Space-Time Representation of People Based on 3D Skeletal Data: A Review

Fei Han*, Brian Reily*, William Hoff, and Hao Zhang

Abstract—Spatiotemporal human representation based on 3D visual perception data is a rapidly growing research area. Based on the information sources, these representations can be broadly categorized into two groups based on RGB-D information or 3D skeleton data. Recently, skeleton-based human representations have been intensively studied and kept attracting an increasing attention, due to their robustness to variations of viewpoint, human body scale and motion speed as well as the realtime, online performance. This paper presents a comprehensive survey of existing space-time representations of people based on 3D skeletal data, and provides an informative categorization and analysis of these methods from the perspectives, including information modality, representation encoding, structure and transition, and feature engineering. We also provide a brief overview of skeleton acquisition devices and construction methods, enlist a number of public benchmark datasets with skeleton data, and discuss potential future research directions.

Index Terms—Human representation, skeleton data, 3D visual perception, space-time features, survey

1 Introduction

Human representation in spatiotemporal space is a fundamental research problem extensively investigated in computer vision and machine intelligence over the past few decades. The objective of building human representations is to extract compact, descriptive information (i.e., features) to encode and characterize a human's attributes from perception data (e.g., human shape, pose, and motion), when developing recognition or other human-centered reasoning systems. As an integral component of reasoning systems, approaches to construct human representations have been widely used in a variety of real-world applications, including video analysis [1], surveillance [2], robotics [3], human-machine interaction [4], augmented and virtual reality [5], assistive living [6], smart homes [7], education [8], and many others [9], [10], [11].

During recent years, human representations based on 3D perception data have been attracting an increasing amount of attention [12], [13], [14], [15]. Comparing with 2D visual data, additional depth information provides several advantages for building 3D human representations. Depth images provide geometric information of pixels that encode the external surface of the scene in 3D space. Features extracted from depth images and 3D point clouds are robust to variations of illumination, scale, and rotation [16], [17]. Thanks to the emergence of affordable structured-light color-depth sensing technology, such as the Microsoft Kinect [18] and Asus Xtion PRO LIVE [19] RGB-D cameras, it is much easier and cheaper to obtain depth data. In addition, structured-light cameras enable us to retrieve the 3D human skeletal information in real time [20], which used to be only possible when using expensive and complex vision systems (e.g., motion capture systems [21]), thereby significantly popularizing skeleton-based human representations. Moreover, the vast increase in computational power allows researchers to develop advanced computational algorithms (e.g., deep learning [22]) to process visual data at an acceptable speed. The advancements contribute to the boom of utilizing 3D perception data to construct reasoning systems in computer vision and machine learning communities.

Since the performance of machine learning and reasoning methods heavily relies on the design of data representation [23], human representations are intensively investigated to address human-centered research problems (e.g., human detection, tracking, pose estimation, and action recognition). Among a large number of human representation approaches [24], [25], [26], [27], [28], [29], most of the existing 3D based methods can be broadly grouped into two categories: representations based on local features [30], [31] and skeleton-based representations [32] [33] [34]. Methods based on local features detect points of interest in space-time dimensions, describe the patches centered at the points as features, and encode them (e.g., using bag-of-word models) into representations, which can locate salient regions and are relatively robust to partial occlusion. However, methods based on local features ignore spatial relationships among the features. These approaches are often incapable of identifying feature affiliations, and thus the methods are generally incapable to represent multiple individuals in the same scene. These methods are also computationally expensive because of the complexity of the procedures including keypoint detection, feature description, dictionary construction, etc.

On the other hand, human representations based on 3D skeleton information provide a very promising alternative. The concept of skeleton-based representation can be traced back to the early seminal research of Johansson [49], which demonstrated that a small number of joint positions can effectively represent human behaviors. 3D skeleton-based representations also demonstrate promising performance in

• *F. Han*, B. Reily*, W. Hoff, and H. Zhang are with the Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO, 80401. * These authors contributed equally to this work, E-mail: {fhan, breily, whoff, hzhang}@mines.edu*

TABLE 1
Existing Surveys in Related Fields

Year	Review Papers	Focus
2015	Lun and Zhao [35]	Human motion recognition with Kinect
2014	Aggarwal and Xia [16]	Human activity recognition from 3D data
2014	Ruffieux et al. [36]	Datasets for human gesture recognition
2013	Borges et al. [37]	Video-based human behavior understanding
2013	Chen et al. [38]	Human motion analysis using depth imagery
2013	Han et al. [17]	Computer vision with Kinect
2013	Ke et al. [39]	Video-based human activity recognition
2013	LaViola [40]	3D gestural interaction
2013	Ye et al. [41]	Human activity recognition from depth data
2012	Chaaraoui et al. [42]	Human behaviour analysis for ambient-assisted living
2011	Aggarwal and Ryoo [43]	Human activity analysis
2010	Ji and Liu [44]	View-invariant human motion analysis
2010	Poppe [45]	Vision-based human action recognition
2008	Zhou and Hu [46]	Human motion tracking for rehabilitation
2006	Moeslund et al. [47]	Vision-based human motion capture and analysis
2001	Moeslund and Granum [48]	Computer vision-based human motion capture

real-world applications including Kinect-based gaming, as well as in computer vision research [22], [50]. 3D skeleton-based representations are able to model the relationship of human joints and encode whole body configuration. They are also robust to scale and illumination changes, and can be invariant to camera view as well as human body rotation and motion speed. In addition, many skeleton-based representations can be computed at a high frame rate, which can significantly facilitate online, real-time applications. Given the advantages and previous success of 3D skeleton-based representations, we have witnessed a significant increase of new techniques to construct such representations in recent years, as demonstrated in Fig. 1, which underscores the need of this survey paper focusing on the review of 3D skeleton-based human representations.

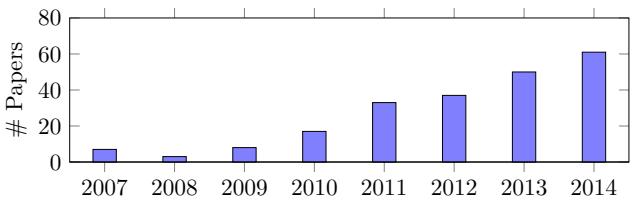


Fig. 1. Number of 3D skeleton-based human representations published in recent years according to our comprehensive review.

Several survey papers were published in related research areas such as motion and activity recognition. For example, Han et al. [17] introduced how Kinect works and its general applications in computer vision and machine intelligence. Aggarwal and Xia [16] recently published a review paper on human activity recognition from 3D visual data, which summarized five categories of representations based on 3D silhouettes, skeletal joints or body part locations, local spatio-temporal features, scene flow features, and local occupancy features. Several earlier surveys were also published to review methods to recognize human poses, motions, gestures, and activities [35], [36], [37], [38], [39], [40], [41], [43], [47], [48], as well as their applications [42], [46], as summarized by the complete list in Table 1. However, none of the survey papers specifically focused on the 3D human representation based on skeletal data, which was investigated by numerous research papers in literature and continues to gain popularity in recent years.

The objective of this survey is to provide a comprehensive overview of 3D skeleton-based human representations published in the computer vision and machine intelligence communities. We categorize and compare the reviewed approaches from multiple perspectives, including information modality, representation coding, structure and transition, and feature engineering methodology, and analyze the pros and cons of each category. A comprehensive review on methods to acquire and estimate 3D human skeleton and a complete list of available benchmark datasets are also included. Compared with the existing surveys, the main contributions of this review include:

- To the best of our knowledge, this is the first survey dedicated to *human representations based on 3D skeleton data*, which fills the current void in the literature.
- The survey is *comprehensive* and covers the *most recent and advanced* approaches. We review 158 3D skeleton-based human representations, including 142 papers that were published in the recent five years, thereby providing readers with the complete, state-of-the-art methods.
- This paper provides an insightful categorization and analysis of the 3D skeleton-based representation construction approaches from multiple perspectives, and summarizes and compares attributes of all reviewed representations.

The remainder of this review is organized as follows. The background information including 3D skeleton acquisition and construction as well as benchmark datasets is presented in Section 2. Sections 3 to 5 discuss the categorization of 3D skeleton-based human representations from four perspectives, including information modality in Section 3, encoding in Section 4, hierarchy and transition in Section 5, and feature construction methodology in Section 6. After discussing the advantages of skeleton-based representations and pointing out future research directions in Section 7, the review paper is concluded in Section 8.

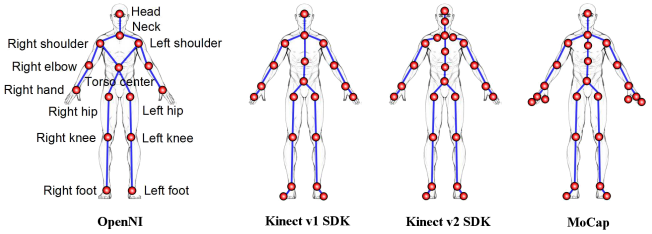


Fig. 2. Examples of skeletal human body models obtained from different sensors. The OpenNI library tracks 15 joints; The Kinect V1 SDK tracks 20 joints; The Kinect v2 SDK tracks 25; and MoCap systems can track various numbers of joints.

2 Background

The objective of building 3D skeleton-based human representations is to extract compact, discriminative description to characterize human's attributes from 3D human skeletal information. The 3D skeleton data encodes human body as an articulated system of rigid segments connected by joints. This section discusses how 3D skeletal data can be acquired, including devices that directly provide the skeletal data and computational methods to construct the skeleton. Available benchmark datasets including 3D skeleton information are also summarized in this section.

2.1 Direct Acquisition of 3D Skeletal Data

Several commercial sensors, including motion capture systems, time-of-flight sensors, and structured-light cameras, allow for direct retrieval of 3D skeleton data. The 3D skeletal kinematic human body models provided by the devices are illustrated in Fig. 2.

2.1.1 Motion Capture Systems (MoCap)

Motion capture systems identify and track markers that are attached to a human subject's joints or body parts to obtain 3D skeleton information. There are two main categories of MoCap systems, based on either visual cameras or inertia sensors. Optical-based systems employ multiple cameras positioned around a subject to track, in 3D space, reflective markers attached to the human body. In MoCap systems based on inertial sensors, each 3-axis inertial sensor estimates the rotation of a body part with respect to a fixed point. This information is collected to obtain the skeleton data without any optical devices around a subject. Software to collect skeleton data is provided with commercial MoCap systems, such as Nexus for Vicon MoCap[1], NatNet SDK for OptiTrack[2], etc. MoCap systems, especially based on multiple cameras, can provide very accurate 3D skeleton information at a very high speed. On the other hand, such systems are typically expensive and can only be used in well controlled indoor environments.

2.1.2 Structured-Light Cameras

Structured-light color-depth sensors are a type of camera that uses infrared light to capture depth information about a scene, such as Microsoft Kinect v1 [18], ASUS Xtion PRO LIVE [19], and PrimeSense [74], 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. Several drivers are available to provide the access to the color-depth data acquired by the sensor, including the Microsoft Kinect SDK [18], the OpenNI library [75], and the OpenKinect library [76]. The Kinect SDK also provides 3D human skeletal data produced using the method described by Shotton et.al [77]. OpenNI applies NITE [78] – a skeleton generation framework developed as proprietary software by PrimeSense, to generate a similar 3D human skeleton model. Markers are not necessary for structured-light sensors. They are also cheap and can provide 3D skeleton information in real time. On the other hand, 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.

2.1.3 Time-of-Flight (ToF) Sensors

ToF sensors are able to acquire accurate depth data at a high frame rate, by emitting light and measuring the amount of time it takes for that light to return – similar in principle to the established depth sensing technology, such as radar and LiDAR. Comparing to other ToF sensors, the Microsoft Kinect v2 camera offers an affordable alternative to acquire depth data using this technology. In addition, a color camera is integrated into the sensor to provide registered color data. The color-depth data can be accessed by the Kinect SDK 2.0 [79] or the OpenKinect library (using the libfreenect2 driver) [76]. The Kinect v2 camera provides a higher resolution of depth images (512×424) at 30 Hz. Moreover, the camera is able to provide 3D skeleton data by estimating positions of 25 human joints, with an improved tracking accuracy than the Kinect V1 sensor. Similar to the first version, Kinect v2 has a working range from approximately 0.5 to 5 meters.

2.2 3D Joints Estimation and Skeleton Construction

Besides manual human skeletal joint annotation [56], [80], [81], a number of approaches have been designed to automatically construct a skeleton model from perception data. Some of these are based on methods using in RGB imagery, while others take advantage of the extra information available in a depth or RGB-D image. 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 [31], [82], [83]. This section provides a brief review of autonomous skeleton construction methods based on visual data according to their used information. A summary of the reviewed skeleton construction techniques is presented in Table 2.

1. Vicon: http://www.vicon.com/products/software/nexus.
2. OptiTrack: http://www.optitrack.com/products/natnet-sdk.

TABLE 2
Summary of Recent Skeleton Construction Techniques.

Reference	Approach	Input Data	Performance
Shotton et al. [20], [51]	Pixel-by-pixel classification	Single depth image	3D skeleton, 16 joints, real-time, 200 fps
Ye et al. [52]	Motion exemplars	Single depth image	3D skeleton, 38mm accuracy
Jung et al. [53]	Random tree walks	Single depth image	3D skeleton, real-time, 1000fps
Sun et al. [54]	Conditional regression forests	Single depth image	3D skeleton, over 80% average precision
Charles and Everingham [55]	Limb-based shape models	Single depth image	2D skeleton, robust to occlusions
Holt et al. [56]	Decision tree poselets with pictorial structures prior	Single depth image	3D skeleton, only need small amount of training data
Grest et al. [57]	ICP using optimized Jacobian	Single depth image	3D skeleton, over 10 fps
Baak et al. [58]	Matching previous joint positions	Single depth image	3D skeleton, 20 joints, real-time, 100 fps, robust to sensor noise and occlusions
Taylor et al. [59]	Regression to predict correspondences	Single depth image and multiple silhouette images	3D skeleton, 19 joints, real-time, 120fps
Zhu et al. [60]	ICP on individual parts	Depth image sequence	3D skeleton, 10fps, robust to occlusion
Ganapathi et al. [61]	ICP with physical constraints	Depth image sequence	3D skeleton, real-time, 125fps, robust to self collision
Plagemann et al. [62], Ganapathi et al. [25]	Haar features and Bayesian prior	Depth image sequence	3D skeleton, real-time
Zhang et al. [63]	3D non-rigid matching based on MRF deformation model	Depth image sequence	3D skeleton
Schwarz et al. [64]	Geodesic distance & optical flow	Depth and RGB image streams	3D skeleton, 16 joints, robust to occlusions
Wang et al. [65]	Recurrent 2D/3D pose estimation	Single RGB images	3D skeleton, robust to viewpoint changes and occlusions
Fan et al. [66]	Dual-source deep CNN	Single RGB images	2D skeleton, robust to occlusions
Toshev and Szegedy [67]	Deep neural networks	Single RGB images	2D skeleton, robust to appearance variations
Dong et al. [68]	Parselets/grid layout feature	Single RGB images	2D skeleton, robust to occlusions
Akhter and Black [69]	Prior based on joint angle limits	Single RGB images	3D skeleton
Tompson et al. [70]	CNN/Markov random field	Single RGB images	2D skeleton, close to real-time
Elhayek et al. [71]	ConvNet joint detector	Multi-perspective RGB images	2D skeleton, nearly 95% accuracy
Gall et al. [72], Liu et al. [73]	Skeleton tracking and surface estimation	Multi-perspective RGB images	3D skeleton, deal with rapid movements and apparel like skirts

2.2.1 Construction from Depth Imagery

Due to the additional 3D geometric information that depth imagery can provide, many methods are developed to build 3D human skeleton model based on a single depth image or a sequence of depth frames.

Human joint estimation via body parts recognition is one popular approach to construct the skeleton model [20], [51], [53], [54], [55], [56], [62], [64]. A seminal paper by Shotton et al. [20] in 2011 provided an extremely effective skeleton construction algorithm based on body part recognition, that was able to work in real time. 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 the simple 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 5 ms. An extended work was published in [51], with both accuracy and speed improved. Plagemann et al. [62] introduced an approach to recognize body parts using Haar features [84] 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 [25]. Holt et al. [56] proposed Connected Poselets to estimate 3D human pose from depth data. The approach utilizes the idea of poselets [85], which is 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 [86], [87], the method begins with a torso point and connect outwards to body parts. By applying kinematic inference 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 [52], [57], [58], [59], [60], [63]. Several approaches for whole-skeleton matching are based on the Iterative Closest Point (ICP) method [88], 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 [57], [60] 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.

2.2.2 Construction from RGB Imagery

Early approaches and several recent methods based on deep learning focused on 2D or 3D human skeleton construction from traditional RGB or intensity images, typically by identifying human body parts using visual features (e.g., image gradients, deeply learned features, etc.), or matching known poses to a segmented silhouette.

Methods based on a single image: Many algorithms were proposed to construct human skeletal model using a single color or intensity image acquired from a monocular camera [65], [68], [69], [89]. Wang et al. [65] constructs a 3D human skeleton from a single image using a linear combination of known skeletons with physical constraints on limb lengths. Using a 2D pose estimator [89], the algorithm begins with a known 2D pose and a mean 3D pose, and calculates camera parameters from this estimation. The 3D joint positions are recalculated using the estimated parameters, and the camera parameters are updated. The steps continue iteratively until convergence. This approach was demonstrated to be robust to partial occlusions and errors in the 2D estimation. Dong et al. [68] considered the human parsing and pose estimation problems simultaneously. The authors introduced a unified framework based on semantic parts using a tailored And-Or graph. The authors also employed parselets and Mixture of Joint-Group Templates as the representation.

Recently, deep neural networks have proven their ability in human skeleton construction [66], [67], [70]. Toshev and Szegedy [67] employed Deep Neural Networks (DNNs) for human pose estimation. The proposed cascade of DNN regressors obtains pose estimation results with high precision. Fan et al. [66] uses Dual-Source Deep Convolutional Neural Networks (DS-CNNs) for estimating 2D human poses from a single image. This method takes a set of image patches as the input and learns the appearance of each local body part by considering their previous views in the full body, which successfully addresses the joint recognition and localization issue. Tompson et al. [70] proposed a unified learning framework based on deep Convolutional Networks (ConvNets) and Markov Random Fields, which can generate a heat-map to encode a per-pixel likelihood for human joint localization from a single RGB image.

Methods based on multipe images: When multiple images of a human are acquired from different perspectives by a multi-camera system, traditional stereo vision techniques can be employed to estimate depth maps of the human. After obtaining the depth image, human skeleton model can be constructed using the methods based on depth information (Section 2.2.1). Although there exists a commercial solution that uses marker-less multi-camera systems to obtain highly precious skeleton data with 120 frames per second (FPS) and approximately 50-25ms latency [90], computing depth maps is usually slow and often suffers from problems such as failures of correspondence search and noisy depth information. To address these problems, algorithms were also studied to construct human skeleton models directly from the multi-images without calculating the depth image [71], [72], [73]. For example, Gall et al. [72] introduced an approach to fully-automatically estimate the 3D skeleton model from a multi-perspective video sequence, where an articulated template model and silhouettes are obtained from the sequence. Another method was also proposed by Liu et al. [73], which uses a modified global optimization method to handle occlusions.

2.3 Benchmark Datasets With Skeletal Data

In the past five years, a large number of benchmark datasets containing 3D human skeleton data were collected in different scenarios and made available to the public. This section provides a complete review of the datasets as listed in Table 3. We categorize and discuss these datasets according to the type of devices used to acquire the skeleton information.

2.3.1 Datasets Collected Using MoCap Systems

Early 3D human skeleton datasets were usually collected by a MoCap system, which can provide accurate locations of a various number of skeleton joints by tracking the markers attached on human body, typically in indoor environments. The CMU MoCap dataset [91] is one of the earliest resources that consists of a wide variety of human actions, including interaction between two subjects, human locomotion, interaction with uneven terrain, sports, and other human actions. The recent Human3.6M dataset [92] is one of the largest MoCap datasets, which consists of 3.6 million human poses and corresponding images captured by a high-speed MoCap system. It contains activities by 11 professional actors in 17 scenarios: discussion, smoking, taking photo, talking on the phone, etc., as well as provides accurate 3D joint positions and high-resolution videos. The PosePrior dataset [69] is the newest MoCap dataset that includes an extensive variety of human stretching poses performed by trained athletes and gymnasts. Many other MoCap datasets were also released, including the Pictorial Human Spaces [93], CMU Multi-Modal Activity (CMU-MMAC) [94] Berkeley MHAD [95], Standford ToFMCD [25], HumanEva-I [96], and HDM05 MoCap [97] datasets.

2.3.2 Datasets Collected by Structured-Light Cameras

Affordable structured-light cameras are widely used for 3D human skeleton data acquisition. Numerous datasets were collected using the Kinect v1 camera in different scenarios. The MSR Action3D dataset [121], [126] was captured using the Kinect camera at Microsoft Research, which consists of subjects performing American Sign Language gestures and a variety of typical human actions, such as making a phone call or reading a book. The dataset provides RGB, depth, and skeleton information generated by the Kinect v1 camera for each data instance. A large number of approaches used this dataset for evaluation and validation [127]. The MSRC-12 Kinect gesture dataset [120], [128] is one of the largest gesture databases available. Consisting of nearly seven hours of data and over 700,000 frames of a variety of subjects performing different gestures, it provides the pose estimation and other data that was recorded with a Kinect v1 camera. The Cornell Activity Dataset (CAD) includes CAD-60 [117] and CAD-120 [110], which contains 60 and 120 RGB-D videos of human daily activities, respectively. The dataset was recorded by aKinect v1 in different environments, such as an office, bedroom, kitchen, etc. The SBU-Kinect-Interaction dataset [123] contains skeleton

TABLE 3
Publicly Available Benchmark Datasets Providing 3D Human Skeleton Information.

Release Year	Dataset and Reference	Acquisition device	Other Data Source	Scenario
2015	M^2I [98]	Kinect v1	RGB + depth	human daily activities
2015	Multi-View TJU [99]	Kinect v1	RGB + depth	human daily activities
2015	PosePrior [69]	MoCap	color	extreme motions
2015	SYSU 3D HOI [100]	Kinect v1	color + depth	human daily activities
2015	TST Intake Monitoring [101]	Kinect v2 + IMU	depth	human daily activities
2015	TST TUG [102]	Kinect v2 + IMU	depth	human daily activities
2015	UTD-MHAD [103]	Kinect v1 + IMU	RGB + depth	atomic actions
2014	CMU-MAD [104]	Kinect v1	RGB + depth	atomic actions
2014	G3Di [105]	Kinect v1	RGB + depth	gaming
2014	Human3.6M [92]	MoCap	color	movies
2014	Northwestern-UCLA Multiview [106]	Kinect v1	RGB + depth	human daily activities
2014	ORGBD [107]	Kinect v1	RGB + depth	human-object interactions
2014	SPHERE [108]	Kinect	depth	human daily activities
2014	TST Fall Detection [109]	Kinect v2 + IMU	depth	human daily activities
2013	Berkeley MHAD [95]	MoCap	RGB + depth	human daily activities
2013	CAD-120 [110]	Kinect v1	RGB + depth	human daily activities
2013	ChaLearn [111]	Kinect v1	RGB + depth	Italian gestures
2013	KTH Multiview Football [112]	3 cameras	color	professional football activities
2013	MSR Action Pairs [113]	Kinect v1	RGB + depth	activities in pairs
2013	Multiview 3D Event [114]	Kinect v1	RGB + depth	indoor human activities
2013	Pictorial Human Spaces [93]	MoCap	color	human daily activities
2013	UCF-Kinect [115]	Kinect v1	color	human daily activities
2012	3DIG [116]	Kinect v1	color + depth	iconic gestures
2012	CAD-60 [117]	Kinect v1	RGB + depth	human daily activities
2012	Florence 3D-Action [118]	Kinect v1	color	human daily activities
2012	G3D [119]	Kinect v1	RGB + depth	gaming
2012	MSRC-12 Gesture [120]	Kinect v1	N/A	gaming
2012	MSR Daily Activity 3D [121]	Kinect v1	RGB + depth	human daily activities
2012	RGB-D Person Re-identification [122]	Kinect v1	RGB + 3D mesh	person re-identification
2012	SBU-Kinect-Interaction [123]	Kinect v1	RGB + depth	human interaction activities
2012	UT Kinect Action [124]	Kinect v1	RGB + depth	atomic actions
2011	CDC4CV pose [56]	Kinect v1	depth	basic activities
2010	HumanEva [96]	MoCap	color	human daily activities
2010	MSR Action 3D [121]	Kinect v1	depth	gaming
2010	Stanford ToFMCD [25]	MoCap + ToF sensor	depth	human daily activities
2009	TUM kitchen [125]	4 cameras	color	manipulation activities
2008	CMU-MMAC [94]	MoCap	color	cooking in kitchen
2007	HDM05 MoCap [97]	MoCap	color	human daily activities
2001	CMU MoCap [91]	MoCap	N/A	gaming + sports + movies

data of a pair of subjects performing different interaction activities, one person acting and the other reacting. Many other datasets captured using a Kinect v1 camera were also released to the public, including the MSR Daily Activity 3D [121], MSR Action Pairs [113], Online RGBD Action (ORGBD) [107], UTKinect-Action [124], Florence 3D-Action [118], CMU-MAD [104], UTD-MHAD [103], G3D/G3Di [105], [119], SPHERE [108], ChaLearn [111], RGB-D Person Re-identification [122], Northwestern-UCLA Multiview Action 3D [106], Multiview 3D Event [114], CDC4CV pose [56], SBU-Kinect-Interaction [123], UCF-Kinect [115], SYSU 3D Human-Object Interaction [100], Multi-View TJU [99], M^2I [98], and 3D Iconic Gesture [116] datasets. The complete list of human-skeleton datasets are presented in Table 3.

2.3.3 Datasets Collected by Other Techniques

Besides the datasets collected by MoCap or structured-light cameras, additional technologies were also applied to collect datasets containing 3D human skeleton information, such as multiple camera systems, ToF cameras such as the Kinect v2 camera, or even manual annotation.

Due to the low price and improved performance of the Kinect v2 camera, it has become increasingly widely adopted to collect 3D skeleton data. The Telecommunication Systems Team (TST) created a collection of datasets using Kinect v2 ToF cameras, which include three datasets for different purposes. The TST fall detection dataset [109] contains eleven different subjects performing falling activities and activities of daily living in various ways; The TST TUG dataset [102] contains 20 different individuals standing up and walking around; and the TST intake monitoring dataset contains food intake actions performed by 35 subjects [101].

Manual annotation approaches are also widely used to provide skeleton data. The KTH Multiview Football dataset [112] contains images of professional football players during real matches, which is obtained using color sensors from 3 views. There are 14 annotated joints for each frame. Several other skeleton datasets are collected based on manual annotation, including the LSP dataset [81], and the TUM Kitchen dataset [125], etc.

3 Information Modality

Skeleton-based human representations are constructed from various features computed from raw 3D skeletal data, where each feature source is called a *modality*. From the perspective

TABLE 4
Summary of 3D Skeleton-Based Representations Based on Joint Displacement Features.

Notation: In the *feature encoding* column: Concatenation-based encoding, Statistics-based encoding, Bag-of-words encoding. In the *structure and transition* column: Low-level features, Body parts models, Manifolds; In the *feature engineering* column: Hand-crafted features, Dictionary learning, Unsupervised feature learning, Deep learning. In the *representation properties* column: 'T' indicates that temporal information is used in feature extraction; 'VI' stands for View-Invariant; 'ScI' stands for Scale-Invariant; 'SpI' stands for Speed-Invariant; 'OL' stands for OnLine; 'RT' stands for Real-Time.

Reference	Approach	Feature Encoding	Structure and Transition	Feature Engineering	T	VI	ScI	SpI	OL	RT
Hu et al. [100]	JOULE	Ba	Ll	Un	✓	✓	✓			
Wang et al. [106]	Cross View	Ba	Bp	Di	✓	✓				
Wei et al. [114]	4D Interaction	Co	Ll	Hc	✓	✓		✓		
Ellis et al. [115]	Latency Trade-off	Co	Ll	Hc	✓	✓			✓	✓
Wang et al. [121], [129]	Actionlet	Co	Ll	Hc		✓	✓	✓		
Barbosa et al. [122]	Soft-biometrics Feature	Co	Bp	Hc						
Xia et al. [124]	Hist. of 3D Joints	St	Ll	Hc		✓				✓
Yun et al. [123]	Joint-to-Plane Distance	Co	Ll	Hc	✓	✓				✓
Yang and Tian [130], [131]	EigenJoints	Co	Ll	Un	✓	✓	✓		✓	✓
Chen and Koskela [132]	Pairwise Joints	Co	Ll	Hc		✓			✓	✓
Rahmani et al. [133]	Joint Movement Volumes	St	Ll	Hc						✓
Luo et al. [134]	Sparse Coding	Ba	Ll	Di		✓		✓		
Jiang et al. [135]	Hierarchical Skeleton	Ba	Ll	Hc	✓	✓			✓	✓
Yao and Li [136]	2.5D Graph Representation	Ba	Ll	Hc		✓	✓			
Vantigodi and Babu [137]	Variance of Joints	St	Ll	Hc	✓	✓				
Zhao et al. [138]	Motion Templates	Ba	Ll	Di		✓	✓		✓	✓
Yao et al. [139]	Coupled Recognition	Co	Ll	Hc	✓					
Zhang et al. [140]	Star Skeleton	Ba	Ll	Hc	✓	✓		✓	✓	✓
Zou et al. [141]	Key Segment Mining	Ba	Ll	Di	✓	✓	✓			
Kakadiaris and Metaxas [142]	Physics Based Model	Co	Ll	Hc	✓					
Nie et al. [143]	ST Parts	Ba	Bp	Di	✓	✓				
Anirudh et al. [144]	TVSRF Space	Co	Mf	Hc	✓	✓	✓	✓		
Koppula and Saxena [145]	Temporal Relational Features	Co	Ll	Hc	✓					
Wu and Shao [146]	EigenJoints	Co	Ll	De	✓	✓	✓			✓
Kerola et al. [147]	Spectral Graph Skeletons	Co	Ll	Hc	✓	✓	✓			

of information modality, 3D skeleton-based human representations can be classified into four categories, based on joint displacement, orientation, raw position, and combined information. Existing approaches falling in each categories are summarized in detail in Tables 4–7, respectively.

3.1 Displacement-Based Representations

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

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 [121] [129] [130] [132] [136] [138]. Within the same skeleton model obtained at a 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 [121], [129], [130]. Chen and Koskela [132] implemented a similar feature extraction method based on pairwise relative position of skeleton joints with normalization calculated by $\|\boldsymbol{p}_i - \boldsymbol{p}_j\| / \sum_{i \neq j} \|\boldsymbol{p}_i - \boldsymbol{p}_j\|, i \neq j$, which is illustrated in Fig. 10(a).

Another group of joint displacement features extracted from the same frame for skeleton-based representation construction is 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 reference 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. [133] 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. 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. [134] applied similar position information for feature extraction. Since the joint hip center has relatively small motions for most actions, they used that joint as the reference. Lu et al. [124] 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 displacements 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 used in [124] is oriented with the x axis aligned with the direction a person is facing, their

TABLE 5
Summary of 3D Skeleton-Based Representations Based on Joint Orientation Features.

Notation: In the *feature encoding* column: Concatenation-based encoding, Statistics-based encoding, Bag-of-words encoding. In the *structure and transition* column: Low-level features, Body parts models, Manifolds; In the *feature engineering* column: Hand-crafted features, Dictionary learning, Unsupervised feature learning, Deep learning. In the *representation properties* column: 'T' indicates that temporal information is used in feature extraction; 'VI' stands for View-Invariant; 'ScI' stands for Scale-Invariant; 'SpI' stands for Speed-Invariant; 'OL' stands for OnLine; 'RT' stands for Real-Time.

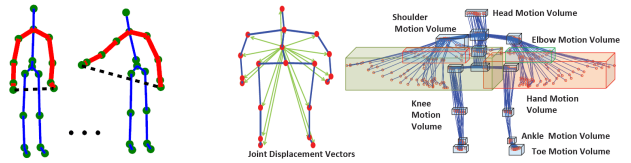
Reference	Approach	Feature Encoding	Structure and Transition	Feature Engineering	T	VI	ScI	SpI	OL	RT
Sung et al. [117]	Orientation Matrix	Co	Ll	Hc		✓	✓	✓		
Fothergill et al. [128]	Joint Angles	Co	Ll	Hc	✓	✓	✓		✓	✓
Gu et al. [148]	Gesture Recognition	Ba	Ll	Di		✓			✓	✓
Sung et al. [149]	Orientation Matrix	Co	Ll	Hc		✓	✓	✓		
Jin and Choi [150]	Pairwise Orientation	St	Ll	Hc		✓	✓	✓	✓	✓
Zhang and Tian [151]	Pairwise Features	St	Ll	Hc		✓			✓	✓
Kapsouras and Nikolaidis [152]	Dynemes Representation	Ba	Ll	Di	✓					
Vantigodi and Radhakrishnan [153]	Meta-cognitive RBF	St	Ll	Hc	✓	✓	✓	✓		
Ohn-Bar and Trivedi [154]	HOG2	Co	Ll	Hc	✓	✓	✓			
Chaudhry et al. [155]	Shape from Neuroscience	Ba	Bp	Di	✓	✓				
Ofli et al. [156]	SMIJ	Co	Ll	Un	✓	✓	✓			
Miranda et al. [157]	Joint Angle	Ba	Ll	Di		✓	✓		✓	✓
Fu and Santello [158]	Hand Kinematics	Co	Ll	Hc					✓	✓
Zhou et al. [159]	4D quaternions	Ba	Ll	Di	✓			✓	✓	✓
Campbell and Bobick [160]	Phase Space	Co	Ll	Hc	✓	✓	✓			
Boubou and Suzuki [161]	HOVV	St	Ll	Hc	✓	✓	✓	✓		✓
Sharaf et al. [162]	Joint angles and velocities	St	Ll	Hc	✓	✓	✓	✓	✓	✓
Salakhutdinov et al. [163]	HD Models	Co	Ll	De	✓	✓	✓	✓		
Parameswaran and Chellappa [164]	ISTs	Co	Ll	Hc	✓	✓	✓	✓		

approach is view invariant.

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 [130], [131] introduced a novel feature based on the position difference of joints, called EigenJoints, which combines three categories of features including static posture, motion, and offset features. In particular, the joint displacement of current frame with respect to the previous frame and initial frame is calculated. Ellis et al. [115] introduced an algorithm to reduce latency for action recognition using the 3D skeleton-based representation that depends on spatial-temporal features computed from the information in three frames: the current frame, the frame collected 10 time steps ago, and the frame collected 30 frames 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 [114]. 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. Hu et al. [100] introduced the joint heterogeneous features learning (JOULE) model through extracting the pose dynamics using skeleton data from a sequence of depth images. A real-time skeleton tracker is used to extract the trajectories of human joints. Then relative positions of each trajectory pair is used to construct features to distinguish different human actions.



(a) Displacement of pairwise joints [132]

(b) Relative joint displacement and joint motion volume features [133]

Fig. 3. Examples of 3D human representations based on joint displacements.

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 [133]. 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 Fig. 10(b). 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.

3.2 Orientation-Based Representations

Another widely used information modality for human representation construction is based on the joint orientations,

TABLE 6
Summary of Representations Based on Raw Position Information.

Notation: In the *feature encoding* column: Concatenation-based encoding, Statistics-based encoding, Bag-of-words encoding. In the *structure and transition* column: Low-level features, Body parts models, Manifolds; In the *feature engineering* column: Hand-crafted features, Dictionary learning, Unsupervised feature learning, Deep learning. In the *representation properties* column: 'T' indicates that temporal information is used in feature extraction; 'VI' stands for View-Invariant; 'ScI' stands for Scale-Invariant; 'SpI' stands for Speed-Invariant; 'OL' stands for OnLine; 'RT' stands for Real-Time.

Methods	Approach	Feature Encoding	Structure and Transition	Feature Engineering	T	VI	ScI	SpI	OL	RT
Du et al. [22]	BRNNs	Co	Bp	De	✓					
Kazemi et al. [112]	Joint Positions	Co	Ll	Hc		✓				
Seidenari et al. [118]	Multi-Part Bag of Poses	Ba	Ll	Di	✓	✓	✓			
Chaaraoui et al. [165]	Evolutionary Joint Selection	Ba	Ll	Di		✓				
Reyes et al. [166]	Vector of Joints	Co	Ll	Hc		✓		✓		
Patsadu et al. [167]	Vector of Joints	Co	Ll	Hc			✓	✓		
Huang and Kitani [168]	Cost Topology	St	Ll	Hc						
Devanne et al. [169]	Motion Units	Co	Mf	Hc		✓				
Wang et al. [170]	Motion Poselets	Ba	Bp	Di		✓				
Wei et al. [171]	Structural Prediction	Co	Ll	Hc		✓		✓		
Gupta et al. [172]	3D Pose w/o Body Parts	Co	Ll	Hc		✓		✓		
Amor et al. [173]	Skeleton's Shape	Co	Mf	Hc		✓	✓	✓		
Sheikh et al. [174]	Action Space	Co	Ll	Hc	✓	✓	✓	✓		
Yilma and Shah [175]	Multiview Geometry	Co	Ll	Hc	✓	✓				
Gong et al. [176]	Structured Time	Co	Mf	Hc	✓	✓		✓		
Rahmani and Mian [177]	Knowledge Transfer	Ba	Ll	Di		✓				
Munsell et al. [178]	Motion Biometrics	St	Ll	Hc	✓	✓				
Lillo et al. [179]	Composable Activities	Ba	Ll	Di	✓	✓	✓			
Wu et al. [180]	Watch-n-Patch	Ba	Ll	Di	✓	✓			✓	✓
Gong and Medioni [181]	Dynamic Manifolds	Ba	Mf	Di	✓	✓		✓		
Han et al. [182]	Hierarchical Manifolds	Ba	Mf	Di	✓	✓	✓	✓		
Slama et al. [183], [184]	Grassmann Manifolds	Ba	Mf	Di	✓	✓		✓	✓	✓
Devanne et al. [185]	Riemannian Manifolds	Co	Mf	Hc	✓	✓	✓	✓	✓	✓
Huang et al. [186]	Shape Tracking	Co	Ll	Hc	✓	✓	✓		✓	✓
Devanne et al. [187]	Riemannian Manifolds	Co	Mf	Hc	✓	✓	✓	✓		
Zhu et al. [188]	RNN with LSTM	Co	Ll	De	✓					
Chen et al. [189]	EnwMi Learning	Ba	Ll	Di	✓	✓	✓			
Hussein et al. [190]	Covariance of 3D Joints	St	Ll	Hc	✓	✓	✓	✓		
Shahroudy et al. [191]	Fourier Temporal Pyramid	Ba	Bp	Un	✓	✓	✓			
Jung and Hong [192]	Elementary Moving Pose	Ba	Ll	Di	✓	✓	✓	✓		
Evangelidis et al. [193]	Skeletal Quad	Co	Ll	Hc	✓	✓	✓			
Azary and Savakis [194]	Grassmann Manifolds	Co	Mf	Hc	✓	✓	✓	✓		
Barnachon et al. [195]	Hist. of Action Poses	St	Ll	Hc					✓	✓
Shahroudy et al. [196]	Feature Fusion	Ba	Bp	Un		✓	✓			

since in general orientation-based features are invariant to human position, body size, and orientation to the camera.

3.2.1 Spatial Orientation of Pairwise Joints

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

A popular orientation-based human representation computes the orientation of each joint to the human centroid in 3D space. For example, Gu et al. [148] collected the skeleton data with fifteen joints and extracted features representing joint angles with respect to the person's torso. Sung et al. [117] 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 [149] based on the orientation matrix.

Another approach is to calculate the orientation of two joints, called relative joint orientations. Jin and Choi [150] 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 the robustness to noise. Zhang and Tian [151] 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. Orientations 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 geometry distance between two joints i and j in 3D space.

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. Campbell and Bobick [160] introduced a mapping from the Cartesian space to the "phase space". By modeling the joint trajectory in the new space, the approach is able to represent a curve that can be easily visualized and quantifiably compared to other motion curves. Boubou and Suzuki [161] described a representation based on the so-called Histogram of Ori-

ented Velocity Vectors (HOVV), which is a histogram of the velocity orientations computed from 19 human joints in a skeleton kinematic model acquired from the Kinect v1 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 movements, this method is able to effectively represent humans with still poses in 3D space in human action recognition applications.

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.

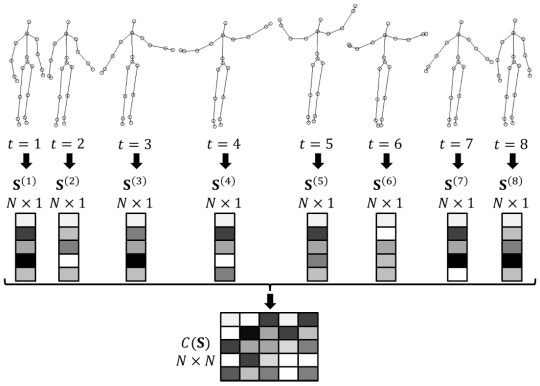


Fig. 4. 3D human representation based on the Cov3DJ descriptor [190].

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 naively encode the sequence with each column containing the flattened joint coordinates obtained at a specific time point. Following this direction, Hussein et al. [190] computed the statistical Covariance of 3D Joints (Cov3DJ) as their features, as illustrated in Fig. 4. Specifically, given K human joints with each joint denoted by $\boldsymbol{p}_i = (x_i, y_i, z_i), i = 1, \ldots, K$, a feature vector is formed to encode the skeleton acquired at time t: $\boldsymbol{S}^{(t)} = [x_1^{(t)}, \ldots, x_K^{(t)}, y_1^{(t)}, \ldots, y_K^{(t)}, z_1^{(t)}, \ldots, z_K^{(t)}]^\top$. 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 $\boldsymbol{S}$. Since not all the joints are the same informative, several methods were proposed to select key joints that are more descriptive [165], [166], [167], [168]. Chaaraoui et al. [165] 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. [166] selected 14 joints in 3D human skeleton models without normalization for feature extraction in gesture recognition applications.

Another group of representation construction techniques utilize the raw joint position information to form a trajectory, and then extract features from this trajectory, which are often called the trajectory-based representation. For example, Wei et al. [171] used a sequence of 3D human skeletal joints to construct joint trajectories, and applied wavelet to encode each temporal joint sequence into features, which is demonstrated in Fig. 5. Gupta et al. [172] proposed a cross-view human representation, which matches trajectory features of videos to MoCap joint trajectories and uses these matches to generate multiple motion projections as features. Junejo et al. [212] used trajectory-based self-similarity matrices (SSMs) to encode humans observed from different views. This method showed great cross-view stability to represent humans in 3D space using MoCap data.

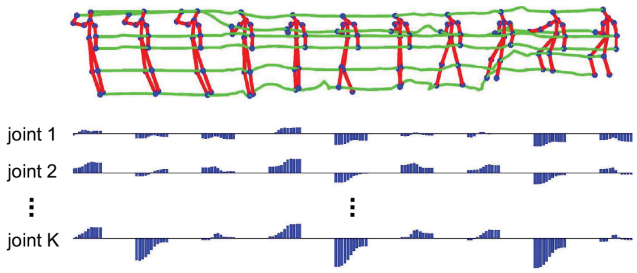


Fig. 5. Trajectory-based representation based on wavelet features [171].

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. [22] 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. [188] used raw 3D joint coordinates as the input to a RNN with Long Short-Term Memory (LSTM) to automatically learn human representations.

3.4 Multi-View Representations

Since multiple information modalities are available, an intuitive way to improve the description power of a human representation is to integrate multiple information sources and build a multi-view 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 [211] 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. [207] 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.

Multi-view spatio-temporal human representations were also actively studied, which is able to integrate both spatial and temporal information and represent human motions in 3D space. Yu et al. [107] 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. [206] implemented a similar representation by incorporating both pairwise joint distances and temporal joint location variations. Zanfir et al. [197] introduced the so-called moving pose feature that integrates raw 3D joint positions as well as first and second derivatives of the joint trajectories, based on the

TABLE 7
Summary of Representations Based on Multi-View Information.

Notation: In the *feature encoding* column: Concatenation-based encoding, Statistics-based encoding, Bag-of-words encoding. In the *structure and transition* column: Low-level features, Body parts models, Manifolds; In the *feature engineering* column: Hand-crafted features, Dictionary learning, Unsupervised feature learning, Deep learning. In the *representation properties* column: 'T' indicates that temporal information is used in feature extraction; 'VI' stands for View-Invariant; 'ScI' stands for Scale-Invariant; 'SpI' stands for Speed-Invariant; 'OL' stands for OnLine; 'RT' stands for Real-Time.

Methods	Approach	Feature Encoding	Structure & Transition	Feature Engineering	T	VI	ScI	SpI	OL	RT
Ganapathi et al. [25]	Kinematic Chain	Co	Ll	Hc						✓
Akhter and Black [69]	Joint Position with Limits	Co	Bp	Hc						✓
Ionescu et al. [92]	MPJPE & MPJAE	Co	Ll	Hc	✓	✓	✓			✓
Marinoiu et al. [93]	Visual Fixation Pattern	Co	Ll	Hc						
Sigal et al. [96]	Parametrization of the Skeleton	Co	Ll	Hc		✓				
Huang et al. [104]	SMMED	Co	Ll	Hc	✓				✓	✓
Bloom et al. [105]	Pose Based Features	Co	Ll	Hc	✓				✓	✓
Yu et al. [107]	Orderlets	Co	Ll	Hc	✓				✓	✓
Paiement et al. [108]	Normalized Joints	Co	Mf	Hc	✓	✓	✓		✓	✓
Koppula and Saxena [110]	Node Feature Map	Co	Ll	Hc	✓				✓	✓
Sadeghipour et al. [116]	Spatial Positions & Directions	Co	Ll	Hc					✓	
Bloom et al. [119]	Dynamic Features	Co	Ll	Hc	✓				✓	✓
Tenorth et al. [125]	Set of Nominal Features	Co	Ll	Hc						
Zanfir et al. [197]	Moving Pose	Ba	Ll	Di	✓	✓				✓
Lehrmann et al. [198]	Vector of Joints	Co	Ll	Hc		✓			✓	✓
Bloom et al. [199]	Dynamic Features	Co	Ll	Hc	✓				✓	✓
Vemulapalli et al. [200]	Lie Group Manifold	Co	Mf	Hc	✓	✓	✓	✓		
Zhang and Parker [201]	BIPOD	St	Bp	Hc	✓	✓	✓		✓	✓
Lv and Nevatia [202]	HMM/Adaboost	Co	Ll	Hc	✓	✓	✓			
Pons-Moll et al. [203]	Posebits	Co	Ll	Hc						
Herda et al. [204]	Quaternions	Co	Bp	Hc		✓	✓		✓	✓
Negin et al. [205]	RDF Kinematic Features	Co	Ll	Un	✓	✓	✓			
Masood et al. [206]	Pairwise Joint Displacement & Temporal Location Variations	Co	Ll	Hc	✓				✓	✓
Gowayyed et al. [207]	HOD	St	Ll	Hc	✓	✓	✓	✓		
Meshry et al. [208]	Angle & Moving Pose	Ba	Ll	Un	✓	✓			✓	✓
Tao and Vidal [209]	Moving Poselets	Ba	Bp	Di	✓					
Eweiwi et al. [210]	Discriminative Action Features	Co	Ll	Un	✓	✓	✓			
Guerra-Filho and Aloimonos [211]	Visuo-motor Primitives	Co	Ll	Hc		✓	✓	✓		

assumption that the speed and acceleration of human joint motions can be described accurately by quadratic functions.

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 without involving temporal information are suitable to address the problems such as pose estimation and gesture recognition. However, if we want the representations to be capable of encoding dynamic human motions, temporal information needs to be integrated. Applications such as activity recognition can benefit from spatio-temporal representations that incorporate time and space information simultaneously. Among space-time human representations, approaches based on joint trajectories can be designed to be insensitive to motion speed invariance.

4 Representation Encoding

Feature encoding is a necessary and important component in representation construction [213], 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. The encoding technique used by each reviewed human representation is summarized in the *Feature Encoding* column in Table 4–7.

4.1 Concatenation-Based Approach

We loosely define feature concatenation as a representation encoding approach, which is a popular method to integrate multiple features into a single feature vector during human

representation construction. 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 [107], [114], [117], [128], [129], [130], [131], [132], [154], [160], [166], [167], [174], [175], [176], [202]. For example, Fothergill et al. [128] encoded the 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 big final feature vector that is fed into a classifier for reasoning. Similarly, Gong et al. [176] directly concatenated 3D joint positions into a 1D vector as a representation at each frame to address the time series segmentation problem.

4.2 Statistics-Based Encoding

Statistics-based encoding is a common but 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 [190], as illustrated in Fig. 4, 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 in [190]. 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. [124] 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 from 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 were also introduced, including Histogram of Joint Position Differences (HJPD) [133], Histogram of Oriented Velocity Vectors (HOVV) [161], and Histogram of Oriented Displacements (HOD) [207], among others [178] [153] [195] [168] [151] [201]. When multi-view skeleton-based features are involved, concatenation-based encoding is usually employed to incorporate multiple histograms into a single final feature vector [201].

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 vector of code occurrence as the final feature vector. Bag-of-words encoding is widely employed by a large number of skeleton-based human representations [100], [106], [118], [134], [135], [138], [141], [143], [148], [152], [155], [157], [165], [170], [177], [179], [180], [181], [183], [197], [208], [209]. 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 algorithm is a popular unsupervised learning method that is commonly used to construct a dictionary. Wang et al. [170] 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 [152] 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.

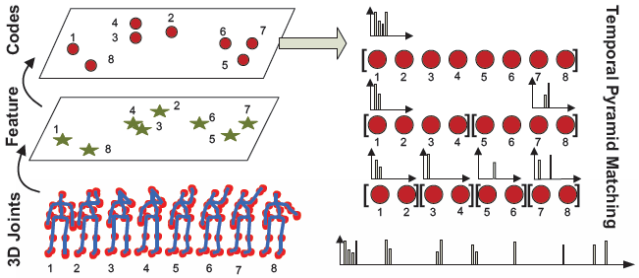


Fig. 6. Dictionary learning based on sparse coding for skeleton-based human representation construction [134].

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. [138] introduced a sparse coding approach regularized by $l_{2,1}$ norm to construct a dictionary of templates from the so-called Structured Streaming Skeletons (SSS) features in a gesture recognition application. Luo et al. [134] 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 Fig. 6.

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, the statistics-based encoding 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 automatically find a good over-complete basis and encode a feature vector using a sparse solution to minimize approximation error. Bag-of-words encoding is also validated to be robust to data noise. However, dictionary construction and feature quantization require additional computation.

5 Structure and Transition

While most of the skeleton-based human representations are based on pure low-level features extracted from the skeleton data in 3D Euclidean space, several works investigated mid-level features or feature transition to other topological space. This section categorizes the reviewed approaches from the structure and transition perspective into three groups: representations using low-level features in Euclidean space, representations using mid-level features based on body parts, and manifold-based representations. The major class of each representation categorized from this perspective is listed in the *Structure and Transition* column in Table 4–7.

5.1 Representations Based on Low-level Features

A simple, straightforward framework to construct skeleton-based representations is to use low-level features computed from 3D skeleton data in Euclidian space, without considering human body structures or applying feature transition. Most of the existing representations fall in this category. The representations can be constructed by single-layer methods, or by approaches with multiple layers.

An example of the single-layer representation construction method is the EigenJoints approach introduced by Yang and Tian [130], [131]. This approach extracts low-level features from skeletal data, such as pairwise joint displacements, and uses Principal Component Analysis (PCA) to perform dimension reduction. Many other existing human representations are also based on low-level skeleton-based features [117], [118], [119], [123], [134], [139], [140], [141], [154], [157], [158], [162], [199] without modeling the hierarchy of the data.

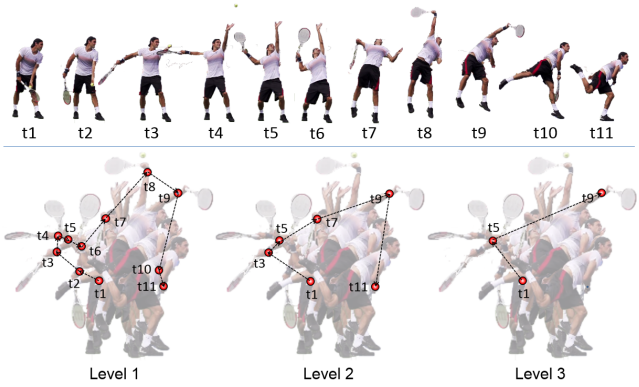


Fig. 7. Temporal pyramid techniques to incorporate multi-layer temporal information for space-time human representation construction based on a sequence of 3D skeleton frames [201].

Several multi-layer techniques were also implemented to create skeleton-based human representations from low-level features. In particular, deep learning approaches inherently consist of multiple layers with the intermediate and output layers encoding different levels of features [214]. The multi-layer deep learning approaches have attracted an increasing attention in recent several years to learn human representations directly from human joint positions [146], [188]. Inspired by the spatial pyramid method [215] to incorporate multi-layer image information, temporal pyramid methods were introduced and used by several skeleton-based human representations to capture the multi-layer information in the time dimension [121], [129], [190], [201], [207]. For example, a temporal pyramid method was proposed by Zhang et al. [201] to capture long-term independencies, as illustrated in Fig. 7. In this example, a temporal sequence of eleven frames is used to represent a tennis-serve motion, and the joint of interest is the right wrist, as denoted by the red dots in Fig 7. 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 joints 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.

5.2 Representations based on Body Part Models

Mid-level features based on body part models are actively studied to construct skeleton-based human representations. Since these mid-level features partially take into account the physical structure of human body, they can usually result in improve discrimination power to represent humans [201], [209].

Wang et al. [170] decomposed a kinematic human body model into five parts, including the left/right arms/legs and torso, each consisting of a set joints. Then, the authors used a data mining technique to obtain a spatiotemporal human representation, by capturing spatial configurations of body parts in one frame (by spatial-part-sets) as well as body part movements across a sequence of frames (by temporal-part-sets), as illustrated in Fig. 8. With this human representation, the approach was able to obtain a hierarchical data that can simultaneously model the correlation and motion of human joints and body parts. Nie et al. [143] implemented a spatial-temporal And-Or graph model to represent humans at three levels including poses, spatiotemporal-parts, and parts. The hierarchical structure of this body model captures the geometric and appearance variations of humans at each frame. Du et al. [22] introduced a deep neural network to create a body part model and investigate the correlation of body parts.

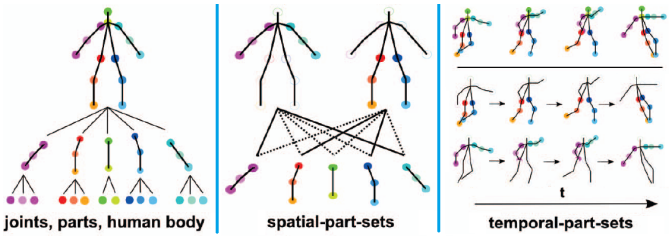


Fig. 8. Spatiotemporal human representations based mid-level features extracted from human body parts [170].

Bio-inspired body part methods were also introduced to extract mid-level features for skeleton-based representation construction, based on body kinematics or human anatomy. Chaudhry et al. [155] implemented a bio-inspired mid-level feature to represent people based on 3D skeleton information through leveraging findings in the area of static shape encoding in the neural pathway of primate cortex [216]. By showing the primates various 3D shapes and measuring the neural response when changing different parameters of the

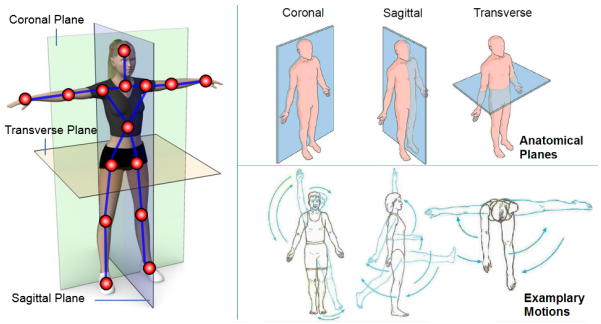


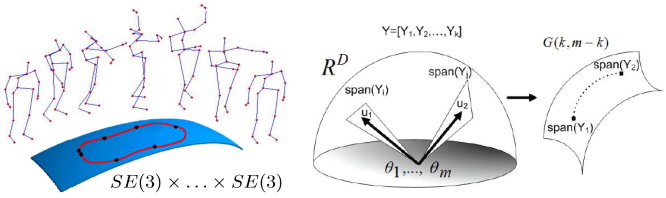
Fig. 9. Representations based on mid-level features extracted from bio-inspired body part models, inspired by human anatomy research [201].

shapes, the primates' internal shape representation can be estimated, which was then applied to extract body parts to construct skeleton-based representations. Zhang and Parker [201] implemented a bio-inspired predictive orientation decomposition (BIPOD) using mid-level features to construct representations of people from 3D skeleton trajectories, which is inspired by biological research in human anatomy. This approach decomposes a human body model into five body parts, and then projects 3D human skeleton trajectories onto three anatomical planes (i.e., coronal, transverse and sagittal planes), as illustrated in Fig. 9. By estimating future skeleton trajectories, the BIPOD representation possesses the ability to predict future human motions.

5.3 Manifold-based Representations

A number of methods in the literature transited the skeleton data in 3D Euclidean space to another topological space (i.e., manifold) in order to process skeleton trajectories as curves within the new space. This category of methods are typically utilize the trajectory-based representation.

Vemulapalli et al. [200] introduced a skeletal representation that was created in the Lie group $SE(3) \times \ldots \times SE(3)$, which is a curved manifold, based on the observation that 3D rigid body motions are members of the space. Using this representation, joint trajectories can be modeled as curves in the Lie group. This manifold-based representation can model 3D geometric relationships between joints using rotations and translations in 3D space. Since analyzing curves in the Lie group is not easy, the approach maps the curves from the Lie group to its Lie algebra, which is a vector space. Gong and Medioni [181] introduced a spatial-temporal manifold and a dynamic manifold warping method, which is an adaptation of dynamic time warping methods for the manifold space. Spatial alignment is also used to deal with variations of viewpoints and body scales. Slama et al. [183] introduced a multi-stage method based on a Grassmann manifold. Body joint trajectories are represented as points on the manifold, and clustered to find a 'control tangent' defined as the mean of a cluster. Then a query human joint trajectory is projected against the tangents to form a final representation. This manifold was also applied by Azary and Savakis [194] to build sparse human representations. Anirudh et al. [144] introduced the transport square-root velocity function (TSRVF) to encode humans in 3D space, which provides an elastic metric to model joint trajectories on Riemannian manifolds. Amor et al. [173] proposed to model the evolution of human skeleton shapes as trajectories on Kendall's shape manifolds, and used a parameterization-invariant metric [217] for aligning, comparing, and modeling skeleton joint trajectories, which can deal with noise caused by large variability of execution rates within and across humans. Devanne et al. [185] introduced a human representation by comparing the similarity between human skeletal joint trajectories in a Riemannian manifold [218].



(a) Lie group [200] (b) Grassmann manifold [194]

Fig. 10. Examples of skeleton-based representations created by transiting joint trajectories in 3D Euclidean space to a manifold.

5.4 Summary

Single or multi-layer human representations based on low-level features directly extract features from 3D skeletal data without considering the physical structure of human body. The kinematic body structure is coarsely encoded by human representations based on mid-level features extracted from body part models, which can capture the relationship of not only joints but also body parts. Manifold-based representations map motion joint trajectories into a new topological space, in the hope of finding a more descriptive representation in the new space. Good performance of all these human representations was reported in the literature.

6 Feature Engineering

Feature engineering is one of the most fundamental research problems in computer vision and machine learning research. Early feature engineering techniques for human representation construction are manual; features are hand-crafted and their importance are manually decided. In recent years, we have been witnessing a clear transition from manual feature engineering to automated feature learning and extraction. In this section, we categorize and analyze the human representations based on 3D skeleton data from the perspective of feature engineering. The feature engineering approach used by each human representation is summarized in the *Feature Engineering* column in Table 4–7.

6.1 Hand-crafted Features

Hand-crafted features are manually designed and constructed to capture certain geometric, statistical, morphological, or other attributes of 3D human skeleton data, which dominated the early skeleton-based feature extraction methods and are still intensively studied in modern research.

Lv and Nevatia [202] decomposed the high dimensional 3D joint space into a set of feature spaces where each of them corresponds to the motion of a single joint or a combination of related multiple joints. Ofli et al. [156] proposed a human

representation called the Sequence of the Most Informative Joints (SMIJ), by selecting a subset of skeletal joints to extract category-dependent features. Zhao et al. [132] described a method of representing humans using the similarity of current and previously seen skeletons in a gesture recognition application. Pons-Moll et al. [203] used qualitative attributes of the 3D skeleton data, called posebits, to estimate human poses, by manually defining features such as joints distance, articulation angle, relative position, etc. Huang et al. [186] proposed to utilize hand-crafted features including skeletal joint positions to locate key frames and track humans from a multi-camera video. In general, the majority of the existing skeleton-based human representations employ hand-crafted features, especially, the methodologies based on histograms and manifolds, as presented by Tables 4–7.

6.2 Representation Learning

In many vision and reasoning tasks, good performance is all about the right representation. Thus, automated learning of skeleton-based features has become highly active in the task of human representation construction based on 3D skeletal data. These skeleton-based representation learning methods can be broadly divided into three groups: dictionary learning, unsupervised feature learning, and deep learning.

6.2.1 Dictionary Learning

Dictionary learning aims at learning a basis set (dictionary) to encode a feature vector as a sparse linear combination of basis elements, as well as to adapt the dictionary to the data in a specific task. Learning a dictionary is the foundation of the bag-of-words encoding. In the literature of 3D skeleton-based representation creation, the k-means algorithm [152], [170] and sparse coding [134], [138] are the most commonly used techniques for dictionary learning. A number of these methods are reviewed in Section 4.3.

6.2.2 Unsupervised Feature Learning

The objective of unsupervised feature learning is to discover low-dimensional features that capture the underlying structure of the input data in a higher dimension. For example, the traditional PCA method is applied for dimension reduction to extract low-dimensional features from raw skeleton features [130], [131], [208]. Negin et al. [205] designed a feature selection method to build human representations from 3D skeletal data. This approach describes humans via a collection of feature time-series computed from the skeletal data, and discriminatively optimizes a random decision forest model over this collection to identify the most effective subset of features in time and space dimensions.

Very recently, several multi-view feature learning methods via sparsity-inducing norms were proposed to integrate different types of features, such as color-depth and skeleton-based features, to produce a compact, informative representation of people. Shahroudy et al. [196] recently developed a multi-view feature learning method to fuse the RGB-D and skeletal information into an integrated set of discriminative features. This approach uses the group-l_1 norm to force that features from the same view can be activated or deactivated together, and applies the $l_{2,1}$ norm to allow a single feature within a deactivated view can be activated. The authors also introduced a multi-modal multi-part human representation based on a hierarchical mixed norm [191], which regularizes structured features of each joint subset and applies sparsity between them. Another heterogenous feature learning algorithm was introduced by Hu et al. [100]. The approach casted joint feature learning as a least-square optimization problem that applies the Frobenius matrix norm as the regularization term to provide a closed-form solution.

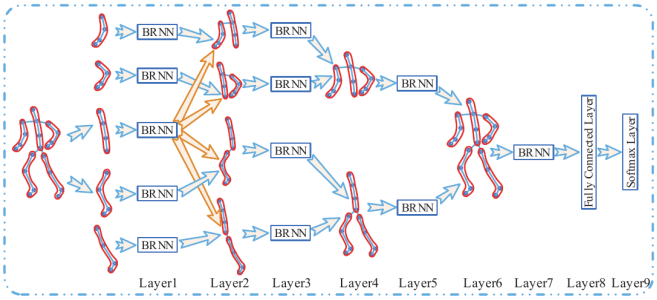


Fig. 11. Hierarchical RNNs for human representation learning based on skeletal joint locations [22].

6.2.3 Deep Learning

While unsupervised feature learning allows for assigning a weight to each feature element, this methodology still relies on manually crafted features as the initial set. Deep learning, on the other hand, attempts to automatically learn a multi-level representation directly from raw data, by exploring a hierarchy of factors that may explain the data. Several such approaches were developed to learn human representations from 3D skeletal joint positions directly acquired by sensors in recent several years. For example, Du et al. [22] proposed an end-to-end hierarchical recurrent neural network (RNN) to construct a skeleton-based human representation. In this method, the whole skeleton is divided into five parts according to human physical structure, and separately fed into five bidirectional RNNs. As the number of layers increases, the representations extracted by the subnets are hierarchically fused to build a higher-level representation, as illustrated in Fig. 11. Zhu et al. [188] introduced a method based on RNNs with the Long Short-Term Memory (LSTM) to automatically learn human representations and model long-term temporal dependencies. In this method, joint positions are used as the input at each time slot to the LST-RNNs that can model the joint co-occurrences to characterize human motions. Wu and Shao [146] proposed to utilize deep belief networks to model the distribution of skeleton joint locations and extract high-level features to represent humans at each frame in 3D space. Salakhutdinov et al. [163] proposed a compositional learning architecture that integrates deep learning models with structured hierarchical Bayesian models. Specifically, this approach learns a hierarchical Dirichlet process (HDP) prior over top-level features in a deep Boltzmann machine (DBM), which simultaneously learns low-level generic features, high-level features that capture the correlation among the low-level features, and a category hierarchy for sharing priors over the high-level features.

6.3 Summary

Hand-crafted features still dominate human representations based on 3D skeletal data in the literature. Although several approaches showed great performance various applications, hand-crafting these manual features typically requires significant domain knowledge and careful parameter tuning. Unsupervised dictionary and feature learning approaches can automatically determine which types of skeleton-based features or templates are more representative, although they typically use hand-craft features as the input. Deep learning, on the other hand, can directly work with the raw skeleton information, and automatically discover and create features. However, the complicated deep learning methods are typically computationally expensive, which currently might not be suitable for online, real-time applications.

7 Future Research Directions

Human representations based on skeleton data can possess several desirable attributes, including the ability to incorporate spatio-temporal information, invariance to variations of viewpoint, human body scale, and motion speed, and real-time, online performance. The characteristics of each review representation are presented in Table 4–7. While significant progress has been achieved on human representations based on 3D skeletal data, there are still numerous research opportunities. Here we briefly summarize some of the prevalent problems and provide possible future research directions.

- *Fusing skeleton data with RGB-D images.* Although 3D skeleton data can be applied to construct descriptive representations of humans, it is incapable to encode texture information, thus cannot effectively represent human-object interaction. Fusing RGB-D information with skeleton data to build a multisensory representation has the potential to address this problem [107], [126], [191] and improve the descriptive power of the existing space-time human representations.
- *General representation construction via cross-training.* A variety of devices can provide skeleton data but with different kinematic models. It is desirable to develop cross-training methods that can utilize skeleton data from different devices to build a general representation that works with different skeleton models [201]. A method of unifying skeleton data to the same format is also useful to integrate available benchmarks dataset and provide sufficient data to modern data-driven, large-scale representation learning methods such as deep learning.
- *Representation of multiple individuals.* Most of existing skeleton-based methods focus on representing a single person, and only a few approaches addressed the representation of a pair of humans [123]. Although multiple sensors including Kinect v2 and MoCap can acquire human skeleton data of multiple individuals simultaneously, no datasets or methods are available to address the essential problem of multi-individual skeleton-based human representations,
- *Protocol for representation evaluation*. There is a strong need of a protocol to benchmark skeleton-based human representations, which must be independent of learning and application-level evaluations. Although the representations have been qualitatively assessed based on their characteristics (e.g., scale-invariance, etc.), a beneficial future direction is to design quantitative evaluation metrics to facilitate evaluating and comparing the human representations.
- *Automated skeleton-based representation learning*. Deep learning and multi-view feature learning have shown very compiling performance in a variety of computer vision and machine learning tasks, but are not well investigated in skeleton-based representation learning and can be a promising future research direction. Moreover, as human skeletal data contains kinematic structures, an interesting problem is how to integrate this structure as a priori in representation learning.
- *Real-time, anywhere skeleton estimation of arbitrary poses.* Skeleton-based human representations heavily rely on the quality of 3D skeleton tracking. A possible future direction is to extract skeleton information of unconventional human poses (e.g., beyond gaming related poses using a Kinect sensor). Another future direction is to reliably extract skeleton information in an outdoor environment using depth data acquired from other sensors such as stereo vision and LiDAR. Although recent works based on deep learning [66], [67], [70] showed promising skeleton tracking results, real-time processing must be ensured for real-word online applications.

8 Conclusion

This paper presents a unique and comprehensive survey of the state-of-the-art space-time human representations based 3D skeleton data that is now widely available. We provide a categorization of the representations from four key perspectives, and compare the pros and cons of the methods in each perspective. A brief overview of 3D skeleton acquisition and construction methods is also included in this paper. Potential future topics are discussed with the hope to facilitate the ongoing research on skeleton-based human representations that keep attracting an increasing attention.

References

[1] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, 2012.

[2] B. Jun, I. Choi, and D. Kim, "Local transform features and hybridization for accurate face and human detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1423–1436, 2013.

[3] E. Demircan, D. Kulic, D. Oetomo, and M. Hayashibe, "Human movement understanding," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 22–24, 2015.

[4] Y. Song, D. Demirdjian, and R. Davis, "Continuous body and hand gesture recognition for natural human-computer interaction," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 1, p. 5, 2012.

[5] S. Green, M. Billinghurst, X. Chen, and G. Chase, "Human-robot collaboration: A literature review and augmented reality approach in design," *International Journal of Advanced Robotic Systems*, pp. 1–18, 2007.

[6] K. Okada, T. Ogura, A. Haneda, J. Fujimoto, F. Gravot, and M. Inaba, "Humanoid motion generation system on HRP2-JSK for daily life environment," in *IEEE International Conference on Mechatronics and Automation*, vol. 4, pp. 1772–1777, 2005.

[7] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Transactions on Automation Science and Engineering*, vol. 6, no. 4, pp. 588–597, 2009.

[8] F. Mondada, M. Bonani, X. Raemy, J. Pugh, C. Cianci, A. Klaptocz, S. Magnenat, J.-C. Zufferey, D. Floreano, and A. Martinoli, "The e-puck, a robot designed for education in engineering," in *Conference on Autonomous Robot Systems and Competitions*, vol. 1, pp. 59–65, 2009.

[9] E. Broadbent, R. Stafford, and B. MacDonald, "Acceptance of healthcare robots for the older population: Review and future directions," *International Journal of Social Robotics*, vol. 1, no. 4, pp. 319–330, 2009.

[10] K. I. Kang, S. Freedman, M. J. Matarić, M. J. Cunningham, and B. Lopez, "A hands-off physical therapy assistance robot for cardiac patients," in *International Conference on Rehabilitation Robotics*, pp. 337–340, 2005.

[11] M. Fujita, "Digital creatures for future entertainment robotics," in *IEEE International Conference on Robotics and Automation*, vol. 1, pp. 801–806, 2000.

[12] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 252–259, 2012.

[13] I. Kviatkovsky, E. Rivlin, and I. Shimshoni, "Online action recognition using covariance of shape and motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[14] N. Siddharth, A. Barbu, and J. M. Siskind, "Seeing what you're told: Sentence-guided activity recognition in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[15] S.-Z. Li, B. Yu, W. Wu, S.-Z. Su, and R.-R. Ji, "Feature learning based on SAE–PCA network for human gesture recognition in RGBD images," *Neurocomputing*, vol. 151, pp. 565–573, 2015.

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

[17] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[18] "Microsoft Kinect." https://dev.windows.com/en-us/kinect.

[19] "ASUS Xtion PRO LIVE." http://www.asus.com/Multimedia/XtionPRO/.

[20] 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*, pp. 1297–1304, 2011.

[21] R. Tobon, *The Mocap Book: A Practical Guide to the Art of Motion Capture*. Foris Force, 2010.

[22] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1110–1118, 2015.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[24] A. O. Bălan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker, "Detailed human shape and pose from images," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

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

[26] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *European Conference on Computer Vision*, 2014.

[27] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *European Conference on Computer Vision*, pp. 872–885, 2012.

[28] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3D pictorial structures for multiple human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1669–1676, 2014.

[29] M. Burenius, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3618–3625, 2013.

[30] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3361–3368, 2011.

[31] H. 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*, pp. 2044–2049, 2011.

[32] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *IEEE International Conference on Computer Vision*, pp. 3456–3462, 2013.

[33] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: Structured estimation of 3D articulated hand posture," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3786–3793, 2014.

[34] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 824–832, 2015.

[35] R. Lun and W. Zhao, "A survey of applications and human motion recognition with Microsoft Kinect," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, no. 5, 2015.

[36] S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled, "A survey of datasets for human gesture recognition," in *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, pp. 337–348, 2014.

[37] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1993–2008, 2013.

[38] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013.

[39] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[40] J. J. LaViola, "3D gestural interaction: The state of the field," *International Scholarly Research Notices*, vol. 2013, 2013.

[41] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pp. 149–187, 2013.

[42] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10873–10888, 2012.

[43] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, p. 16, 2011.

[44] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 1, pp. 13–24, 2010.

[45] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[46] H. Zhou and H. Hu, "Human motion tracking for rehabilitationa survey," *Biomedical Signal Processing and Control*, vol. 3, no. 1, pp. 1–18, 2008.

[47] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[48] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[49] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[50] A. Yao, J. Gall, G. Fanelli, and L. V. Gool, "Does human action recognition benefit from pose estimation?," in *British Machine Vision Conference*, 2011.

[51] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *IEEE International Conference on Computer Vision*, pp. 415–422, 2011.

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

[53] H. Y. Jung, S. Lee, Y. S. Heo, and I. D. Yun, "Random tree walk toward instantaneous 3D human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2467–2474, 2015.

[54] M. Sun, P. Kohli, and J. Shotton, "Conditional regression forests for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3394–3401, 2012.

[55] J. Charles and M. Everingham, "Learning shape models for monocular human pose estimation from the Microsoft Xbox Kinect," in *IEEE International Conference on Computer Vision*, pp. 1202–1208, 2011.

[56] B. Holt, E.-J. Ong, H. Cooper, and R. Bowden, "Putting the pieces together: Connected poselets for human pose estimation," in *Workshops on IEEE International Conference on Computer Vision*, pp. 1196–1201, 2011.

[57] D. Grest, J. Woetzel, and R. Koch, "Nonlinear body pose estimation from depth images," in *Pattern Recognition*, pp. 285–292, 2005.

[58] A. Baak, M. Muller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," *IEEE International Conference on Computer Vision*, pp. 1092–1099, 2011.

[59] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 103–110, 2012.

[60] Y. Zhu, B. Dariush, and K. Fujimura, "Controlled human pose estimation from depth image streams," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[61] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real-time human pose tracking from range data," *European Conference on Computer Vision*, pp. 738–751, 2012.

[62] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *IEEE International Conference on Robotics and Automation*, pp. 3108–3113, 2010.

[63] Q. Zhang, X. Song, X. Shao, R. Shibasaki, and H. Zhao, "Unsupervised skeleton extraction and motion capture from 3D deformable matching," *Neurocomputing*, vol. 100, pp. 170–182, 2013.

[64] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow," *Image and Vision Computing*, vol. 30, no. 3, pp. 217–226, 2012.

[65] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3D human poses from single images," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2369–2376, 2014.

[66] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1347–1355, 2015.

[67] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014.

[68] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan, "Towards unified human parsing and pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 843–850, 2014.

[69] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1446–1455, 2015.

[70] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Annual Conference on Neural Information Processing Systems*, pp. 1799–1807, 2014.

[71] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, "Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3810–3818, 2015.

[72] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1746–1753, 2009.

[73] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of interacting characters using multi-view image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1249–1256, 2011.

[74] "PrimeSense." https://en.wikipedia.org/wiki/PrimeSense.

[75] "The OpenNI Library." http://structure.io/openni.

[76] "The OpenKinect Library." https://github.com/openkinect/libfreenect.

[77] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[78] "NITE." https://wiki.debian.org/PrimeSenseNite.

[79] "Kinect V2 SDK." https://dev.windows.com/en-us/kinect/tools.

[80] M. W. Lee and R. Nevatia, "Dynamic human pose estimation using Markov chain Monte Carlo approach," in *IEEE Workshops on Application of Computer Vision*, pp. 168–175, 2005.

[81] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference*, 2010.

[82] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features," in *Asian Conference on Computer Vision*, pp. 25–38, 2011.

[83] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3838–3843, 2011.

[84] S. Gould, O. Russakovsky, I. Goodfellow, and P. Baumstarck, "STAIR Vision Library," tech. rep., 2011.

[85] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *IEEE International Conference on Computer Vision*, pp. 1365–1372, 2009.

[86] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, no. 1, pp. 67–92, 1973.

[87] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited people detection and articulated pose estimation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[88] P. BESL and N. MCKAY, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[89] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1385–1392, 2011.

[90] A. L. Brooks and A. Czarowicz, "Markerless motion tracking: MS Kinect & Organic Motion OpenStage®," in *International Conference on Disability, Virtual Reality and Associated Technologies*, pp. 435–437, 2012.

[91] "CMU Graphics Lab Motion Capture Database." http://mocap.cs.cmu.edu. NSF EIA-0196217.

[92] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.

[93] E. Marinoiu, D. Papava, and C. Sminchisescu, "Pictorial human spaces: How well do humans perceive a 3D articulated pose?," in *IEEE International Conference on Computer Vision*, pp. 1289–1296, 2013.

[94] "CMU Multi-Modal Activity Database." http://kitchen.cs.cmu.edu. NSF EEEC-0540865.

[95] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *IEEE Workshop on Applications of Computer Vision*, pp. 53–60, 2013.

[96] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[97] M. Muller, T. Roder, M. Clausen, B. Eberhardt, B. Kruger, and A. Weber, "Documentation Mocap Database HDM05," tech. rep., Universität Bonn, June 2007.

[98] N. Xu, A. Liu, W. Nie, Y. Wong, F. Li, and Y. Su, "Multi-modal & multi-view & interactive benchmark dataset for human action recognition," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pp. 1195–1198, ACM, 2015.

[99] A.-A. Liu, N. Xu, Y.-T. Su, H. Lin, T. Hao, and Z.-X. Yang, "Single/multi-view human action recognition via regularized multi-task learning," *Neurocomputing*, vol. 151, pp. 544–553, 2015.

[100] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5344–5352, 2015.

[101] E. Cippitelli, S. Gasparrini, A. De Santis, L. Montanini, L. Raffaeli, E. Gambi, and S. Spinsante, "Comparison of RGB-D mapping solutions for application to food intake monitoring," in *Ambient Assisted Living*, pp. 295–305, 2015.

[102] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, J. Wahsleny, I. Orhany, and T. Lindhy, "Time synchronization and data fusion for RGB-depth cameras and inertial sensors in AAL applications," in *Workshop on IEEE International Conference on Communication*, pp. 265–270, 2015.

[103] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *IEEE International Conference on Image Processing*, 2015.

[104] D. Huang, S. Yao, Y. Wang, and F. De La Torre, "Sequential max-margin event detectors," in *European Conference on Computer Vision*, pp. 410–424, 2014.

[105] V. Bloom, V. Argyriou, and D. Makris, "G3Di: A gaming interaction dataset with a real time detection and evaluation framework," in *Workshops on Computer Vision on European Conference on Computer Vision*, pp. 698–712, 2014.

[106] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649–2656, 2014.

[107] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Asian Conference on Computer Vision*, pp. 50–65, 2014.

[108] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, and M. Mirmehdi, "Online quality assessment of human movement from skeleton data," in *British Machine Vision Conference*, 2014.

[109] S. Gasparrini, E. Cippitelli, S. Spinsante, and E. Gambi, "A depth-based fall detection system using a Kinect® sensor," *Sensors*, vol. 14, no. 2, p. 27562775, 2014.

[110] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[111] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *ACM on International Conference on Multimodal Interaction*, pp. 445–452, 2013.

[112] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forests," in *British Machine Vision Conference*, 2013.

[113] O. Oreifej and Z. Liu, "HON4D: histogram of oriented 4D normals for activity recognition from depth sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2013.

[114] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for event and object recognition," in *IEEE International Conference on Computer Vision*, 2013.

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

[116] A. Sadeghipour, L.-P. Morency, and S. Kopp, "Gesture-based object recognition using histograms of guiding strokes," in *British Machine Vision Conference*, 2012.

[117] J. 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.

[118] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Workshop on IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[119] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *Workshops on IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[120] "MSRC-12 Kinect Gesture Data Set." http://research.microsoft.com/en-us/um/cambridge/projects/msrc12.

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

[122] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *International Workshop on Re-Identification*, 2012.

[123] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Workshops on IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[124] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Workshops on IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[125] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Workshops on IEEE International Conference on Computer Vision*, pp. 1089–1096, 2009.

[126] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Workshops on IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9–14, 2010.

[127] J. R. P. Lopez, A. A. Chaaraoui, and F. F. Revuelta, "A discussion on the validation tests employed to compare human action recognition methods using the MSR Action3D dataset," *arXiv preprint arXiv:1407.7390*, 2014.

[128] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *The SIGCHI Conference on Human Factors in Computing Systems*, pp. 1737–1746, 2012.

[129] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, 2014.

[130] X. Yang and Y. Tian, "EigenJoints-based action recognition using Näive-Bayes-Nearest-Neighbor," in *Workshops on IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[131] X. Yang and Y. Tian, "Effective 3D action recognition using eigenjoints," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2–11, 2014.

[132] X. Chen and M. Koskela, "Online RGB-D gesture recognition with extreme learning machines," in *ACM International Conference on Multimodal Interaction*, pp. 467–474, 2013.

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

[134] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *IEEE International Conference on Computer Vision*, 2013.

[135] X. Jiang, F. Zhong, Q. Peng, and X. Qin, "Online robust action recognition based on a hierarchical model," *The Visual Computer*, vol. 30, no. 9, pp. 1021–1033, 2014.

[136] B. Yao and L. Fei-Fei, "Action recognition with exemplr based 2.5D graph matching," in *European Conference on Computer Vision*, pp. 173–186, 2012.

[137] S. Vantigodi and R. V. Babu, "Real-time human action recognition from motion capture data," in *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pp. 1–4, 2013.

[138] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng, "Online human gesture recognition from motion data streams," in *ACM International Conference on Multimedia*, pp. 23–32, 2013.

[139] A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 16–37, 2012.

[140] Z. Fan, G. Li, L. Haixian, G. Shu, and L. Jinkui, "Star skeleton for human behavior recognition," in *International Conference on Audio, Language and Image Processing*, pp. 1046–1050, 2012.

[141] W. Zou, B. Wang, and R. Zhang, "Human action recognition by mining discriminative segment with novel skeleton joint feature," in *Advances in Multimedia Information Processing*, pp. 517–527, 2013.

[142] I. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1453–1459, 2000.

[143] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1293–1301, 2015.

[144] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3147–3155, 2015.

[145] H. Koppula and A. Saxena, "Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation," in *International Conference on Machine Learning*, pp. 792–800, 2013.

[146] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints action segmentation and recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[147] T. Kerola, N. Inoue, and K. Shinoda, "Spectral graph skeletons for 3D action recognition," in *Asian Conference on Computer Vision*, pp. 417–432, Springer, 2014.

[148] Y. Gu, H. Do, Y. Ou, and W. Sheng, "Human gesture recognition through a Kinect sensor," in *IEEE International Conference on Robotics and Biomimetics*, pp. 1379–1384, 2012.

[149] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Workshops on AAAI Conference on Artificial Intelligence*, 2011.

[150] S.-Y. Jin and H.-J. 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*, pp. 148–159, 2013.

[151] C. Zhang and Y. Tian, "RGB-D camera-based daily living activity recognition," *Journal of Computer Vision and Image Processing*, vol. 2, no. 4, p. 12, 2012.

[152] I. Kapsouras and N. Nikolaidis, "Action recognition on motion capture data using a dynemes and forward differences representation," *Journal of Visual Communication and Image Representation*, vol. 25, no. 6, pp. 1432–1445, 2014.

[153] S. Vantigodi and V. B. Radhakrishnan, "Action recognition from motion capture data using meta-cognitive RBF network classifier," in *International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 1–6, 2014.

[154] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Workshop on IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[155] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3D discriminative skeletal features for human action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

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

[157] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, A. W. Vieira, and M. F. Campos, "Real-time gesture recognition from depth data through key poses learning and decision forests," in *SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 268–275, 2012.

[158] Q. Fu and M. Santello, "Tracking whole hand kinematics using extended Kalman filter," in *Annual International Conference on Engineering in Medicine and Biology Society*, pp. 4606–4609, 2010.

[159] F. Zhou, F. De la Torre, and J. K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 582–596, 2013.

[160] L. Campbell and A. Bobick, "Recognition of human body motion using phase space constraints," in *IEEE International Conference on Computer Vision*, pp. 624–630, 1995.

[161] S. Boubou and E. Suzuki, "Classifying actions based on histogram of oriented velocity vectors," *Journal of Intelligent Information Systems*, vol. 44, no. 1, pp. 49–65, 2015.

[162] A. Sharaf, M. Torki, M. E. Hussein, and M. El-Saban, "Real-time multi-scale action detection from 3D skeleton data," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 998–1005, IEEE, 2015.

[163] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "Learning with hierarchical-deep models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1958–1971, 2013.

[164] V. Parameswaran and R. Chellappa, "View invariance for human action recognition," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 83–101, 2006.

[165] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices," *Expert Systems with Applications*, vol. 41, no. 3, pp. 786–794, 2014.

[166] M. Reyes, G. Domínguez, and S. Escalera, "Feature weighting in dynamic timewarping for gesture recognition in depth data," in *Workshops on IEEE International Conference on Computer Vision*, pp. 1182–1188, 2011.

[167] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," in *International Joint Conference on Computer Science and Software Engineering*, pp. 28–32, 2012.

[168] D.-A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *European Conference on Computer Vision*, pp. 489–504, 2014.

[169] M. Devanne, H. Wannous, P. Pala, S. Berretti, M. Daoudi, and A. Del Bimbo, "Combined shape analysis of human poses and motion units for action segmentation and recognition," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–6, 2015.

[170] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 915–922, 2013.

[171] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu, "Concurrent action detection with structural prediction," in *IEEE International Conference on Computer Vision*, pp. 3136–3143, 2013.

[172] A. Gupta, J. L. Martinez, J. J. Little, and R. J. Woodham, "3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2601–2608, 2014.

[173] B. Amor, J. Su, and A. Srivastava, "Action recognition using rate-invariant analysis of skeletal shape trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 1–13, 2016.

[174] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *IEEE International Conference on Computer Vision*, vol. 1, pp. 144–149, 2005.

[175] A. Yilma and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *IEEE International Conference on Computer Vision*, vol. 1, pp. 150–157, 2005.

[176] D. Gong, G. Medioni, and X. Zhao, "Structured time series analysis for human action segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1414–1427, 2014.

[177] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2458–2466, 2015.

[178] B. C. Munsell, A. Temlyakov, C. Qu, and S. Wang, "Person identification using full-body motion and anthropometric biometrics from Kinect videos," in *European Conference on Computer Vision*, pp. 91–100, 2012.

[179] I. Lillo, A. Soto, and J. C. Niebles, "Discriminative hierarchical modeling of spatio-temporally composable human activities," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[180] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "Watch-n-Patch: Unsupervised understanding of actions and relations," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4362–4370, 2015.

[181] D. Gong and G. Medioni, "Dynamic manifold warping for view invariant action recognition," in *IEEE International Conference on Computer Vision*, 2011.

[182] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia, "Discriminative human action recognition in the learned hierarchical manifold space," *Image and Vision Computing*, vol. 28, no. 5, pp. 836–849, 2010.

[183] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3D action recognition using learning on the grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556–567, 2015.

[184] R. Slama, H. Wannous, and M. Daoudi, "Grassmannian representation of motion depth for 3D human gesture and action recognition," in *International Conference on Pattern Recognition*, pp. 3499–3504, 2014.

[185] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *IEEE Transactions on Cybernetics*, vol. 45, pp. 1340–1352, July 2015.

[186] C.-H. Huang, E. Boyer, N. Navab, and S. Ilic, "Human shape and pose tracking using keyframes," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3446–3453, 2014.

[187] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "Space-time pose representation for 3D human action recognition," in *New Trends in Image Analysis and Processing*, pp. 456–464, 2013.

[188] W. Zhu, C. Lan, J. Xing, Y. Li, L. Shen, W. Zeng, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *AAAI Conference on Artificial Intelligence, to appear*, 2016.

[189] G. Chen, M. Giuliani, D. Clarke, A. Gaschler, and A. Knoll, "Action recognition using ensemble weighted multi-instance learning," in *IEEE International Conference on Robotics and Automation*, pp. 4520–4525, 2014.

[190] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *International Joint Conference on Artificial Intelligence*, pp. 2466–2472, 2013.

[191] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, accepted, 2016.

[192] H.-J. Jung and K.-S. Hong, "Enhanced sequence matching for action recognition from 3D skeletal data," in *Asian Conference on Computer Vision*, pp. 226–240, Springer, 2015.

[193] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *International Conference on Pattern Recognition*, pp. 4513–4518, IEEE, 2014.

[194] S. Azary and A. Savakis, "Grassmannian sparse representations and motion depth surfaces for 3D action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 492–499, 2013.

[195] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Ongoing human action recognition with motion capture," *Pattern Recognition*, vol. 47, no. 1, pp. 238–247, 2014.

[196] A. Shahroudy, G. Wang, and T.-T. Ng, "Multi-modal feature fusion for action recognition in RGB-D sequences," in *International Symposium on Communications, Control and Signal Processing*, 2014.

[197] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *IEEE International Conference on Computer Vision*, pp. 2752–2759, 2013.

[198] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "A non-parametric bayesian network prior of human pose," in *IEEE International Conference on Computer Vision*, pp. 1281–1288, 2013.

[199] V. Bloom, V. Argyriou, and D. Makris, "Dynamic feature selection for online action recognition," in *Human Behavior Understanding*, pp. 64–76, 2013.

[200] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[201] H. Zhang and L. E. Parker, "Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction," in *International Conference on Robotics and Automation*, 2015.

[202] F. Lv and R. Nevatia, "Recognition and segmentation of 3-D human action using HMM and multi-class adaboost," in *European Conference on Computer Vision*, pp. 359–372, 2006.

[203] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn, "Posebits for monocular human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2345–2352, 2014.

[204] L. Herda, R. Urtasun, and P. Fua, "Hierarchical implicit surface joint limits for human body tracking," *Computer Vision and Image Understanding*, vol. 99, no. 2, pp. 189–209, 2005.

[205] F. Negin, F. Ozdemir, C. B. Akgul, K. A. Yuksel, and A. Ercil, "A decision forest based feature selection framework for action recognition from RGB-Depth cameras," in *International Conference on Image Analysis and Recognition*, 2013.

[206] S. Z. Masood, C. Ellis, A. Nagaraja, M. F. Tappen, J. J. L. Jr., and R. Sukthankar, "Measuring and reducing observational latency when recognizing actions," in *Workshop on IEEE International Conference on Computer Vision*, 2011.

[207] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *International Joint Conference on Artificial Intelligence*, 2013.

[208] M. Meshry, M. E. Hussein, and M. 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.

[209] L. Tao and R. Vidal, "Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition," in *Workshops on IEEE International Conference on Computer Vision*, pp. 61–69, 2015.

[210] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall, "Efficient pose-based action recognition," in *Asian Conference on Computer Vision*, pp. 428–443, Springer, 2015.

[211] G. Guerra-Filho and Y. Aloimonos, "Understanding visuo-motor primitives for motion synthesis and analysis," *Computer Animation and Virtual Worlds*, vol. 17, no. 3-4, pp. 207–217, 2006.

[212] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172–185, 2011.

[213] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: A comprehensive study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 493–506, 2014.

[214] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, pp. 818–833, 2014.

[215] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, 2006.

[216] Y. Yamane, E. T. Carlson, K. C. Bowman, Z. Wang, and C. E. Connor, "A neural code for three-dimensional object shape in macaque inferotemporal cortex," *Nature Neuroscience*, vol. 11, no. 11, pp. 1352–1360, 2008.

[217] J. Su, A. Srivastava, F. D. de Souza, and S. Sarkar, "Rate-invariant analysis of trajectories on Riemannian manifolds with application in visual speech recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 620–627, 2014.

[218] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on Pure and Applied Mathematics*, vol. 30, no. 5, pp. 509–541, 1977.