A Survey on Human Action Recognition Using Depth Sensors

Bin Liang and Lihong Zheng
School of Computing and Mathematics, Charles Sturt University
Wagga Wagga, NSW 2650, Australia
{bliang, lzheng}@csu.edu.au

***Abstract*—The recent advent of depth sensors opens up new opportunities to advance human action recognition by providing depth information. Many different approaches have been proposed for human action recognition using depth sensors. The main purpose of this paper is to provide a comprehensive study and an updated review on human action recognition using depth sensors. We give an overview of recent works in this field from the viewpoints of data modalities, feature extraction and classification. In terms of data modalities from depth sensors, recent approaches can be roughly categorized into depth map based and skeleton based approaches. Since depth maps encode 3D shape and appearance information, approaches based on depth maps are suitable for short simple actions and can achieve high performance. In contrast, due to the discriminative power and more concise form of skeletal joints, skeleton based approaches can model more complex actions, even in real time. This paper further provides a summary of the results obtained in the last couple of years on the public datasets. Moreover, we discuss limitations of the state of the art and outline promising directions of research in this area. The review assists in guiding both researchers and practitioners in the selection and development of approaches for human action recognition using depth sensors.**

I. Introduction

We as humans have a remarkable ability to understand human actions purely through visual information. Our lives would be revolutionized if machines could automatically interpret and recognize human actions. Human action recognition has raised considerable interest and has become a widely studied research topic in the computer vision community during the last two decades [1]–[7]. It mainly aims to analyze human motions, and to perceive various kinds of human actions. Based on either RGB data, depth maps, skeletal joints or combination of these modalities, many action recognition approaches have been developed and applied in a large variety of practical applications, such as human-computer interaction, video surveillance, health-care and content-based video retrieval [8].

The recognition of human movement can be performed at various levels of abstraction [4]. Similarities exist among different taxonomies: *gesture*, *action* and *activity*. According to the literature [3], [9], we conceptually give our definitions. *Gestures* are elementary motions of the human body. Basically they are primary meaningful components that can be described at the level of body parts. “Arm stretches” and “leg kicks” are examples of gestures. *Actions* are single-person activities, consisting of multiple gestures that have a temporal ordering. They are characterized as whole-body movements. Examples of actions are “waving”, “walking” and “running”. *Activities* typically contain a number of subsequent actions often performed by several persons, with or without objects. One example of an activity could be “playing basketball” or “using a laptop”.

Nevertheless, there are no clear boundary between *gesture* and *action*. “Gesture recognition” and “action recognition” are always interchangeably used in computer vision and other AI fields [10], [11]. In this paper, we ignore the subtle difference between them, and do not consider “activity recognition” that could introduce contexts such as environment and interactions with objects.

The purpose of this paper is to provide a complete overview of state-of-the-art human action recognition methodologies using depth sensors. We discuss various types of such approaches designed for the recognition from viewpoints of available modalities provided by depth sensors, feature extraction and classification. The previous reviews [6], [7] covered several essential components for the understanding of human motion, such as tracking, detection, and body pose modeling. However, with the rapidly developing technology, the motion analysis techniques were insufficient to illustrate ongoing human actions. Specifically, there is concentration on introducing an up-to-date review of the recent literature and presenting some insights into the studies of the latest action recognition methods using depth sensors. Moreover, the recent methods in this area are divided into two categories, *i.e.,* depth map based and skeleton based methods. Subsequently, corresponding typical approaches in different categories are presented and their distinctness is explicitly revealed. Finally, we present summaries of publicly available datasets for evaluation, and discuss recent research trends in action recognition using depth sensors.

II. Background

A. Challenges

For human action recognition, the common approach is to extract motion features from the video sequences and to predict corresponding action class labels. However, different people perform the same action differently, and even the same person could perform it differently. According to [12], [13], the performance of action recognition could be affected by several sources of variations, *i.e., viewpoint*, *anthropometry*, *execution rate*, and *personal style*. *Viewpoint* illustrates the relationship between people and the camera. The same action, observed from various viewpoints, could lead to very different results. *Anthropometry* encodes the physical attributes of human body sizes, and it is movement invariant. *Execution rate* is related to

978-1-4673-6795-0/15/$31.00 ©2015 IEEE

the speed of action performance or frame rates. The execution rate at which the action is performed has an important effect on the temporal information of an action, especially when temporal features are used. *Personal style* is dependent on each single person, since different people could choose their own way to perform the same action. A robust human action recognition approach should be able to handle these variations. With the increasing number of action classes, the task of automatically recognizing human actions will be more challenging.

Additionally, many other potential difficulties exist in captured data. Robust action recognition approaches require large amounts of training data, as the lack of training data often implies a lack of high-quality discriminative information. Typically, this will result in poor performances. Furthermore, the environment where actions are performed is an important source of variations for data recording, such as different lighting conditions, camera settings, and partial body occlusions. Therefore, most approaches are restricted to certain scenarios.

Recently, the launch of depth sensors provides the possibility of depth capture, which facilitates a number of visual recognition tasks including human action recognition [14]. Recognizing human actions using depth sensors overcomes some of the limitations of the previous works using conventional cameras.

B. Advances

In the past, research on human action recognition has concentrated on learning and recognizing human actions from video sequences captured by conventional cameras [1], [4], [5], [9]. Data captured by conventional cameras encode rich texture and color information, which is useful for image processing. However, they have the following limitations: Data captured by conventional cameras are very sensitive to illumination changes. Robust background subtraction is a difficult task because the foreground objects, by necessity, blend with the background. Capturing 3D action movements from conventional cameras has become a much more challenging problem. As human actions are performed in 3D space, access to 3D information plays an important role in action recognition.

In contrast to conventional cameras, there have been vision technologies that can capture depth information. Depth information has long been regarded as an essential part of successful action recognition [15]. Compared with video sequences captured by conventional cameras, depth data have shown several advantages in the context of action recognition. Depth data can provide 3D structural information so that the motion information of actions can be more discriminative. Moreover, depth data are insensitive to illumination changes, and the huge color and texture variability induced by clothing, hair, skin and background could be reduced.

C. Depth Sensing

Nowadays, there are a large number of depth sensors available in the market. Different types of depth sensors operate using different techniques. We classify depth sensors into three main categories based on [6]: *stereo triangulation*, *time-of-flight*, and *structured light*. A stereo camera system based on *stereo triangulation* infers the 3D structure of a scene from two or more images from various viewpoints [16]. Due to the complexity of geometry calculation and sensitivity to light changes, this kind of depth sensor is still difficult to set up real-time applications. A *time-of-flight* (ToF) camera estimates the distance to an object surface using active light pulses from a single camera. ToF cameras are able to achieve high frame rates, which makes them suitable for real-time applications. However, the relatively high price of ToF cameras is a major practical issue. By illuminating the scene with a specially designed light pattern, *structured light*, depth can be estimated using only a single image of the reflected light. Microsoft released the first-generation Kinect in 2010, and the second-generation Kinect in 2014, both of which rely on the technique of structured light.

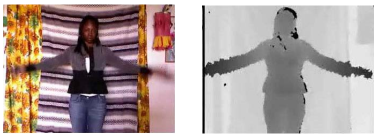
The introduction of cost-effective depth sensor, Kinect, provides a new possibility to address difficult issues we mentioned in Section II-A in the computer vision field. In particular, a growing amount of research has focused on recognizing human actions using Kinect rather than other depth sensors due to its immense popularity. Therefore, most of the work reviewed in this paper has been carried out using Kinect as the main depth sensor.

III. Data Modalities

With the development of computing ability and the improvement of sensor techniques, a large number of methods have been proposed to address action recognition using various data modalities from depth sensors. Depth maps and skeletal joints are two most commonly used data modalities in this area. These data modalities not only facilitate a rather powerful human motion capturing technique, but also make it possible to efficiently model human movements.

A. Depth Maps

The introduction of the depth sensors greatly extends the ability of computer systems to sense the 3D visual world and capture low-level visual information. Intrinsically different from RGB data, pixels in a depth map encode the distance information of a scene rather than a measure of the intensity of color. Texture and color information from depth maps is much less than that in RGB data. Fig. 1 shows RGB image and corresponding depth image from one action sequence in ChaLearn Gesture Dataset [17]. Working in low light environment, and being color and texture invariant, depth maps offer several advantages over RGB data as mentioned earlier in Section II-B. In a depth map, the depth silhouettes of a object can usually be extracted more easily and accurately. However, flicker noise and occlusions in 3D silhouettes could affect the performance of action recognition negatively. Hence, many robust depth map based features are proposed to address these issues, which will be reviewed in Section IV-A1.



(a) RGB image (b) Depth image

Fig. 1. Comparison between RGB image and depth image.

B. Skeletal Joints

Additionally, another data modality, skeletal joints, is also accessible. Skeletal joints encode 3D human joint positions for each frame in real-time. Since the movements of the human skeleton can distinguish many actions, exploiting skeleton data for action recognition is a promising direction. Compared with modeling the skeleton structures from RGB data, the depth information makes the modeling more feasible and stable. Several algorithms have been proposed and applied to model the skeleton from the depth data [18]–[20]. The basic idea underlying these methods is to segment the depth data of the human body into multiple parts with dense probabilistic labeling. The segmentation of the body parts can be considered as a classification task for each pixel in the depth data. The 3D joint positions are obtained based on the spatial modes of the inferred per-pixel distribution. The first-generation Kinect provides 20 joints for each video frame, while the second-generation Kinect allows up to 25 joints. The 3D human joint positions output by Kinect are usually noisy when self-occlusions or object occlusions occur, or when the camera is facing the side of the human subject. As a consequence, directly exploiting skeletal joints usually does not provide encouraging results. It is necessary to develop skeleton based features that are robust to noise and occlusions. These approaches will be discussed in Section IV-A2.

IV. Action Recognition

Due to the wide range of applications for action recognition, researchers have been actively studying this topic and have achieved promising results. Some excellent surveys have been published based on conventional cameras [1], [2], [4], [9], but there has been relatively little review literature devoted based on depth sensors. In this section, we review the state-of-the-art approaches over the past few years for action recognition using depth sensors. The overall framework of action recognition using depth sensors is illustrated in Fig. 2. According to the modalities from depth sensors, various remarkable feature extraction and representation approaches can be classified into two categories: *depth map based* and *skeleton based* approaches. Furthermore, each category is further divided into two types depending on how they model human actions: *space-time features* and *sequential features*. *Space-time features* are extracted by representing an action sequence as a space-time volume, while *sequential features* are based on frame-level or subsequence-level information by considering an action as a sequence of temporal observations.

A. Feature Extraction and Representation

1) Depth Map Based Features: Depth maps provide complementary information of the 3D body shape and appearance to color data. Many algorithms have been proposed to recognize actions using features from depth maps.

Depth Map Based Space-Time Features mainly utilize features, either local or global, from the space-time volume. It is intuitive to treat depth maps as gray images and extract 2D video features. The widely used 2D video features include SITF [21], HOG [22], HOF [23], STIP [23], and kernel descriptors [24]. Those features are shown good performance for 3D object recognition using depth silhouettes in RGB-D object dataset [25]. However, depth silhouette encode 3D shapes and geometric relationships inherently. Oreifej and Liu [26] proposed to describe the depth sequence using a histogram capturing the distribution of the surface normal orientation in the 4D space of time (HON4D), depth, and spatial coordinates. The HON4D features treat a depth map sequence as a 4D spatio-temporal shape, compute a 4D normal for each point on this shape, and construct a histogram of the 4D normal vectors. It is able to capture the observed changing structure. Yang and Tian [27] proposed to subdivide a depth video into a set of space-time grids, and then adopted a novel scheme of aggregating the low-level polynormals into the super normal vector (SNV). Additionally, an adaptive spatial-temporal pyramid was introduced to capture the spatial layout and temporal order in a global way. This is shown to be better adapted to retain the spatial and temporal orders. Motivated by how 3D computer (or human) vision effectively performs geometric reasoning based on the spatial configuration, Lu [28] proposed binary range-sample features from depth maps. It is based on τ tests and achieves reasonable invariance with respect to possible changes in scale, viewpoint, and background. Due to the binary property of the range-sample features, the speed and storage efficiency is impressive. Wang *et al.* [29] presented semi-local features called Random Occupancy Pattern (ROP) features for the purpose of dealing with occlusions. The ROP features are extracted from randomly sampled 4D subvolumes with different sizes and at different locations. Since ROP features are extracted at a larger scale and they only encode information of the most discriminative regions, ROP features are robust to noise and occlusions.

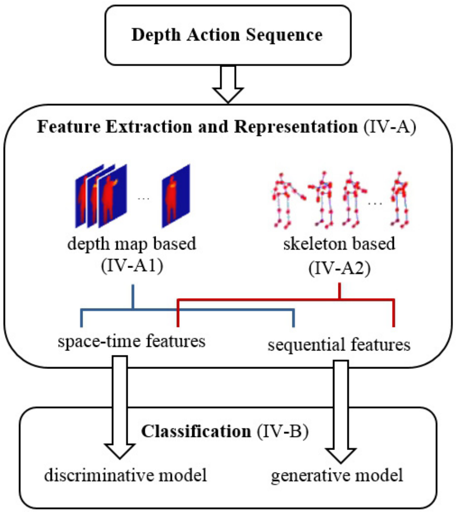


Fig. 2. Framework of action recognition using depth sensors. The framework mainly consists of two parts: feature extraction and representation in Section IV-A and classification in Section IV-B.

Inspired by the success of template matching methods (*e.g.,* MHI [30]) on 2D video sequences, many approaches have been proposed to transform the problem from 3D to 2D, and to perform recognition on projected 2D planes. Yang *et al.* [31] developed Depth Motion Maps (DMM) to capture the aggregated temporal motion energies. The 3D silhouettes are projected onto three pre-defined orthogonal Cartesian planes and then normalized. The DMM-HOG descriptors are constructed by concatenating HOG features from summed binary maps on each plane. Liang and Zheng [32] proposed a 3DMTM-PHOG descriptor to perform human action recognition on depth sequences. The 3D Motion Trail Model (3DMTM) is generated through the entire depth video sequence to encode additional motion information from three projected orthogonal planes. By adding pyramid representation, 3DMTM-PHOG descriptor is able to represent the 3DMTM in different degree of details according to the selected pyramid levels.

To briefly summarize, depth map based space-time features are extracted by considering each action sequence as a 4D volume along spatial (x, y, z) and temporal t direction. The sequence can be processed either as a whole, or as a set of local feature points. The features based on 2D projected planes have shown promising results. Nevertheless, the major limitation of these features lies in that they are highly dependent on very large motions, which means that subtle motions will decrease the performance.

Depth Map Based Sequential Features are extracted through explicitly modeling temporal dynamics from depth sequences.

Bag-of-words models have been previously used in text recognition and object recognition. For action recognition, local features are extracted from the spatial and temporal space. Li *et al.* [33] employed an action graph approach on a bag of 3D points to model actions. Each bag of points extracted from a 3D silhouette represents a salient posture which is one node in an expandable graphical model of actions. However, this method is view-dependent. Moreover, due to noise and occlusions in the depth maps, the projections of silhouettes may not be reliable. To address these issues, Vieira *et al.* [34] proposed a new representation for 3D action recognition, named Space-Time Occupancy Pattens (STOP). It roughly characterizes the 4D space-time patterns of human actions by partitioning the 4D video volume into 4D space-time cells, and aggregating the occupancy information in each cell. STOP descriptors leverage the spatial and temporal contextual information while allowing for intra-class variations. Jalal *et al.* [35] utilized depth silhouettes and $\Re$ transformation to continuously recognize human actions in an indoor environment. $\Re$ transformation is applied on the depth silhouettes to compute a 2D angular projection map of an action silhouette. PCA and LDA are then applied to extract robust features from the $\Re$ transformed profiles of depth silhouettes. Malgireddy *et al.* [36] implemented HOG and HOF features to obtain a dense population of descriptors. The visual words clustered from descriptors are then re-expressed in terms of topics so that each frame can be seen to be made up of visual words which belong to different topics through LDA process. LDA is used here to reduce the size and refine the meaning of the observable features within each frame.

In contrast to depth map based space-time features, relative less work has been presented for action recognition using depth map based sequential features at the time of writing. Analyzing human motions is difficult because there exist great varieties. Directly applying the sequential features for 2D video may not be appropriate, so depth map specific techniques should be considered. This leaves many possibilities for future research.

2) Skeleton Based Features: Skeleton based features are typically easier to extract and require less computational cost and memory than depth map based features.

Skeleton Based Space-Time Features are usually extracted from the skeletal joints to encode temporal motion information. Yang *et al.* [37] developed the EigenJoints features from skeleton data of action sequences. Eigenjoint employs position differences between joints to represent human actions. PCA is applied to the concatenated feature vector within the current frame and consecutive frames to extract "EigenJoints", which encode the most important human pose information for action recognition. The work by Wang *et al.* [38] utilized both skeleton and point cloud information. They proposed to extract Local Occupancy Patterns (LOP features) at each joint and learn an actionlet ensemble model to represent each action and to capture the intra-class invariance. The features are robust to noise, invariant to translational and temporal misalignments. Luo *et al.* [39] proposed a new discriminative dictionary learning algorithm (DL-GSGC) that incorporated both group sparsity and geometry constraints to better represent skeleton features. In addition, the temporal pyramid matching method was applied on each action sequence to keep the temporal information in the representation.

Most of the previous skeleton based space-time features use either the joint locations or the joint angles to represent human actions. Recently, Vemulapalli *et al.* [40] proposed a new skeletal representation that explicitly models the 3D geometric relationships between various body parts using rotations and translations in 3D space. They represented a human skeleton as a point in the Lie group so that human actions are modeled as curves in the Lie group. The experimental results have shown the proposed representation based on Lie group performs better than many existing skeleton representations.

Skeleton Based Sequential Features are typically extracted from skeletal joints within frames or subsequences to model temporal dynamics of actions. Skeleton based sequential features are able to offer low-latency responses that would allow the rapid identification of an action long before it ends.

Xia *et al.* [41] proposed a feature called Histogram of 3D Joint Locations (HOJ3D) using modified spherical coordinates. HOJ3D is able to essentially encode spatial occupancy information relative to the skeleton root, *i.e.,* hip center. The main advantage is the real-time performance. In order to continuously recognize human actions from unsegmented sequences, Zhao *et al.* [13] proposed an effective and efficient framework for online human action recognition. In addition, their proposed Structured Streaming Skeletons (SSS) feature is able to handle intra-class variations including viewpoint, anthropometry, execution rate, and personal style. To cope with unsegmented action sequences, Zanfir *et al.* [42] proposed a fast, simple, yet powerful non-parametric Moving Pose (MP) descriptor for low-latency human action recognition. The proposed MP descriptor is a novel frame-based dynamic representation. It captures not only the 3D body pose but also differential properties like the speed and acceleration of the human body joints within a short time window around the current frame. More recently, Wu and Shao [10] proposed a hierarchical dynamic framework based on high-level skeletal joints features to segment and recognize actions simultaneously. They make feature extraction from skeleton data an implicit approach by deep belief networks. The features themselves are discovered by building a multi-layer generative model of much richer information in the skeletal joints configurations.

TABLE I. SUMMARY OF THE WORK DISCUSSED IN THIS SURVEY.

Approach	Modality	Space-Time	Sequential	Classifier	VI^a	RT^b
[26]	depth maps	HON4D		SVM		
[27]	depth maps	SNV		SVM		
[28]	depth maps	Range-Sample Feature		SVM	✓	✓
[29]	depth maps	ROP		SVM		
[31]	depth maps	DMM-HOG		SVM		
[32]	depth maps	3DMTM-PHOG		SVM		
[33]	depth maps		Bag-of-3D-Points	Action Graph		
[34]	depth maps		STOP	Action Graph		
[35]	depth maps		$\Re$ Transformation	HMM		
[36]	depth maps		HOG, HOF	mcHMM		
[37]	skeletal joints	EigenJoints		NBNN		
[38]	skeletal joints	Actionlet Ensemble		MKL SVM	✓	
[39]	skeletal joints	DL-GSGC		BoVW + SVM		
[40]	skeletal joints	Lie Group Based Feature		SVM		
[41]	skeletal joints		HOJ3D	HMM	✓	✓
[13]	skeletal joints		SSS Feature	Least Square Regression	✓	✓
[42]	skeletal joints		MP Descriptor	Modified kNN		✓
[10]	skeletal joints		High-level Skeleton Feature	HMM		✓

^{a}viewpoint invariant
^{b}real-time

B. Classification

Constructing a good feature representation for depth sequences is the first step for a human action recognition framework. The framework also requires an effective machine learning algorithms to understand and interpret the semantic meanings of human actions. This subsection will give a brief overview of the classification for human action recognition using depth sensors after feature extraction and representation.

Space-time features represent an action sequence as a space-time volume, in which local or global features are extracted for discriminative models (*e.g.,* SVM). Direct classification is the simplest classification. If robust space-time features are extracted from action sequences, a classification model can be learned by directly applying discriminative machine learning algorithms. SVM is widely applied in the reviewed space-time features, *i.e.,* ROP [29], HON4D [26], DMM-HOG [31], 3DMTM-PHOG [32], SNV [27], range-sample feature [28], actinolet ensemble [38] and Lie group based feature [40]. The main advantages of direct classification are easy implementation and discriminative power. However, this is only feasible when there is little noise in the depth and skeleton data, and the actions have relatively simple temporal structures. To address these issues, some approaches perform classification using Bag-of-Visual-Words (BoVW) framework, such as DL-GSGC [39]. BoVW framework can build action classifiers that are robust to variations in subject location, background, and action speed. Its performance, nevertheless, is largely limited by the discriminative power of local descriptors, and it fails to capture the temporal structure of the actions. Nearest neighbor-based method can also be used for action recognition, such as EigenJoints [37].

On the other hand, modeling the temporal structure is essential to successfully represent complex human actions. Sequential features view an action as a sequence of temporal observations. Local features extracted from data of each temporal observation can be used for a generative model. Action graph approach used for bag-of-3D-points [33] and STOP features [34] models an action as an acyclic directed graph, where every node represents a hidden posture state and every edge represents a transition between the states. Another generative model, HMM-based approach, is also utilized to model the temporal dynamics of sequential features, such as $\Re$ Transformation based feature [35], HOJ3D [41] and high-level skeleton feature [10]. Some other approaches have also been proposed for sequential features. Classic least square regression was used in [13] to transfer SSS features to gesture labels. In [42], kNN classifier was modified by considering both the temporal location as well as the discriminative power of MP descriptor.

C. Summary

Table I gives an overall summary of the mentioned approaches using depth maps and skeletal joints. The columns labeled VI and RT indicate whether the paper claims viewpoint invariant or real-time performance, which will be important for some practical applications. From the table, it is noted that there are more skeleton based approaches with viewpoint invariant and real-time performance than depth map based approaches. Depth map based features are suitable for short simple actions such as walking, running and arm waving. It is challenging to address the issues of complex actions, especially with the presence of serve self-occlusions. Furthermore, the depth map in fact only gives the 3D silhouettes of the object facing the camera, so the depth map based features are usually view-dependent. On the other hand, skeleton based features can model more complex human actions. Apparently, skeletal joints are showing more discriminative power and require less computational costs to process than depth maps. This makes it possible to perform real-time continuous action recognition, even along with temporal segmentation. The limitation of the skeletal feature is that it does not encode information of the surroundings.

In addition, sequential features have shown their better capability to handle real-time issues than space-time features. Therefore, there is still a wide gap that remains between existing capabilities of the space-time features and the needs of real-world applications. The gap is particularly meaningful when comparing action recognition using sequential features. Many improvements can be leveraged from sequential features for the design of better space-time features.

V. Datasets and Results

A. Datasets

Since depth sensors, *e.g.,* Kinect, have attracted attention from researchers in many areas, several research groups have built human action datasets by depth sensors using various settings for different purposes. These datasets have been made publicly available. Table II summarizes the currently available human action datasets recorded by depth sensors. We only summarize action datasets recorded by depth sensors, and some other activity datasets are not covered here.

In these datasets, MSR datasets ([26], [29], [33]) are more popular and widely used. This is probably because they have the comprehensive settings and relative simple actions with constrained environmental parameters. By providing unsegmented action sequences along with background, the UTKinect-Action dataset [41] is more competitive. The three Chalearn datasets ([17], [43], [44]) seem to remain challenging due to their one-short learning, unsegmented sequences, and huge number of sequences. However, there are a relatively small number of classes in theses datasets compared to the current trends in image classification which could provide thousands of classes.

B. Results

Since some surveys have summarized results on Chalearn datasets, we omit the results on these datasets. Interested readers are referred to [17], [43], [44] for more details. Here we summarize experimental results on MSR datasets ([26], [29], [33]) and UTKinect-Action dataset [41] due to their popularities. Experimental results of the approaches on these datasets are shown in Table III. As mentioned earlier, MSR Action 3D dataset provides comprehensive experiments for evaluations. One setting is the division of whole classes of actions. The 20 classes of actions in the dataset are further divided into three subsets, each of which is deliberately constructed so that similar actions are included within the same subset. Another setting is the different number of training samples. For each subset, there are three different tests, *i.e.,* Test One (T1), Test Two (T2), and Cross Subject Test (CST). T1 and T2 choose 1/3 and 2/3 samples for training respectively. CST uses the cross subject setting.

The highest records published on these datasets are 96.70%, 94.74%, 97.08% and 98.89%, respectively. The framework proposed in [39] has achieved 96.70% on cross subject test in MSR Action 3D dataset. Although the skeleton based space-time features used in this work are easy to generate, the proposed discriminative class-specific dictionary learning algorithm (DL-GSGC) is able to achieve state-of-the-art performance by using these features. DL-GSGC incorporates both group sparsity and geometry constraints. Since the geometry relationship among local features is considered during the process of dictionary learning, the features from the same class with high similarity are forced to have similar coefficients. This process improves the discriminative power of the recognition performance. In order to capture the local motion and geometry cues, SNV [27] was proposed and achieved highest accuracies in both MSR Gesture 3D dataset and MSR Action Pairs dataset. This is because polynormals in SNV obtain more discriminative local motion and shape information than individual normals. Moreover, the aggregation scheme, the spatial average pooling and temporal max pooling of weighted difference feature vectors, is more representative. Inspired by the observation that the relative geometry between various body parts provides a more meaningful description than their absolute locations, [40] explicitly models the relative 3D geometry using Lie groups and obtains the best performance on UTKinect-Action dataset. These state-of-the-art approaches are all based on space-time features and utilize SVM as the classifier. Furthermore, their performances demonstrate that the 3D geometry information is indispensable for the robust recognition using depth sensors.

These promising performances have shown that current action recognition approaches can work efficiently under constrained settings where the clear background, segmented sequences and well-positioned subjects are provided. For an analysis of approaches for real-word applications, besides the accuracy of measurement, it is imperative that the needs of temporal segmentation, real-time performance, and view invariance should be considered as well. Future works for real-world applications may better explore this aspect to meet the aforementioned needs with the help of depth sensors.

VI. Discussion and Future Directions

The advent of depth sensors has facilitated the task of human action recognition. An increasing number of techniques have employed depth maps and skeletal joints for action recognition. Researchers of computer vision are exploring an extended research field with more potential applications. The survey presented here provides an overview of this emerging field.

This survey demonstrates that contemporary action recognition approaches using depth sensors can now perform very well in controlled settings. Unconstrained action recognition in real-world, however, remains a challenging problem. Most of the state-of-the-art methods are currently not suitable for practical applications for several reasons. Firstly, due to the lack of large and realistic action datasets, most current approaches are limited to controlled action samples with few action classes. In contrast, the recent development of image classification can use huge datasets for learning and prediction, which could provide thousands of classes and millions of photo samples. There is one more point, we should touch on, that it is promising to consider real-time performance besides the high

TABLE II. HUMAN ACTION DATASETS RECORDED BY DEPTH SENSORS.

Dataset	Year	#cls.[a]	#seq.[b]	Modalities	Setting	Technical details	Description
MSR Action3D dataset [33]	2010	20	567	depth maps, skeletal joints	10 subjects, 2 or 3 repetitions	Resolution: 320×240; Without background; Segmented sequences	Actions are chosen in the context of gaming; Full body
Chalearn Gesture dataset [17]	2011	8∼12	23500	RGB data, depth maps	20 subjects; Only one labeled example of one gesture	Resolution: 320×240; With background; Unsegmented sequences	Mainly hand and arm gestures; One-shot learning
MSR Gesture3D dataset [29]	2012	12	336	depth maps	10 subjects, 2 or 3 repetitions	Various resolutions; With background; Segmented sequences	Dynamic hand gestures from American Sign Language (ASL)
UTKinect-Action dataset [41]	2012	10	200	RGB data, depth maps, skeletal joints	10 subjects; 2 repetitions	RGB resolution: 640×480; Depth Resolution: 320×240; With background; Unsegmented action sequences	Common indoor actions; This dataset is designed to investigate variations in the viewpoint
2013 Chalearn Mutli-Modal dataset [43]	2013	20	13858	RGB data, depth maps, user mask, skeletal joints, audio	20 subjects; Each sequence contains 8∼20 gesture samples	Resolution: 640×480; With background; Unsegmented sequences	Italian gestures; Multiple instances, user independent learning; Involve distracter gestures
MSR Action Pairs dataset [26]	2013	12	360	RGB data, depth maps, skeletal joints	10 subjects, 2 or 3 repetitions	Resolution: 320×240; Without background; Segmented sequences	Within each pair the motion and the shape cues are similar; Evaluate the performance of action representation in the context of capturing the prominent cues.
2014 Chalearn Mutli-Modal dataset [44]	2014	20	13858	RGB data, depth maps, user mask, skeletal joints	20 subjects; Each sequence contains 8∼20 gesture samples	Resolution: 640×480; With background; Unsegmented action sequences	Italian gestures; Multiple instances, user independent learning; Involve distracter gestures; More ground-truth annotations than [43]

[a] the number of action classes

[b] the number of video sequences

TABLE III. SUMMARY OF EXPERIMENTAL RESULTS.

(a) Action3D

	Acc.		
	T1	T2	CST
[26]	–	–	88.89%
[27]	–	–	93.09%
[28]	–	–	95.62%
[29]	–	–	86.50%
[31]	95.80%	97.30%	91.70%
[32]	97.8%	100.0%	90.70%
[33]	91.60%	94.20%	74.70%
[34]	96.80%	98.25%	84.80%
[37]	95.80%	97.80%	81.40%
[38]	–	–	88.20%
[39]	98.90%	98.90%	**96.70%**
[40]	–	–	92.46%
[41]	96.20%	97.15%	78.97%
[13]	–	–	81.70%
[42]	–	–	91.70%
[10]	–	–	82.00%

(b) Gesture3D

	Acc.
[26]	92.45%
[27]	**94.74%**
[29]	88.50%

(c) UTKinect

	Acc.
[40]	**97.08%**
[41]	90.92%

(d) Action Pairs

	Acc.
[26]	96.67%
[27]	**98.89%**

accuracy of performance. Depth information contains a great amount of data which could result in high computational costs. It is challenging to develop an effective action recognition approach. The last but not the least, only a few techniques are able to cope with unsegmented action sequences, which is required in realistic scenario. In this paper we show the wide gap that remains between the current state-of-the-art approaches and practical applications. We do so in an effort to motivate subsequent research into action recognition using depth sensors on larger, more challenging action samples, reflecting realistic, real-world conditions.

REFERENCES

[1] 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.

[2] S. Mitra and T. Acharya, “Gesture recognition: A survey,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 3, pp. 311–324, 2007.

[3] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, “Machine recognition of human activities: A survey,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[4] R. Poppe, “A survey on vision-based human action recognition,” *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[5] D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[6] 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.

[7] 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*. Springer, 2013, pp. 149–187.

[8] A. Jaimes and N. Sebe, “Multimodal human–computer interaction: A survey,” *Computer vision and image understanding*, vol. 108, no. 1, pp. 116–134, 2007.

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

[10] D. Wu and L. Shao, “Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 724–731.

[11] R. Slama, H. Wannous, and M. Daoudi, “Grassmannian representation of motion depth for 3d human gesture and action recognition,” in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 3499–3504.

[12] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, “The function space of an activity,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 959–968.

[13] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng, “Online human gesture recognition from motion data streams,” in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 23–32.

[14] J. Wang, Z. Liu, and Y. Wu, *Human Action Recognition with Depth Cameras*. Springer, 2014.

[15] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, “A category-level 3d object dataset: Putting the kinect to work,” in *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 141–165.

[16] E. Trucco and A. Verri, *Introductory techniques for 3-D computer vision*. Prentice Hall Englewood Cliffs, 1998, vol. 201.

[17] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner, “Results and analysis of the chalearn gesture challenge 2012,” in *Advances in Depth Image Analysis and Applications*. Springer, 2013, pp. 186–204.

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

[19] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, “Real-time human pose recognition in parts from single depth images,” *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

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

[21] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[23] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[24] L. Bo, K. Lai, X. Ren, and D. Fox, “Object recognition with hierarchical kernel descriptors,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1729–1736.

[25] K. Lai, L. Bo, X. Ren, and D. Fox, “A large-scale hierarchical multi-view rgb-d object dataset,” in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1817–1824.

[26] O. Oreifej and Z. Liu, “Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 716–723.

[27] X. Yang and Y. Tian, “Super normal vector for activity recognition using depth sequences,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 804–811.

[28] C. Lu, J. Jia, and C.-K. Tang, “Range-sample depth feature for action recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 772–779.

[29] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, “Robust 3d action recognition with random occupancy patterns,” in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 872–885.

[30] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[31] X. Yang, C. Zhang, and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 1057–1060.

[32] B. Liang and L. Zheng, “3d motion trail model based pyramid histograms of oriented gradient for action recognition,” in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 1952–1957.

[33] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[34] 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*. Springer, 2012, pp. 252–259.

[35] A. Jalal, M. Z. Uddin, J. T. Kim, and T.-S. Kim, “Recognition of human home activities via depth silhouettes and r transformation for smart homes,” *Indoor and Built Environment*, p. 1420326X11423163, 2011.

[36] M. R. Malgireddy, I. Inwogu, and V. Govindaraju, “A temporal bayesian model for classifying, detecting and localizing activities in video sequences,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 43–48.

[37] X. Yang and Y. Tian, “Eigenjoints-based action recognition using naive-bayes-nearest-neighbor,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 14–19.

[38] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[39] J. Luo, W. Wang, and H. Qi, “Group sparsity and geometry constrained dictionary learning for action recognition from depth maps,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1809–1816.

[40] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 588–595.

[41] L. Xia, C.-C. Chen, and J. Aggarwal, “View invariant human action recognition using histograms of 3d joints,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[42] M. Zanfir, M. Leordeanu, and C. Sminchisescu, “The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2752–2759.

[43] 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 *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 445–452.

[44] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon, “Chalearn looking at people challenge 2014: Dataset and results,” in *Computer Vision-ECCV 2014 Workshops*. Springer, 2014, pp. 459–473.