed.

As seen in Figure 3.11, the application is simple and well organized. Coaches simply have to position the Kinect in front of the pommel horse and begin recording routines. The application is currently being used by coaches at an elite gymnastics facility to record and track their gymnasts.

Figure 3.11(a) shows the recording screen. This gives coaches a live preview of what the Kinect is seeing, with the green vector indicating the current tracked position of the gymnast's feet. Seeing this allowed coaches to better understand how the system was generating the analysis. Figure 3.11(b) shows the analysis screen, which gives coaches a visual representation of their gymnast's performance. The green circles illustrate consistency 'scores' for both spin timing and leg angle, while the graph below shows times for individual spins in the routine. This allows coaches to identify where gymnasts are having problems - for example, increased or erratic spin times near the end of a routine would point to a problem with the gymnast's cardiovascular conditioning. Figure 3.11(c) shows the applications ability to track multiple gymnasts. Coaches can review a gymnast's performances, allowing them to see trends over time and identify gymnasts that are improving versus ones that need coaching adjustments.

Our described three-stage pipeline approach has many desirable characteristics. Some of these allow for increased usability for end users compared to hypothetical approaches, while others arise from the fact that our system is generalizable.

The method is completely automated, which makes it a very flexible system ideal for use by non-technical individuals. Because the approach can automatically localize the gymnast without specific depth or size constraints, users (such as gymnastics coaches) can place the Kinect at a distance and position convenient to them, instead of enforcing a particular pose. Similarly, the method's ability to identify a depth of interest and recognize a human form in that neighborhood is key. A naïve approach may use a background subtraction based approach to remove noise and identify the gymnast, but this would require the user to record an 'empty' segment before actually performing any analysis. Our approach avoids this, allowing coaches to have gymnasts begin training immediately. Additionally, our approach avoids issues that would arise for background subtraction if the camera was moved during the performance - this would cause the 'background' to no longer correctly correspond to



Current Athlete: unselected...must select an athlete to record a performance.

Kinect Accuracy: 15%

Date	Time	Body Angle Consistency	Spin Time Consistency	Number of Spins	Kinect Accuracy	Notes
06/15/2015	08:53:00	84%	90%	16	67%	
06/13/2015	21:05:18	84%	92%	15	88%	
06/11/2015	14:10:20	83%	91%	51	57%	
06/11/2015	14:05:07	83%	92%	48	23%	
06/11/2015	13:44:29	74%	90%	27	0%	
06/11/2015	13:41:02	95%	92%	15	75%	
06/11/2015	13:13:18	83%	90%	15	55%	
06/11/2015	12:52:44	82%	91%	50	62%	
06/11/2015	12:30:00	82%	91%	27	69%	11
06/11/2015	12:28:43	89%	92%	14	68%	
06/11/2015	12:18:59	88%	91%	15	75%	
06/11/2015	12:17:33	95%	91%	15	61%	
06/11/2015	12:15:58	89%	92%	15	72%	
06/11/2015	12:05:52	87%	90%	15	73%	
06/11/2015	11:47:59	87%	92%	48	66%	

(a) Recording Screen

(b) Analysis Screen

(c) Gymnast Tracker

Figure 3.11: Screenshots from the gymnastics analysis application. The application is simple and intuitive, ideal for non-technical users. Figure 3.11(a) shows the main screen used to record a performance. Figure 3.11(b) shows an evaluation of the gymnast's spin consistency and leg angle. Figure 3.11(c) shows the ability to track the progress of multiple gymnasts and play back their performances.

the current frame, causing an incorrect segmentation.

Our method also generalizes very well to a variety of situations. Because it is based solely on the depth information obtained from the Kinect, it is lighting-invariant. This allows it to be used in a variety of spaces that have different lighting conditions from where the original dataset was recorded.

The drawback to our described approach is that it has a lower limit on accuracy, based on the frame rate of the camera used. Currently, our system uses a Kinect 2 which operates at 30 frames per second, or 33 milliseconds between frames. As the RMSE of a detected spin is less than this time, camera frame rate is currently the limitation on our system's accuracy - with a faster depth camera, our system could obtain even more accurate results.

CHAPTER 4
BIPOD SKELETAL REPRESENTATION

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 (EKF) 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 the Baxter humanoid robot 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.

4.1 Motivation

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

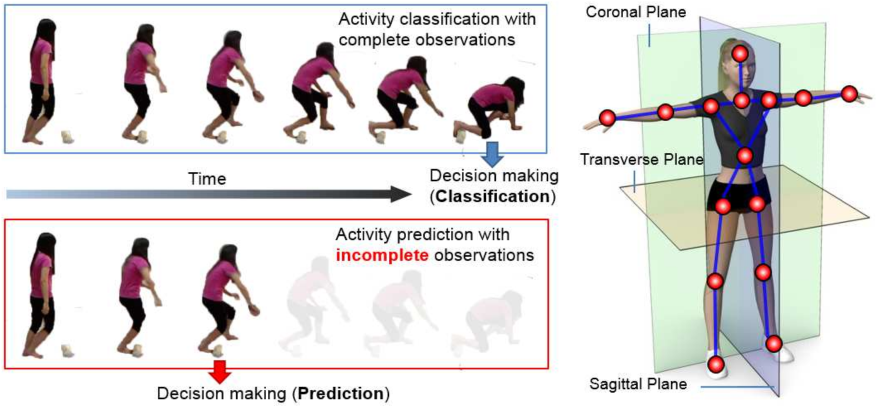

Activity classification with complete observations
Time
Decision making (Classification)
Activity prediction with incomplete observations
Decision making (Prediction)
Coronal Plane
Transverse Plane
Sagittal Plane

Figure 4.1: Motivation for the problem of activity prediction. A robot needs to infer ongoing human activities and make a decision based on incomplete observations.

and interact with people in human social environments. Although many activity recognition methods [135] have been proposed in robotics applications, most of them focus on classification of finished activities [134, 183, 184]. 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 decide 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 shown by the example in Figure 4.1. 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 to infer ongoing activities based on temporally *incomplete information*. 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 [134] 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 [24].

To address the aforementioned challenges, we introduce a novel 3D human representation based on a *Bio-Inspired Predictive Orientation Decomposition* of skeleton trajectories, called *BIPOD*. 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 [2], 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.

4.2 Approach

This section introduces the BIPOD representation. First, we discuss its foundation in human anatomy. Then, we introduce our approaches to estimate anatomical planes and to decompose spatio-temporal joint orientations onto these planes. Finally, we discuss our approach's predicative ability to address activity prediction.

4.2.1 Foundation in Biology

In human anatomy, human motions are described in three dimensions according to a series of planes named anatomical planes [185–187]. There are three anatomical planes of motions that pass through the human body, as demonstrated in Figure 4.2:

- *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.

Examples of human movements in each anatomical plane are demonstrated in Figure 4.2. 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* [186]. 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 [185, 187], 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 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

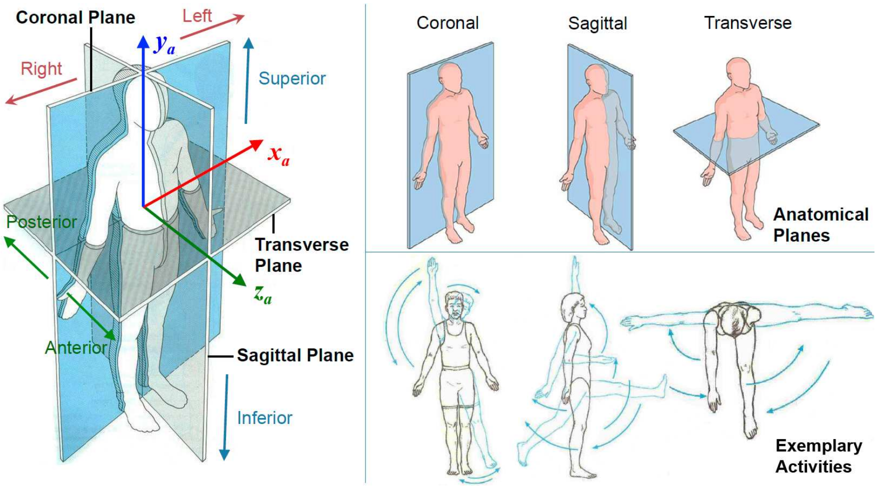


Figure 4.2: Explanation of anatomical planes. Our bio-inspired representation is based on human anatomy research. This figure demonstrates how anatomical planes divide the human body into different portions and illustrates exemplary human motions performed in each anatomical plane [186].

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.

4.2.2 Estimation of Anatomical Planes

A core procedure of our bio-inspired human representation is to estimate anatomical planes, which involves three major steps: inferring the (1) *coronal axis* z_a (intersection of the sagittal and transverse planes), the (2) *transverse axis* y_a (intersection of the coronal and sagittal planes), and the (3) *sagittal axis* x_a (intersection of the coronal and transverse planes). The anatomical axes x_a, y_a, z_a are illustrated in Figure 4.2.

Since the coronal plane is represented by human torso in anatomy [186], torse joints are used to estimate it. 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 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:

$$A(x_j - x_c) + B(y_j - y_c) + C(z_j - z_c) = \epsilon_j \tag{4.1}$$

The estimation of parameters (A, B, C) of the human coronal plane yields the regression problem, solved with the SVD method [188]:

$$R(A, B, C) = \sum_{i=1}^{M} \epsilon_i^2, \tag{4.2}$$

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 [186], as shown in Figure 4.2. 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). Typically, with most datasets we assume that the subject began in a T-pose and is thus facing towards the sensor. However, using an off-the-shelf human face detector, based on Haar cascades [189], it is possible to easily determine whether or not a face exists near the head 'joint' – a positive result means the subject is facing towards the sensor.

The origin of the estimated anatomy coordinate is placed at the human torso center, as shown in Figure 4.2. 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 Figure 4.2.

4.2.3 Anatomy-Based Orientation Decomposition

To construct a discriminative and compact representation. 3D trajectories of each joint of interest are described in relation to the 2D anatomical planes in a spatio-temporal fashion. 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, by projecting the original 3D trajectory onto the 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

$$\boldsymbol{P} = \{\boldsymbol{p}_t^{(x_a y_a)},\ \ \boldsymbol{p}_t^{(y_a z_a)},\ \ \boldsymbol{p}_t^{(z_a x_a)}\}_{t=1}^{T} \tag{4.3}$$

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.

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:

$$\theta_t = \arccos \frac{\overrightarrow{\boldsymbol{p}_{t-1}\boldsymbol{p}_t} \cdot \overrightarrow{\boldsymbol{p}_t\boldsymbol{p}_{t+1}}}{\|\overrightarrow{\boldsymbol{p}_{t-1}\boldsymbol{p}_t}\| \|\overrightarrow{\boldsymbol{p}_t\boldsymbol{p}_{t+1}}\|},\ \ t = 2, \ldots, T-1 \tag{4.4}$$

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. 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.

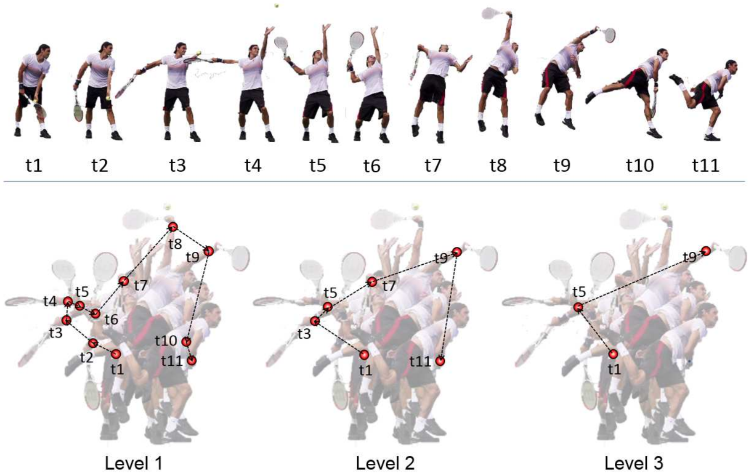


Figure 4.3: Temporal pyramid used to capture long-term independencies; the joint we are interested in is the right wrist, as denoted by the red dots. When three levels are used in the temporal pyramid, level 1 uses human skeleton data at all time points $(t_1, t_2, \ldots, t_{11})$; level 2 selects the 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.

The oriented angles computed based on Equation 4.4 can only capture temporal information within a defined time interval T. 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 Equation 4.4. In level 2 of the pyramid, only half of the temporal joint positions are adopted, for example, $t = 1, \ldots, 2n - 1$ where $n \in {\rm I\!R}$. If a temporal pyramid has three levels, then in level 3, only the joint data that satisfy $t = 1, \ldots, 4n - 1$ where $n \in {\rm I\!R}$ are applied to compute the orientation changes. Figure 4.3 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.

Through capturing both space (anatomy-based spatial decomposition) and time (temporal orientation description) information, our approach provides a spatio-temporal representation of humans and their movements.

4.2.4 Joint Trajectory Refinement and Prediction

Because skeletal data acquired from 3D robot perception systems can be noisy or occluded, it is important to estimate true positions of human joints given the observed skeleton data. In addition, to effectively predict activities, the representation requires the capability to estimate future human joint positions. To solve these problems, Extended Kalman Filters (EKFs) [190] 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. Relevant movement patterns of the body joints in different subspaces are typically assumed to be independent. 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.

The application of EKFs to track true human joint positions provides two advantages. First, it provides the capability to encode 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-dimension spaces is more efficient than calculating state changes for a single high-dimension space), but also makes our implementation flexible - 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 $x_k = A_k x_{k-1}$ and then corrections are made according to the measurement model $z_k = H_k x_k$. These models 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 a linear Kalman Filter, where both the matrices are constant. The output of the filter is used as the 'actual' position. This smooths out noise, can attempt to deal with occluded joints, and can provide a predicted future position for the joints.

4.3 Experimental Results

The performance of BIPOD was evaluated on available benchmark datasets to determine the accuracy of both activity classification and prediction. We also demonstrated the validity of our approach in interacting with a Baxter robot, showing that BIPOD is able to recognize our command actions while they are being performed.

The 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 3.0Ghz CPU with 16Gb of memory was used; the test on the TurtleBot2 was performed on an i3 1.7Ghz CPU with 4Gb 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 SVM with χ^2-kernels [191], which has demonstrated superior performance on the histogram-based feature [192]. To address multi-class classification and prediction, the standard one-against-one methodology is applied [191].

4.3.1 Performance on Benchmark Datasets

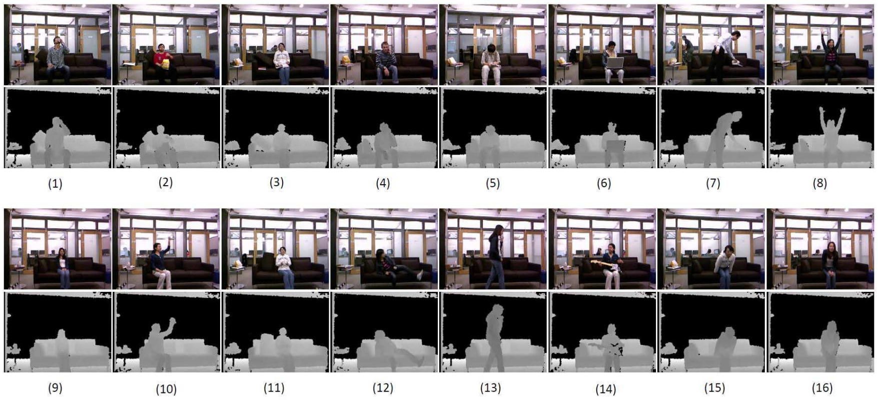


Figure 4.4: Example data from the MSR Daily Activity 3D dataset. The MSR Daily Activity 3D dataset contains 16 activity categories: (1) drink, (2) eat, (3) read book, (4) call cellphone, (5) write on a paper, (6) use laptop, (7) use vacuum cleaner, (8) cheer up, (9) sit still, (10) toss paper, (11) play game, (12) lie down on sofa, (13) walk, (14) play guitar, (15) stand up, (16) sit down.

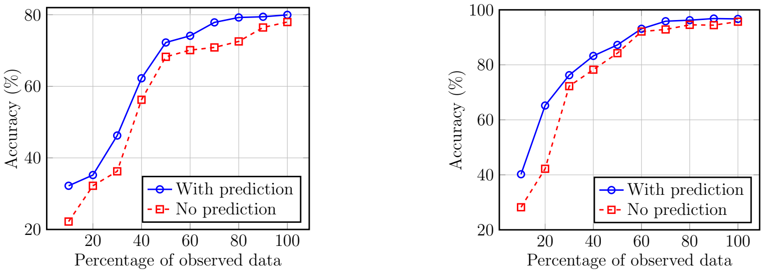
The MSR Daily Activity 3D dataset [181] is a widely used benchmark dataset in human activity recognition tasks, and was used in Chapter 3 to test the SAD3D representation. This dataset contains color-depth and skeleton information of 16 activity categories, as illustrated in Figure 4.4. Each activity 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.

Table 4.1: BIPOD activity recognition accuracy on MSR Daily Activity 3D dataset, compared with previous skeleton-based representations.

Skeleton-based representations	Accuracy
Dynamic Temporal Warping [181]	54.0%
Distinctive Canonical Poses [137]	65.7%
Actionlet Ensemble (3D pose only) [181]	68.0%
Relative Position of Joints [100]	70.0%
Moving Pose [182]	73.8%
Fourier Temporal Pyramid [181]	78.0%
Our BIPOD representation	**79.7%**

BIPOD was evaluated on 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 4.1, with comparisons to existing state-of-the-art methods. When a human activity is complete and all frames are observed, our approach obtains an average recognition accuracy of 79.7%. It is observed that our approach outperforms previous works and obtains the best recognition accuracy over this dataset. In addition, the efficiency of BIPOD was evaluated for the activity classification task. An average processing speed of 53.3 frames-per-second is obtained, which demonstrates the high efficiency of our representation.

BIPOD was also tested to determine it's activity prediction abilities. 15% future unobserved data are predicted by the component procedure of joint trajectory refinement and prediction, as discussed in Section 4.2.4. 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 Figure 4.5(a). It can be observed that BIPOD's predicition capa-



(a) MSR Daily Activity 3D

(b) HDM05 MoCap

Figure 4.5: Performance improvement by predicting future joint states. 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.

bilities allow it to achieve better recognition accuracy.

To validate the generalizability and applicability of our BIPOD representation on skeleton data collected from different sensing technologies, another set of experiments was conducted using skeletal data obtained using motion capture systems. The HDM05 MoCap dataset [30] is used in our experiments. Compared 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 since a multi-camera system can significantly reduce occlusions. Second, the human skeleton obtained by a MoCap system contains 31 joints. Since all motion sequences begin with a T-pose, as explained in [30], we simply assume subjects face toward the view point instead of detecting faces.

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 4.2 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

Table 4.2: BIPOD activity recognition accuracy on the HDM05 MoCap dataset, compared with previous skeleton-based representations.

Skeleton-based representations	Accuracy
Trifocal tensor of joint positions [193]	80.86%
Sequence of Most Informative Joints [194]	84.40%
Subtensor of joint positions [193]	85.71%
Relevant Joint Positions [195]	92.20%
Cov3DJ [72]	95.41%
Our BIPOD representation	**96.70%**

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 4.2. 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.

Similarly, BIPOD's predictive abilities were again tested, with 15% future data being predicted. Figure 4.5(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.

4.3.2 Real World Validation on a Baxter Robot

Finally, we tested BIPOD on a Baxter humanoid robot [196]. The Baxter robot, seen interacting in Figure 4.6(a), is ideally suited for human-robot interaction as it can ably mimic human activities. Additionally, it is designed for safe operation around humans, unlike typical industrial robot arms. Our Baxter robot was equipped with a Kinect sensor mounted on the 'chest', and all processing was done on a a networked Linux desktop. Skeleton data was obtained through OpenNI running on ROS, providing 15 joints.

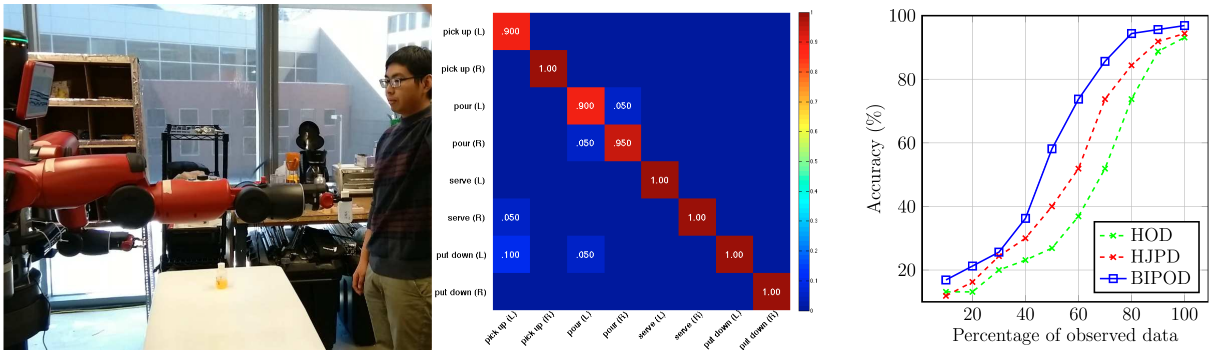
We created four new activites that would be used to interact with Baxter in order to control him through the process of making and serving a drink. Each is able to be performed on either side of the body, for a total of eight new activities: (1) pick up, (2) pour, (3) serve, and (4) put down. Ideally, a user could command Baxter to pick up a glass and a beverage, pour a drink, and serve it - with either arm. These 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 on which they are performed 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 Figure 4.6(a). 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 displacements (HOD) [114], and histograms of joint position differences (HJPD) [115].

Figure 4.6(b) shows the confusion matrix produced by our described BIPOD approach - 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 actions. Overall, our approach classifies only four instances incorrectly - making it 96.88% accurate. A comparison with

HOD and HJPD is shown in Table 4.3. 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 Figure 4.6(c), compared again 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.



(a) Environment Setup (b) Confusion matrix (c) Prediction ability

Figure 4.6: Experimental results on Baxter. Figure 4.6(a) illustrates the experimental setup, with Fei Han interacting with Baxter. Figure 4.6(b) show the confusion matrix for activity classification, and Figure 4.6(c) displays the accuracy rates for early prediction of activities for HOD [114], HJPD [115], and BIPOD.

Table 4.3: BIPOD activity recognition accuracy on Baxter, compared with previous skeleton-based representations.

Representations	Pick Up (L)	Pick Up (R)	Pour (L)	Pour (R)	Serve (L)	Serve (R)	Put Down (L)	Put Down (R)	Overall
HOD [114]	95	90	100	75	90	95	100	100	93.13
HJPD [115]	100	95	100	95	95	90	90	90	94.38
BIPOD	95	100	90	95	100	100	100	100	**96.88**

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 (described earlier in Section 4.2.4) is extremely useful for systems built on BIPOD's abilities. On a 2.7Ghz laptop with4 Gb of memory, the BIPOD representation can be constructed at 3600 'frames' per second - a 'frame' being either a representation built from skeleton data (with EKF processing to reduce noise) or a representation interpolated by the EKFs between frames of actual skeleton data. Using a pre-built SVM, these representations are able to be classified at a rate of 2800 per second.

4.4 Discussion

This skeleton-based representation based upon bio-inspired predictive orientation decomposition possesses several desirable characteristics, many of which make it unique in the field. The 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. 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, this 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. 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; it can be easily altered to represent only specific portions of the body and thus only predict and recognize actions 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 was 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.

CHAPTER 5
CONCLUSION

In this thesis, I demonstrate the applicability of depth imagery to the fields of sports analysis and human activity recognition. In both areas, I describe novel applications that address current limitations in the fields.

I introduce a novel system able to effectively provide an analysis of a gymnast's performance. My system addresses the problem of producing an automated analysis of a gymnast's performance on the pommel horse using a portable 3D camera, the Microsoft Kinect. This problem has been difficult in the past, but depth imagery provides a capable solution. The Kinect allows us to effectively segment the scene to identify a depth of interest, localize a gymnast within that region, identify what parts of a routine when a gymnast is performing circles, and then provide an accurate analysis of their performance. I can identify a depth of interest with 97.8% accuracy, detect spinning with 93.8% accuracy, and then analyze spin consistency with less than a 13ms RMSE, far less than the Kinect's 30 frames per second frame rate. With this method, it is possible to provide a real-world application that makes this analysis available to gymnastics coaches, providing them a quantitative basis for improvement.

I also introduce my contributions to the BIPOD representation, that enables intelligent robots to predict human activities in real time from 3D skeletal data in practical human-centered robotics applications. The 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 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. In addition, our BIPOD representation is compared with methods in previous studies on activity classification and prediction, using the 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 CITED

[1] Microsoft Kinect. https://dev.windows.com/en-us/kinect, 2012.

[2] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. *CVPR*, 2011.

[3] Fei Han, Brian Reily, William Hoff, and Hao Zhang. Space-time representation of people based on 3D skeletal data: A review. *ArXiv e-prints*, January 2016.

[4] Brian Reily, Fei Han, Lynne E. Parker, and Hao Zhang. Skeleton-Based Bio-Inspired Human Activity Prediction for Real-time Human-Robot Interaction. *Submitt. Publ. to IEEE Trans. Cybern.*, 2016.

[5] Brian Reily, William Hoff, and Hao Zhang. Real-time Gymnast Detection and Performance Analysis with a Portable 3D Camera. *Submitt. Publ. to Comput. Vis. Image Underst.*, 2016.

[6] D.G. Lowe. Object Recognition From Local Scale-Invariant Features. *ICCV*, 1999.

[7] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR*, 2005.

[8] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human Detection Using Oriented Histograms of Flow and Appearance. *ECCV*, 2006.

[9] Lubomir Bourdev and Jitendra Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. *ICCV*, 2009.

[10] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting People Using Mutually Consistent Poselet Activations. *ECCV*, 2010.

[11] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial Structures Revisited People Detection and Articulated Pose Estimation. *CVPR*, 2009.

[12] Pedro Felzenszwalb, Ross Girshick, David McAllester, and Deva Ramanan. Object Detection with Discriminative Trained Part Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1627–1645, 2010.

[13] Minh Hoai and Andrew Zisserman. Talking Heads: Detecting Humans and Recognizing Their Interactions. *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 875–882, 2014.

[14] Sho Ikemura and Hironobu Fujiyoshi. Real-Time Human Detection Using Relational Depth Similarity Features. *ACCV*, pages 25–38, 2010.

[15] Luciano Spinello and Kai Arras. People Detection in RGB-D Data. *IROS*, 2011.

[16] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. Human Detection Using Depth Information by Kinect. *CVPR*, 2011.

[17] Omid Hosseini Jafari, Dennis Mitzel, and Bastian Leibe. Real-Time RGB-D Based People Detection and Tracking for Mobile Robots and Head-Worn Cameras. *ICRA*, 2014.

[18] Dragomir Anguelov, Ben Taskar, Vassil Chatalbashev, Daphne Koller, Dinkar Gupta, Geremy Heitz, and Andrew Ng. Discriminative Learning of Markov Random Fields for Segmentation of 3D Scan Data. *CVPR*, 2:169–176, 2005.

[19] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan Barron, Ferran Marques, and Jitendra Malik. Multiscale Combinatorial Grouping. *CVPR*, 2014.

[20] Michael Maire and P Arbeláez. Using contours to detect and localize junctions in natural images. *Comput. Vis. Pattern Recognit.*, pages 1–8, 2008.

[21] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. From Contours to Regions: An Empirical Evaluation. *CVPR*, 2009.

[22] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour Detection and Hierarchical Image Segmentation. *PAMI*, 33(5):898–916, 2011.

[23] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous Detection and Segmentation. *ECCV*, pages 1–16, 2014.

[24] Hao Zhang, Christopher Reardon, and Lynne E. Parker. Real-Time Multiple Human Perception with Color-Depth Cameras on a Mobile Robot. *IEEE Trans. Cybern.*, 43 (5):1429–1441, 2013.

[25] ASUS Xtion PRO LIVE. https://www.asus.com/3D-Sensor/Xtion_PRO/, 2011.

[26] PrimeSense. https://en.wikipedia.org/wiki/PrimeSense, 2011.

[27] The OpenNI Library. http://structure.io/openni, 2013.

[28] The OpenKinect Library. http://www.openkinect.org, 2012.

[29] NITE. http://openni.ru/files/nite/, 2012.

[30] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical report, Universität Bonn, Jun. 2007.

[31] Mun Wai Lee and Ramakant Nevatia. Dynamic human pose estimation using Markov chain Monte Carlo approach. In *IEEE Workshops on Application of Computer Vision*, pages 168–175, 2005.

[32] Brian Holt, Eng-Jon Ong, Helen Cooper, and Richard Bowden. Putting the Pieces Together: Connected Poselets for Human Pose Estimation. *ICCV*, 2011.

[33] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

[34] Hao Zhang and Lynne E Parker. 4-dimensional local spatio-temporal features for human activity recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2044–2049, 2011.

[35] 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*, pages 1297–1304, 2011.

[36] Ross Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon. Efficient Regression of General-Activity Human Poses from Depth Images. *ICCV*, pages 415–422, 2011.

[37] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys. Accurate 3D pose estimation from a single depth image. In *IEEE International Conference on Computer Vision*, pages 731–738, 2011.

[38] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun. Random tree walk toward instantaneous 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2015.

[39] Min Sun, Pushmeet Kohli, and Jamie Shotton. Conditional regression forests for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3394–3401, 2012.

[40] James Charles and Mark Everingham. Learning Shape Models for Monocular Human Pose Estimation from the Microsoft Xbox Kinect. *ICCV*, pages 1202–1208, 2011.

[41] Daniel Grest, Jan Woetzel, and Reinhard Koch. Nonlinear Body Pose Estimation from Depth Images. *CVPR*, pages 285–292, 2005.

[42] Andreas Baak, Meinard Muller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera. *ICCV*, pages 1092–1099, 2011.

[43] James Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 103–110, 2012.

[44] Youding Zhu, Behzad Dariush, and Kikuo Fujimura. Controlled Human Pose Estimation from Depth Image Streams. *CVPR*, 2008.

[45] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real-time human pose tracking from range data. *European Conference on Computer Vision*, pages 738–751, 2012.

[46] Christian Plagemann, Varun Ganapathi, Daphne Koller, and Sebastian Thrun. Real-time Identification and Localization of Body parts From Depth Images. *ICRA*, 2010.

[47] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real Time Motion Capture Using a Single Time-Of-Flight Camera. *CVPR*, pages 755–762, 2010.

[48] Loren Arthur Schwarz, Artashes Mkhitaryan, Diana Mateus, and Nassir Navab. Human skeleton tracking from depth data using geodesic distances and optical flow. In *Image Vis. Comput.*, number 3, pages 217–226. Elsevier B.V., 2012. ISBN 0262-8856. doi: 10.1016/j.imavis.2011.12.001.

[49] Stephen Gould, Olga Russakovsky, Ian Goodfellow, and Paul Baumstarck. STAIR Vision Library. Technical report, 2011.

[50] Martin Fischler and Robert Elschlager. The Representation and Matching of Pictorial Structures. *IEEE Trans. Comput.*, C-22(1):67–92, 1973.

[51] Quanshi Zhang, Xuan Song, Xiaowei Shao, Ryosuke Shibasaki, and Huijing Zhao. Unsupervised skeleton extraction and motion capture from 3D deformable matching. *Neurocomputing*, 100:170–182, 2013.

[52] PJ BESL and ND MCKAY. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[53] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–203, 2013.

[54] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[55] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3D human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):914–927, 2014.

[56] Xiaodong Yang and YingLi Tian. EigenJoints-based action recognition using Näive-Bayes-Nearest-Neighbor. In *Workshops on IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[57] Xi Chen and Markus Koskela. Online RGB-D gesture recognition with extreme learning machines. In *ACM International Conference on Multimodal Interaction*, pages 467–474, 2013.

[58] Bangpeng Yao and Li Fei-Fei. Action recognition with exemplr based 2.5D graph matching. In *European Conference on Computer Vision*, pages 173–186. 2012.

[59] Xin Zhao, Xue Li, Chaoyi Pang, Xiaofeng Zhu, and Quan Z Sheng. Online human gesture recognition from motion data streams. In *ACM International Conference on Multimedia*, pages 23–32, 2013.

[60] Hossein Rahmani, Arif Mahmood, Ajmal Mian, and Du Huynh. Real time action recognition using histograms of depth gradients and random decision forests. In *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[61] Jiajia Luo, Wei Wang, and Hairong Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE International Conference on Computer Vision*, 2013.

[62] Xiaodong Yang and YingLi Tian. Effective 3D action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1):2–11, 2014.

[63] Chris Ellis, Syed Zain Masood, Marshall F Tappen, Joseph J Laviola Jr, and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 101(3):420–436, 2013.

[64] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4D human-object interactions for event and object recognition. In *IEEE International Conference on Computer Vision*, 2013.

[65] Ye Gu, Ha Do, Yongsheng Ou, and Weihua Sheng. Human gesture recognition through a Kinect sensor. In *IEEE International Conference on Robotics and Biomimetics*, pages 1379–1384, 2012.

[66] Jaeyong Sung, C. Ponce, B. Selman, and A Saxena. Unstructured human activity detection from RGBD images. In *IEEE International Conference on Robotics and Automation*, May 2012.

[67] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from RGBD images. In *Workshops on AAAI Conference on Artificial Intelligence*, 2011.

[68] Lu Xia, Chia-Chih Chen, and J.K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *Workshops on IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[69] Sou-Young Jin and Ho-Jin Choi. Essential body-joint and atomic action detection for human activity recognition using longest common subsequence algorithm. In *Workshops on Asian Conference on Computer Vision*, pages 148–159, 2013.

[70] Chenyang Zhang and Yingli Tian. RGB-D camera-based daily living activity recognition. *Journal of Computer Vision and Image Processing*, 2(4):12, 2012.

[71] Somar Boubou and Einoshin Suzuki. Classifying actions based on histogram of oriented velocity vectors. *Journal of Intelligent Information Systems*, 44(1):49–65, 2015.

[72] Mohamed E. Hussein, Marwan Torki, Mohammad a. Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. *IJCAI Int. Jt. Conf. Artif. Intell.*, pages 2466–2472, 2013.

[73] Alexandros Andre Chaaraoui, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert Systems with Applications*, 41(3):786–794, 2014.

[74] Miguel Reyes, Gabriel Domínguez, and Sergio Escalera. Feature weighting in dynamic timewarping for gesture recognition in depth data. In *Workshops on IEEE International Conference on Computer Vision*, pages 1182–1188, 2011.

[75] Orasa Patsadu, Chakarida Nukoolkit, and Bunthit Watanapa. Human gesture recognition using Kinect camera. In *International Joint Conference on Computer Science and Software Engineering*, pages 28–32, 2012.

[76] De-An Huang and Kris M Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *European Conference on Computer Vision*, pages 489–504. 2014.

[77] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[78] Wentao Zhu, Cuiling Lan, Junliang Xing, Yanghao Li, Li Shen, Wenjun Zeng, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *AAAI Conference on Artificial Intelligence, to appear*, 2016.

[79] Gutemberg Guerra-Filho and Yiannis Aloimonos. Understanding visuo-motor primitives for motion synthesis and analysis. *Computer Animation and Virtual Worlds*, 17 (3-4):207–217, 2006.

[80] Mohammad A. Gowayyed, Marwan Torki, Mohamed E. Hussein, and Motaz El-Saban. Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition. In *International Joint Conference on Artificial Intelligence*, 2013.

[81] Gang Yu, Zicheng Liu, and Junsong Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*, pages 50–65, 2014.

[82] Syed Z. Masood, Chris Ellis, Adarsh Nagaraja, Marshall F. Tappen, Joseph J LaViola Jr., and Rahul Sukthankar. Measuring and reducing observational latency when recognizing actions. In *Workshop on IEEE International Conference on Computer Vision*, 2011.

[83] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *IEEE International Conference on Computer Vision*, pages 2752–2759, 2013.

[84] Yongzhen Huang, Zifeng Wu, Liang Wang, and Tieniu Tan. Feature coding in image classification: A comprehensive study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):493–506, 2014.

[85] L.W. Campbell and AF. Bobick. Recognition of human body motion using phase space constraints. In *IEEE International Conference on Computer Vision*, pages 624–630, 1995.

[86] Yaser Sheikh, Mumtaz Sheikh, and Mubarak Shah. Exploring the space of a human action. In *IEEE International Conference on Computer Vision*, volume 1, pages 144–149, 2005.

[87] A Yilma and Mubarak Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *IEEE International Conference on Computer Vision*, volume 1, pages 150–157, 2005.

[88] Dian Gong, Gerard Medioni, and Xuemei Zhao. Structured time series analysis for human action segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1414–1427, 2014.

[89] Eshed Ohn-Bar and Mohan M Trivedi. Joint angles similarities and HOG2 for action recognition. In *Workshop on IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[90] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *The SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746, 2012.

[91] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-D human action using HMM and multi-class adaboost. In *European Conference on Computer Vision*, pages 359–372. 2006.

[92] Brent C Munsell, Andrew Temlyakov, Chengzheng Qu, and Song Wang. Person identification using full-body motion and anthropometric biometrics from Kinect videos. In *European Conference on Computer Vision*, pages 91–100, 2012.

[93] Suraj Vantigodi and Venkatesh Babu Radhakrishnan. Action recognition from motion capture data using meta-cognitive RBF network classifier. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pages 1–6, 2014.

[94] Mathieu Barnachon, Saïda Bouakaz, Boubakeur Boufama, and Erwan Guillou. Ongoing human action recognition with motion capture. *Pattern Recognition*, 47(1):238–247, 2014.

[95] Hao Zhang and Lynne E Parker. Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction. In *International Conference on Robotics and Automation*, 2015.

[96] Xinbo Jiang, Fan Zhong, Qunsheng Peng, and Xueying Qin. Online robust action recognition based on a hierarchical model. *The Visual Computer*, 30(9):1021–1033, 2014.

[97] Hossein Rahmani and Ajmal Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2458–2466, 2015.

[98] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5344–5352, 2015.

[99] Weijia Zou, Baoyuan Wang, and Rui Zhang. Human action recognition by mining discriminative segment with novel skeleton joint feature. In *Advances in Multimedia Information Processing*, pages 517–527. 2013.

[100] Lorenzo Seidenari, Varano Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Weakly Aligned Multi-part Bag-of-Poses for Action Recognition from Depth Cameras. *New Trends Image Anal. Process.*, 8158:446–455, 2013. doi: 10.1109/CVPRW.2013.77. URL ⟨GotoISI⟩://WOS:000343084300048.

[101] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[102] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-Patch: Unsupervised understanding of actions and relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4370, 2015.

[103] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014.

[104] Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. Joint action recognition and pose estimation from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301, 2015.

[105] Dian Gong and Gerard Medioni. Dynamic manifold warping for view invariant action recognition. In *IEEE International Conference on Computer Vision*, 2011.

[106] Rim Slama, Hazem Wannous, Mohamed Daoudi, and Anuj Srivastava. Accurate 3D action recognition using learning on the grassmann manifold. *Pattern Recognition*, 48 (2):556–567, 2015.

[107] Rizwan Chaudhry, Ferda Ofli, Gregorij Kurillo, Ruzena Bajcsy, and Rene Vidal. Bio-inspired Dynamic 3D Discriminative Skeletal Features for Human Action Recognition. *2013 IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, pages 471–478, 2013.

[108] Leandro Miranda, Thales Vieira, Dimas Martinez, Thomas Lewiner, Antonio W Vieira, and Mario FM Campos. Real-time gesture recognition from depth data through key poses learning and decision forests. In *SIBGRAPI Conference on Graphics, Patterns and Images*, pages 268–275, 2012.

[109] Chunyu Wang, Yizhou Wang, and Alan L Yuille. An approach to pose-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2013.

[110] Ioannis Kapsouras and Nikos Nikolaidis. Action recognition on motion capture data using a dynemes and forward differences representation. *Journal of Visual Communication and Image Representation*, 25(6):1432–1445, 2014.

[111] Moustafa Meshry, Mohamed E Hussein, and Marwan Torki. Linear-time online action detection from 3D skeletal data using bags of gesturelets. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, accepted, 2016.

[112] Lingling Tao and René Vidal. Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition. In *Workshops on IEEE International Conference on Computer Vision*, pages 61–69, 2015.

[113] Somar Boubou and Einoshin Suzuki. Classifying actions based on histogram of oriented velocity vectors. *J. Intell. Informations Syst.*, 44(1):49–65, 2015.

[114] Mohammad a. Gowayyed, Marwan Torki, Mohamed E. Hussein, and Motaz El-Saban. Histogram of Oriented Displacements (HOD): Describing trajectories of human joints for action recognition. *IJCAI Int. Jt. Conf. Artif. Intell.*, 25:1351–1357, 2013.

[115] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition. *Eccv*, pages 742–757, 2014.

[116] Hossein Rahmani, Arif Mahmood, Du Q. Huynh, and Ajmal Mian. Real time action recognition using histograms of depth gradients and random decision forests. *2014 IEEE Winter Conf. Appl. Comput. Vision, WACV 2014*, pages 626–633, 2014.

[117] Lu Xia, Chia-Chih Chen, and J K Aggarwal. View Invariant Human Action Recognition Using Histograms of {3D} Joints. *CVPR Work.*, 2012.

[118] Zhang Fan, Guo Li, Lu Haixian, Gui Shu, and Li Jinkui. Star skeleton for human behavior recognition. *2012 Int. Conf. Audio, Lang. Image Process.*, pages 1046–1050, 2012.

[119] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 588–595, 2014.

[120] Xiaodong Yang and Yingli Tian. EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. *Comput. Vis. Pattern Recognit. Work. (CVPRW), 2012*, pages 14–19, 2012.

[121] Xiaodong Yang and Yingli Tian. Effective 3D action recognition using EigenJoints. *J. Vis. Commun. Image Represent.*, 25(1):2–11, 2014.

[122] Hao Zhang and Lynne E Parker. Bio-Inspired Predictive Orientation Decomposition of Skeleton Trajectories for Real-Time Human Activity Prediction. *ICRA*, pages 3053–3060, 2015.

[123] Eshed Ohn-bar and Mohan M Trivedi. Joint Angles Similiarities and HOG 2 for Action Recognition. *Cvpr*, 2013.

[124] Di Wu and Ling Shao. Leveraging Hierarchical Parametric Networks for Skeletal Joints Based Action Segmentation and Recognition. *Conf. Comput. Vis. Pattern Recognit.*, 2014.

[125] Victoria Bloom, Vasileios Argyriou, and Dimitrios Makris. Dynamic feature selection for online action recognition. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 8212 LNCS:64–76, 2013.

[126] Farhood Negin, Firat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil. A decision forest based feature selection framework for action recognition from RGB-depth cameras. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 7950 LNCS:648–657, 2013.

[127] Y Zhai and M Shah. A general framework for temporal video scene segmentation. *Comput. Vision, 2005. ICCV 2005. Tenth IEEE Int. Conf.*, 2:1111 –1116 Vol. 2, 2005.

[128] Hao Zhang, Wenjun Zhou, and Lynne E Parker. Fuzzy Segmentation and Recognition of Continuous Human Activities. *2014 IEEE Int. Conf. Robot. Autom.*, pages 6305–6312, 2014.

[129] F Zhou, F De la Torre, and J.K. Hodgins. Hierarchical Aligned Cluster Analysis For Temporal Clustering of Human Motion. *PAMI*, 35(3):582–596, 2013.

[130] Ekaterina H. Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. *2009 IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2009*, pages 17–24, 2009.

[131] Yu Cheng, Quanfu Fan, Sharath Pankanti, and Alok Choudhary. Temporal Sequence Modeling for Video Event Detection. *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2235–2242, 2014.

[132] Kuan-Ting Lai, Felix X. Yu, Ming-Syan Chen, and Shih-Fu Chang. Video Event Detection by Inferring Temporal Instance Labels. *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2251–2258, 2014.

[133] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal Segmentation of Egocentric Videos Yair Poleg. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 2537–2544, 2014.

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

[135] J.K. Aggarwal and Lu Xia. Human activity recognition from 3D data: A review. *Pattern Recognition Letters*, 48(0):70–80, 2014.

[136] 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.

[137] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola Jr., and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *Int. J. Comput. Vis.*, 101:420–436, 2013.

[138] J.R. Hoare and L.E. Parker. Using on-line conditional random fields to determine human intent for peer-to-peer human robot teaming. In *IROS*, 2010.

[139] Juan Carlos Niebles and Li Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[140] Yelin Kim, Jixu Chen, Ming-Ching Chang, Xin Wang, E.M. Provost, and Siwei Lyu. Modeling transition patterns between events for temporal human action segmentation and classification. *Automatic Face and Gesture Recognition*, 2015.

[141] Alex Pentland and Andrew Liu. Modeling and prediction of human behavior. *Neural Computation*, 11(1):229–242, 1999.

[142] H. Berndt, J. Emmert, and K. Dietmayer. Continuous driver intention recognition with hidden markov models. *Intelligent Transportation Systems*, pages 1189–1194, 2008.

[143] Feng Dai, Jianwu Zhang, and Tongli Lu. The study of driver's starting intentions. *Mechanic Automation and Control Engineering*, pages 2758–2761, 2011.

[144] Lisheng Jin, Haijing Hou, and Yuying Jiang. Driver intention recognition based on continuous hidden markov model. *Transportation, Mechanical, and Electrical Engineering*, pages 739–742, 2011.

[145] Lei He, Chang-fu Zong, and Chang Wang. Driving intention recognition and behaviour prediction based on a double-layer hidden markov model. *Journal of Zhejiang University*, 13:208–217, 2012.

[146] Luzheng Bi, Xuerui Yang, and Cuie Wang. Inferring driver intentions using a driver model based on queuing network. *Intelligent Vehicles Symposium*, pages 1387–1391, 2013.

[147] T. Georgiou and Y. Demiris. Predicting car states through learned models of vehicle dynamics and user behaviours. *Intelligent Vehicles Symposium*, pages 1240–1245, 2015.

[148] Gaetano Bosurgi, Antonino D'Andrea, and Orazio Pellegrino. Prediction of drivers' visual strategy using an analytical model. *Journal of Transportation Safety & Security*, 7, 2014.

[149] Gys Albertus Marthinus Meiring and Hermanus Carel Myburgh. A review of intelligent driving style analysis systems and related artificial intelligence algorithms. *Sensors*, 15:30653–30682, 2015.

[150] Wenshuo Wang, Junqiang Xi, and Huiyan Chen. Modeling and recognizing driver behavior based on driving data: A survey. *Mathematical Problems in Engineering*, 2014.

[151] Yves Boussemart and Mary L. Cummings. Predictive models of human supervisory control behavioral patterns using hidden semi-markov models. *Engineering Applications of Artifical Intelligence*, 24:1252–1262, 2011.

[152] Zhikun Wang, Abdeslam Boularias, Katharina Mulling, Bernhard Scholkopf, and Jan Peters. Anticipatory action selection for human-robot table tennis. *Artificial Intelligence*, 2014.

[153] Kang Li, Jie Hu, and Yun Fu. Modeling complex temporal composition of actionlets for activity prediction. *European Conference on Computer Vision*, pages 286–299, 2012.

[154] Kang Li and Yun Fu. Prediction of human activity by discovering temporal sequence patterns. *Pattern Analysis and Machine Intelligence*, 36:1644–1657, 2014.

[155] Michael S Ryoo, Kristen Grauman, and Jake K Aggarwal. A task-driven intelligent workspace system to provide guidance feedback. *Computer Vision and Image Understanding*, 114(5):520–534, 2010.

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

[157] Gang Yu, Junsong Yuan, and Zicheng Liu. Predicting human activities using spatio-temporal structure of interest points. In *ACM International Conference on Multimedia*, 2012.

[158] Ian Reade, Wendy Rodgers, and Katie Spriggs. New Ideas for High Performance Coaches: A Case Study of Knowledge Transfer in Sport Science. *Int. J. Sport. Sci. Coach.*, 3(3):335–354, 2008.

[159] Ian Reade, Wendy Rodgers, and Nathan Hall. Knowledge Transfer: How do High Performance Coaches Access the Knowledge of Sport Scientists? *Int. J. Sport. Sci. Coach.*, 3(3):319–334, 2009.

[160] S Barris and C Button. A Review of Vision-Based Motion Analysis in Sport. *Sport. Med.*, 38(12):1025–1043, 2008.

[161] Adrian Lees. Technique Analysis in Sports: A Critical Review. *J. Sports Sci.*, 20(10): 813–828, 2002.

[162] S Prassas, YH Kwon, and WA Sands. Biomechanical Research in Artistic Gymnastics: A Review. *Sport. Biomech.*, 2006.

[163] G Irwin and D G Kerwin. The Influence of the Vaulting Table on the Handspring Front Somersault. *Sport. Biomech.*, 8(2):114–128, 2009.

[164] A L Sheets and M. Hubbard. Evaluation of a Subject-specific Female Gymnast Model and Simulation of an Uneven Parallel Bar Swing. *J. Biomech.*, 41(15):3139–3144, 2008.

[165] Jaroslaw Omorczyk, Leszek Nosiadek, Tadeusz Ambrozy, and Andrzej Nosiadek. High Frequency Video Capture and a Computer Program with Frame-by-frame Angle Determination Functionality as Tools that Support Judging in Artistic Gymnastics. *Acta Bioeng. Biomech.*, 17(3):85–93, 2014.

[166] Keith L. Markolf, Matthew S. Shapiro, Bert R. Mandelbaum, and Luc Teurlings. Wrist loading patterns during pommel horse exercises. *J. Biomech.*, 23(10):1001–1011, 1990.

[167] Toshiyuki Fujihara, Takafumi Fuchimoto, and Pierre Gervais. Biomechanical Analysis of Circles on Pommel Horse. *Sport. Biomech.*, 8(1):22–38, 2009.

[168] Toshiyuki Fujihara and Pierre Gervais. Circles on Pommel Horse with a Suspended Aid: Spatio-temporal Characteristics. *J. Sports Sci.*, 30(6):571–581, 2012.

[169] Toshiyuki Fujihara and Pierre Gervais. Circles on Pommel Horse with a Suspended Aid: Influence of Expertise. *J. Sports Sci.*, 30(6):583–589, 2012.

[170] L. Baudry, D. Leroy, and D. Choilet. The effect of combined self- and expert-modelling on the performance of the double leg circle on the pommel horse. *J. Sports Sci.*, 24 (10):1055–1063, 2006.

[171] P Federolf, P Scheiber, E Rauscher, H Schwameder, A Luthi, H Rhyner, and E Muller. Impact of Skier Actions on the Gliding Times in Alpine Skiing. *Scand. J. Med. Sci. Sports*, 2008.

[172] J.K. Moore, J. D. G. Kooijman, A.L. Schwab, and M. Hubbard. Rider Motion Identification during Normal Bicycling by Means of Principal Component Analysis. *Multibody Syst. Dyn.*, 25(2):225–244, 2011.

[173] John Brenkus. SportsScience, 2007. URL http://espn.go.com/espn/sportscience/.

[174] Microsoft Kinect, 2012. URL https://dev.windows.com/en-us/kinect.

[175] CMU Graphics Lab Motion Capture Database, 2001. URL http://mocap.cs.cmu.edu.

[176] An-An Liu, Ning Xu, Yu-Ting Su, Hong Lin, Tong Hao, and Zhao-Xuan Yang. Single/Multi-view Human Action Recognition Via Regularized Multi-task Learning. *Neurocomputing*, 151:544–553, 2015.

[177] David Romer. It's Fourth Down and What Does the Bellman Equation Say? A Dynamic Programming Analysis of Football Strategy. Technical report, National Bureau of Economic Research, 2002.

[178] Emanuel Parzen. On estimation of a probability density function and mode, 1962.

[179] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1—-27:27, 2011.

[180] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human Activity Detection from RGBD Images. In *Assoc. Adv. Artif. Intell. Work. Pattern, Act. Intent Recognit.*, 2011. URL http://pr.cs.cornell.edu/humanactivities/data.php.

[181] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 1290–1297, 2012. ISSN 10636919. doi: 10.1109/CVPR.2012.6247813.

[182] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. *IEEE Int. Conf. Comput. Vis.*, pages 2752–2759, 2013. ISSN 1550-5499. doi: 10.1109/ICCV.2013.342. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6751453$\backslash$npapers3://publication/doi/10.1109/ICCV.2013.342.

[183] Guang Chen, Manuel Giuliani, Daniel Clarke, Andre Gaschler, and Alois Knoll. Action recognition using ensemble weighted multi-instance learning. In *IEEE International Conference on Robotics and Automation*, 2014.

[184] Alessandro Pieropan, Giampiero Salvi, Karl Pauwels, and Hedvig Kjellstrom. Audio-visual classification and detection of human manipulation actions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014.

[185] Henry Gray. *Anatomy of the Human Body*. Lea & Febiger, 1973.

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

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

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

[189] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[190] G.A. Einicke and L.B. White. Robust extended Kalman filtering. *IEEE Trans. Signal Processing*, 47(9):2596–2599, Sept. 1999.

[191] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[192] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

[193] Qiguang Liu and Xiaochun Cao. Action recognition using subtensor constraint. In *European Conference on Computer Vision*, 2012.

[194] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 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, Jan. 2014.

[195] Adolfo López-Mendez, Juergen Gall, Josep R. Casas, and Luc J. Van Gool. Metric learning from poses for temporal clustering of human motion. In *British Machine Vision Conference*, 2012.

[196] Rethink robotics: Baxter, 2015. Accessed: 16 March 2016.