

A Comparative Review of Recent Kinect-based Action Recognition Algorithms

Lei Wang, Du Q. Huynh, *Senior Member, IEEE*, Piotr Koniusz

Abstract—Video-based human action recognition is currently one of the most active research areas in computer vision. Various research studies indicate that the performance of action recognition is highly dependent on the type of features being extracted and how the actions are represented. Since the release of the Kinect camera, a large number of Kinect-based human action recognition techniques have been proposed in the literature. However, there still does not exist a thorough comparison of these Kinect-based techniques under the grouping of feature types, such as handcrafted versus deep learning features and depth-based versus skeleton-based features. In this paper, we analyze and compare ten recent Kinect-based algorithms for both cross-subject action recognition and cross-view action recognition using six benchmark datasets. In addition, we have implemented and improved some of these techniques and included their variants in the comparison. Our experiments show that the majority of methods perform better on cross-subject action recognition than cross-view action recognition, that skeleton-based features are more robust for cross-view recognition than depth-based features, and that deep learning features are suitable for large datasets.

Index Terms—Human action recognition, Kinect-based algorithms, cross-view action recognition, 3D action analysis.

I. INTRODUCTION

HUMAN action recognition has many useful applications such as human computer interaction, smart video surveillance, sports and health care. These applications are one of motivations behind much research work devoted to this area in the last few years. However, even with a large number of research papers found in the literature, many challenging problems, such as different viewpoints, visual appearances, human body sizes, lighting conditions, and speeds of action execution, still affect the performance of these algorithms. Further problems include partial occlusion of the human subjects by other objects in the scene and self-occlusion of human subjects themselves. Among the human action recognition papers presented in the literature, some of the early techniques focus on using conventional RGB videos [1–8]. While these video-based techniques gave promising results,

Lei Wang is pursuing a PhD degree in Computer Vision at the Australian National University (ANU) and Data61/CSIRO, Canberra, Australia, ACT 2601 (e-mail: lei.wang@data61.csiro.au). He is also associated with the University of Western Australia (UWA).

Du Q. Huynh is with the Department of Computer Science and Software Engineering, the University of Western Australia, Crawley, WA 6009, Australia (e-mail: du.huynh@uwa.edu.au).

Piotr Koniusz is with Data61/CSIRO (formerly known as NICTA) and the Australian National University, Canberra, Australia, ACT 2601 (e-mail: piotr.koniusz@data61.csiro.au).

Manuscript received on the 30th of March, 2018, revised on the 19th of April, 2019, accepted on the 24th of June, 2019. Kindly cite the TIP rather than the ArXiV version.

their recognition accuracy is still relatively low, even when the scene is free of clutter.

The Kinect camera introduced by Microsoft in 2001 was an attempt to broaden the 3D gaming experience of the Xbox 360's audience. However, as the Kinect camera can capture real-time RGB and depth videos, and there is a publicly available toolkit for computing the human skeleton model from each frame of a depth video, many research papers on 3D human action recognition using the Kinect camera have emerged. One advantage of using depth videos than the conventional RGB videos is that it is easier to segment the foreground human subject even when the scene is cluttered. As depth videos do not have colour information, the colour of the clothes worn by the human subject has no effect on the segmentation process. This allows action recognition researchers to focus their effort more on getting robust feature descriptors to describe the actions rather than on low level segmentation. Numerous representative methods for 3D action analysis using depth videos include [9–14]. These methods employ advanced machine learning techniques for which good results have been reported. Of course, depth images are also vulnerable to noise due to various factors [15]. Thus, using depth images does not always guarantee good action recognition performance [9]. The algorithm used for computing the 3D joint positions of the human skeletal model by the Kinect toolkit is based on the human skeleton tracking framework (OpenNI) of Shotton et al. [16]. In addition to the availability of the real-time depth video stream, this tracking framework also opens up the research area of skeleton-based human action recognition [17–22].

Human action recognition methods using the Kinect data can be classified into two categories, based on how the feature descriptors are extracted to represent the human actions. The first category is handcrafted features. Action recognition methods using handcrafted features require two complex hand-design stages, namely *feature extraction* and *feature representation*, to build the final descriptor. Both the feature extraction stage and feature representation stage differ from one method to another. The feature extraction stage may involve computing the depth (and/or colour) gradients, histogram, and other more complex transformations of the video data. The feature representation stage may involve simple concatenation of the feature components extracted from the previous stage, a more complex fusion step of these feature components, or even using a machine learning technique, to get the final feature descriptor. These methods usually involve a number of operations that require researchers to carry out careful feature engineering and tuning. Kinect-based human action recognition algorithms using handcrafted features reported in

the literature include [9–13, 16, 17, 19, 23, 24].

The second category is deep learning features. With the huge advance in neural network research in the last decade, deep neural networks have been used to extract high-level features from video sequences for many different applications, including 3D human action analysis. Deep learning methods reduce the need for feature engineering; however, they require a huge amount of labelled training data, which may not be available, and a long time to train. For small human action recognition datasets, deep learning methods may not give the best performance. Recent Kinect-based human action recognition algorithms are: [14, 18, 20–22, 25–38].

Research contributions. Although both handcrafted and deep learning features have been used in human action recognition, to the best of our knowledge, a thorough comparison of recent action recognition methods for these two categories is not found in the literature. Our contributions in this paper are twofold:

- We evaluate the performance of 10 recent state-of-art human action recognition algorithms, with specific focus on comparing the effectiveness of using handcrafted features versus deep learning features and skeleton-based features versus depth-based features. We believe that there is a lack of such a comparison in the literature on human action recognition.
- Furthermore, we evaluate the cross-view versus cross-subject performance of these algorithms and, for the multiview datasets, the impact of the camera view for both small and large datasets on human action recognition with respect to whether the features being used are depth-based, skeleton-based, or depth+skeleton-based. To the best of our knowledge, such evaluation has not been performed before.

The paper is organized as follows. Section II gives a brief review on recent human action recognition techniques. Section III covers the details of the 10 algorithms being compared in this paper. In Section IV, we describe our experimental setting and the benchmark datasets. Sections V and VI summarize our experimental results, comparison, and discussions. The last section concludes the paper.

II. RELATED WORK

Action recognition methods can be classified into three categories based on the type of input data: colour-based [1–8, 39–55], depth-based [9–14, 22, 24, 25, 47, 55–65], and skeleton-based [16–24, 28–38, 55, 66–76]. In this section, we will focus on reviewing recent methods using the last two types of features.

Depth-based action recognition. Action recognition from depth videos [9–11, 13, 56–60, 63] has become more popular because of the availability of real-time cost-effective sensors. Most existing depth-based action recognition methods use global features such as space-time volume and silhouette information. For example, Oreifej and Liu [11] captured the discriminative features by projecting the 4D surface normals obtained from the depth sequence onto a 4D regular space to build the *Histogram of Oriented 4D Normals* (HON4D).

Yang and Tian [60] extended HON4D by concatenating local neighbouring hypersurface normals from the depth video to jointly characterize local shape and motion information. More precisely, they introduced an adaptive spatio-temporal pyramid to subdivide the depth video into a set of space-time cells for more discriminative features. Xia and Aggarwal [57] proposed to filter out the noise from the depth sensor so as to get more reliable spatio-temporal interest points for action recognition. Although these methods have achieved impressive performance for frontal action recognition, they are sensitive to changes of viewpoint. One way to alleviate this viewpoint issue is to directly process the pointclouds, as reported in the paper by Rahmani et al. [13].

Apart from the methods mentioned above which use handcrafted features, the use of deep learning features [14, 22, 25, 47, 61, 62, 64, 65] in human action recognition is on the rise. For example, Wang et al. [61] used a *Hierarchical Depth Motion Maps* (HDMMs) to extract the body shape and motion information and then trained a 3-channel deep *Convolutional Neural Network* (CNN) on the HDMMs for human action recognition. In the following years, Rahmani and Mian [14] proposed to train a single *Human Pose Model* (HPM) from real motion capture data to transfer the human pose from different unknown views to a view-invariant feature space, and Zhang et al. [64] used a multi-stream deep neural networks to jointly learn the semantic relations among action attributes.

Skeleton-based action recognition. Existing skeleton-based action recognition methods can be grouped into two categories: joint-based methods and body part based methods. Joint-based methods model the positions and motion of the joints (either individual or a combination) using the coordinates of the joints extracted by the OpenNI tracking framework. For instance, a reference joint may be used and the coordinates of other joints are defined relative to the reference joint [10, 17, 20, 21], or the joint orientations may be computed relative to a fixed coordinate system and used to represent the human pose [66], etc. For the body part based methods, the human body parts are used to model the human's articulated system. These body parts are usually modelled as rigid cylinders connected by joints. Information such as joint angles [19], temporal evolution of body parts [18, 23, 26, 71], and 3D relative geometric relationships between rigid body parts [17, 19, 26] has all been used to represent the human pose for action recognition.

The method proposed by Vemulapalli et al. [17] falls into the body part based category. They represent the relative geometry between a pair of body parts, which may or may not be directly connected by a joint, as a point in $SE(3)$. Thus, a human skeleton is a point of the Lie group $SE(3) \times \dots \times SE(3)$ where each action corresponds to a unique evolution of such a point in time. The approach of Ke et al. [20] relies on both body parts and body joints. The human skeleton model was divided into 5 body parts. A specific joint was selected for each body part as the reference joint and the coordinates of other joints were expressed as vectors relative to that reference joint. Various distance measures were computed from these vectors to yield a feature vector for each video frame. The features

TABLE I: Ten state-of-the-art action recognition methods evaluated in this paper.

Algorithms	Year	Short descriptions	Kinect data used	Feature dimension
HON4D (Oreifej & Liu) [11]	CVPR 2013	handcrafted (global descriptor)	depth	[17880, 151200]
HDG (Rahmani et al.) [12]	WACV 2014	handcrafted (local + global descriptor)	depth+skeleton	[1662, 1819]
LARP-SO (Vemulapalli & Chellappa) [19]	CVPR 2016	handcrafted (Lie Group)	skeleton	$3 \times 3 \times \#frames$
HOPC (Rahmani et al.) [13]	TPAMI 2016	handcrafted (local descriptor)	depth → pointcloud	depending on #STKs [†]
SCK+DCK (Koniusz et al.) [23]	ECCV 2016	handcrafted (tensor representations)	skeleton	~ 40k
P-LSTM (Shahroudy et al.) [18]	CVPR 2016	deep learning (LSTM)	skeleton	#joints $\times 3 \times 8^{\ddagger}$
HPM+TM (Rahmani & Mian) [14]	CVPR 2016	deep learning (CNN)	depth	4096
Clips+CNN+MTLN (Ke et al.) [21]	CVPR 2017	deep learning (pre-trained VGG19, MTLN)	skeleton	7168
IndRNN (Li et al.) [27]	CVPR 2018	deep learning (independently RNN)	skeleton	512
ST-GCN (Yan et al.) [26]	AAAI 2018	deep learning (Graph ConvNet)	skeleton	256

[†]STK stands for *spatio-temporal keypoint*. [‡]The P-LSTM features include 8 video segments, each of which is composed of a number of 3D joints.

vectors from all video frames were finally appended together and scaled to form a handcrafted greyscale image descriptor fed into a CNN. Somewhat related approach [77] uses kernels formed over body joints to obtain feature maps fed into a CNN for simultaneous action recognition and domain adaptation.

Recent human action recognition papers favour deep learning techniques to perform human action recognition. Apart from the CNN-based approaches [20, 77], *Recurrent Neural Networks* (RNNs) have also been popular [18, 27, 35, 38, 65, 67–70]. Since *Long Short-term Memory* (LSTM) [78] can model temporal dependencies as RNNs and even capture the co-occurrences of human joints, LSTM networks have also been a popular choice in human action recognition [31, 33, 36, 37, 73]. For instance, Zhu et al. [69] presented an end-to-end deep LSTM network with a *dropout* step, Shahroudy et al. [18] proposed a *Part-aware Long Short-term Memory* (P-LSTM) network to learn the long-term patterns of the 3D trajectories for each grouped body part, and Liu et al. [36] introduced the use of *trust gates* in their spatio-temporal LSTM architecture.

Action recognition via a combination of skeleton and depth features. Combining skeleton and depth features together helps overcome situations when there are interactions between human subject and other objects or when the actions have very similar motion trajectories. Various action recognition algorithms [12, 22, 47, 72] that use both depth and skeleton features for robust human action recognition have been reported in recent years. For example, Rahmani et al. [12] proposed to combine 4 types of local features extracted from both depth images and 3D joint positions to deal with local occlusions and to increase the recognition accuracy. We refer to their method as *HDG* from hereon. Another example is the approach of Shahroudy et al. [47] where *Local Occupancy Patterns* (LOP), HON4D, and skeleton-based features are combined with hierarchical mixed norms which regularize the weights in each modality group of each body part. Recently, Rahmani and Bennamoun [22] used an end-to-end deep learning model to learn the body part representation from both skeletal and depth images. To improve the performance of the model, they adopted a bilinear compact pooling [79] layer for the generated depth and skeletal features. Elmadany et al. [72], on the other hand, proposed to use canonical correlation analysis to maximize the correlation of features extracted from different sensors. The features investigated in their paper include bag of angles extracted from skeleton data, depth motion map from

depth video, and optical flow from RGB video. The subspace shared by all the features was learned and average pooling was used to get the final feature descriptor.

III. ANALYZED AND EVALUATED ALGORITHMS

We chose ten action recognition algorithms shown in Table I for our comparison and evaluation as they are recent action recognition methods (from 2013 onward) and they use skeleton-based, depth-based, handcrafted, and/or deep learning features. The technical details of these algorithms are summarized below.

HON4D. Oreifej and Liu [11] presented a global feature descriptor that captures the geometry and motion of human action in the 4D space of spatial coordinates, depth and time. To form the HON4D descriptor, the space was quantized using a 600-cell polychoron with 120 vertices. The vectors stretching from the origin to these vertices were used as projection axes to obtain the distribution of normals for each video sequence. To improve the classification performance, random perturbations were added to those projectors. The dimensions of HON4D features (Table I) vary across different datasets.

HDG. In this algorithm [12], each depth sequence was firstly divided into small subvolumes; the histograms of depth and depth derivatives were computed for each subvolume. For each skeleton sequence, the torso joint was used as a stable reference joint for computing the histograms of joint position differences. In addition, the variations of each joint movement volume were incorporated into the global feature vector to form spatio-temporal joint features. Two *Random Decision Forests* (RDFs) were trained in this algorithm, one for feature pruning and one for classification. More details about feature dimensions of HDG and the feature pruning applied by us will be given in Section IV-D.

HOPC. Approach [13] models depth images as 3D pointclouds. The authors used two types of support volume, namely, so-called spatial support volume and spatio-temporal support volume. The HOPC descriptor was extracted from the pointcloud falling inside the support volume around each point, which may be classified as a *spatio-temporal Keypoint* (STK) if the eigenvalue ratios of the pointcloud around it are larger than some predefined threshold. For each STK, the algorithm further projected eigenvectors onto the axes of the 20 vertices of a regular dodecahedron. The final HOPC descriptor for each STK is a concatenation of 3 small histograms, each of which

captures the distribution of an eigenvector of the pointcloud within the support volume.

LARP-SO. Vemulapalli and Chellappa [19] extended their earlier work [17] by the use of Lie Algebra Relative Pairs via $SO(3)$ for action recognition. We follow the convention adopted in [22] and name this algorithm as LARP-SO. In this algorithm, the rolling map, which describes how a Riemannian manifold rolls over another one along a smooth rolling curve, was used for 3D action recognition. Each skeleton sequence was firstly represented by the relative 3D rotations between various human body parts, and each action was then modelled as a curve in the Lie Group. Since it is difficult to perform the classification of action curves in a non-Euclidean space, the curves were unwrapped by the logarithm map at a single point while a rolling map was used to reduce distortions. The *Fourier Temporal Pyramid* (FTP) representation [56] was used in the algorithm to make the descriptor more robust to noise and less sensitive to temporal misalignments.

SCK+DCK. Koniusz et al. [23] used tensor representations to capture the higher-order relationships between 3D human body joints for action recognition. They applied two different RBF kernels which they referred to as *Sequence Compatibility Kernel* (SCK) and *Dynamics Compatibility Kernel* (DCK). The former kernel captures the spatio-temporal compatibility of joints while the latter models the action dynamics of a sequence. An SVM was then trained on linearized feature maps of such kernels for action classification.

HPM+TM. Approach [14] employs a dictionary containing representative human poses from a motion capture database. A deep CNN architecture which is a modification of [45] was then used to train a view-invariant human pose model. Real depth sequences were passed to the learned model frame-by-frame to extract high-level view-invariant features. Similarly to the LARP-SO algorithm above, the FTP was used to capture the temporal structure of the action videos. The final descriptor for each video sequence is a collection of the Fourier coefficients from all the segments.

P-LSTM. Approach [18] proceeds by transforming 3D coordinates of the body joints from the camera coordinate system to the body coordinate system with the origin set at the spine. The 3D coordinates of all other body joints were then scaled based on the distance between the ‘hip centre’ joint and the ‘spine’ joint. A P-LSTM model was built by splitting the memory cells from the LSTM model into body part based sub-cells. For each video sequence, the pre-processed human joints were grouped into 5 parts (torso, two hands, and two legs) and the video was divided into 8 equal-sized video segments. Then, for a randomly selected frame per video segment, 3D coordinates of the joints inside each grouped part were concatenated and passed as input to the P-LSTM network to learn common temporal patterns of the parts and combine them into a global representation.

Clips+CNN+MTLN. Ke et al. [21] presented a skeletal representation referred to as *clips*. The method proceeds by transforming the Cartesian coordinates of human joints (per skeleton sequence) into the cylindrical coordinates to generate 3 clips, with each clip corresponding to one channel of the

cylindrical coordinates. To encode the temporal information for the whole video sequence, four stable joints (left shoulder, right shoulder, left hip and right hip) were selected as reference joints to produce 4 coordinate frames. The pre-trained VGG19 network [80] was used as a feature extractor to learn the long-term spatio-temporal features from intermediate images formed from the 4 coordinate frames. Moreover, approach [21] also employs the *Multi-task Learning Network* (MTLN) proposed by [81] to incorporate the spatial structural information from the CNN features.

IndRNN. Li et al. [27] proposed a new RNN method, an Independently Recurrent Neural Network, for which neurons per layer are independent of each other but they are reused across layers. Finally, multiple IndRNNs were stacked to build a deeper network than the traditional RNN.

ST-GCN. The spatio-temporal graph representation for skeleton sequences proposed by Yan et al. [26] is an extension of *Graph Convolutional Networks* (GCN) [82–84] tailored to perform human action recognition. Firstly, the spatio-temporal graph is constructed by inserting edges between neighbouring body joints (nodes) of the human body skeleton as well as along the temporal direction. Subsequently, GCN and a classifier are applied to infer dependencies in the graphs (a single graph corresponds to a single action sequence) and perform classification.

IV. EXPERIMENTAL SETTING

To perform experiments, we obtained off-the-shelf codes for HON4D [11], HOPC [13], LARP-SO [19], HPM+TM [14], IndRNN [27] and ST-GCN [26] from the respective authors’ websites. For SCK+DCK [23], HDG [12], P-LSTM [18] and Clips+CNN+MTLN [21], we used our own Matlab implementations given that codes for these methods are not publicly available. Moreover, we employed ten variants of the HDG [12] representation so as to evaluate the performance with respect to different combinations of its individual descriptor types. We also implemented the traditional RNN and LSTM as baseline methods, and added four variants of P-LSTM to evaluate the impact of using different numbers of video segments for skeletal representation as well as different numbers of hidden neurons.

A. Benchmark Datasets

Listed in Table II are six benchmark datasets used in our evaluation, each of which is detailed below.

MSRAAction3D [9] is one of the earliest action datasets captured with the Kinect depth camera. It contains 20 human sport-related activities such as *jogging*, *golf swing* and *side boxing*. Each action in this dataset was performed 2 or 3 times by 10 people. This dataset is challenging because of high interaction similarities.

3D Action Pairs [11] contains 6 selected pairs of actions that have very similar motion trajectories, e.g., *put on a hat* and *take off a hat*; *pick up a box* and *put down a box*; *stick a poster* and *remove a poster*. Each action was performed 3 times by 10 people. There are two challenging aspects of this dataset:

TABLE II: Six publicly available benchmark datasets used in our experiments for 3D action recognition.

Datasets	Year	Classes	Subjects	#Views	#videos	Sensor	Modalities	#joints
MSRACTION3D [9]	2010	20	10	1	567	Kinect v1	Depth + 3DJOints	20
3D Action Pairs [11]	2013	12	10	1	360	Kinect v1	RGB + Depth + 3DJOnts	20
CAD-60 [42]	2011	14	4	—	68	Kinect v1	RGB + Depth + 3DJOnts	15
UWA3D Activity Dataset [58]	2014	30	10	1	701	Kinect v1	RGB + Depth + 3DJOnts	15
UWA3D Multiview Activity II [13]	2015	30	9	4	1070	Kinect v1	RGB + Depth + 3DJOnts	15
NTU RGB+D Dataset [18]	2016	60	40	80	56880	Kinect v2	RGB + Depth + IR + 3DJOnts	25

(The number of views is not stated in the CAD-60 dataset.)

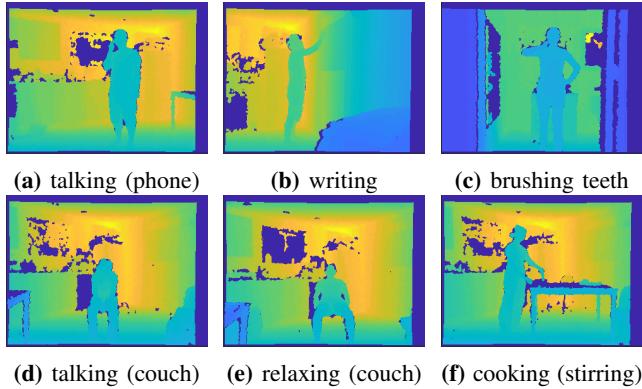


Fig. 1: Sample depth images from the CAD-60 dataset.

- (i) the actions in each pair have similar motion trajectories;
- (ii) the object that is interacted by the subject in each video is only present in the RGB-D data but not the skeleton data.

Cornell Activity Dataset (CAD) [42] comprises two sub-datasets, CAD-60 and CAD-120. Both sub-datasets contain RGB-D and tracked skeleton video sequences of human activities captured by a Kinect sensor. In this paper, only CAD-60 was used in the experiments. Fig. 1 illustrates depth images from the CAD-60 dataset and demonstrates that this dataset exhibits high levels of noise in its depth videos.

UWA3D Activity Dataset [58] contains 30 actions performed by 10 people of various height at different speeds in cluttered scenes. This dataset has high between-class similarity and contains frequent self-occlusions.

UWA3D Multiview Activity II [13] contains 30 actions performed by 9 people in a cluttered environment. In this dataset, the Kinect camera was moved to different positions to capture the actions from 4 different views (see Fig. 2(a)-(c)): front view (V_1), left view (V_2), right view (V_3), and top view (V_4). This dataset is therefore more challenging than the previous four datasets. Fig. 2(d) shows sample video frames from this dataset.

NTU RGB+D Dataset [18] is so far the largest Kinect-based action dataset which contains 56,880 video sequences and over 4 million frames. There are 60 action classes performed by 40 subjects captured from 80 views with 3 Kinect v.2 cameras. This dataset has variable sequence lengths for different sequences and exhibits high intra-class variations.

Dataset usage. Below we detail how the above six datasets were used in our experiments.

The MSRACTION3D, 3D Action Pairs, CAD-60 and UWA3D Activity datasets were used for cross-subject (single-view)

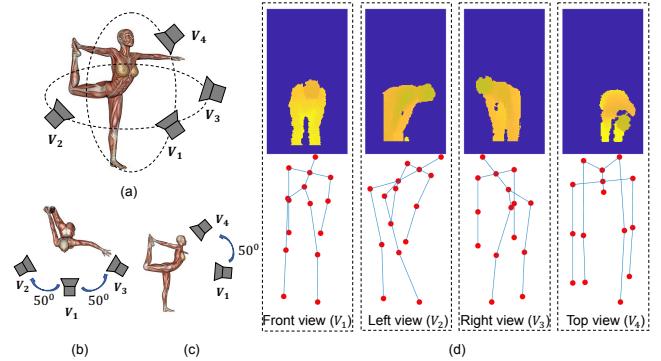


Fig. 2: (a) A perspective view of the camera setup in the UWA3D Multiview Activity II dataset. The views V_1 , V_2 and V_3 are at the same height. (b) and (c) show the top and side views of the setup. The angles between V_1 and V_2 , between V_1 and V_3 , and between V_1 and V_4 are all approximately 50 degrees [58]. (d) An example video frame of the depth and skeleton data for the *bending* action.

experiments. For every dataset, we used half of the subjects' data for training and the remaining half for testing. We tested all the possible combinations of subjects for the training and testing splits to obtain the average recognition accuracy of each algorithm. For example, for 10 subjects in the MSRACTION3D dataset, $\binom{10}{5} = 252$ experiments were carried out.

The UWA3D Multiview Activity II dataset was used for cross-view experiments, with two views of the samples being used for training and the remaining views for testing. There were 12 different view combinations in the experiments.

The NTU RGB+D dataset was used in both cross-subject and cross-view experiments. Despite indications that this dataset has 80 views of human action recognition, the data samples were grouped according to three camera sets. For cross-view action recognition, we used the video sequences captured by two cameras as our training data and the remaining sequences for testing. A total of 3 different camera combinations were experimented with.

B. Evaluation Settings

Below we detail the experimental settings of the algorithms.

HON4D. According to [11], HON4D has the frame size of 320×240 and each video is divided into $4 \times 3 \times 3$ (width \times height \times #frames) spatio-temporal cells. In our evaluations, we used these same settings for all the datasets.

HOPC. Rahmani et al. [13] used different spatial and temporal scales for different datasets. In this paper, a constant temporal scale and spatial scale were used for all the datasets. For the MSRACTION3D and 3D Action Pairs datasets, the temporal

and spatial scales were set to 2 and 19, respectively. For the remaining datasets, we used 2 for the temporal scale and 140 as the spatial scale. Moreover, we divided each depth video into $6 \times 5 \times 3$ spatio-temporal cells (along the X , Y and *time* axes) to extract features.

LARP-SO. The desired number of frames [19] used for computing skeletal representation varies depending on the datasets used in the experiments. The desired frame numbers for the UWA3D Activity, UWA3D Multiview Activity II, and NTU RGB+D datasets were all set to 100. For the MSRAction3D, 3D Action Pairs datasets, and CAD-60, they were set to 76, 111, and 1,000, respectively.

SCK+DCK. We followed the experimental settings described in [23, 85] for all the datasets and we used authors' newest model which aggregates over subsequences (not just sequences). For SCK, we normalized all human body joints with respect to the hip joints across frames as well as the lengths of all body parts. For DCK, we used the unnormalized body joints, and assumed that the displacements of body joint coordinates across frames captured their temporal evolution.

HDG. According to [12], the number of used subvolumes has no significant effect to the discriminative features, we divided each video sequence into $10 \times 10 \times 5$ subvolumes (along X , Y and *time*) for computing the histograms of depth as well as the depth gradients. For the joint movement volume features, each joint volume was divided into $1 \times 1 \times 5$ cells (along X , Y and *time*). There are four individual feature representations encapsulated by HDG:

- (i) histogram of depth (hod),
- (ii) histogram of depth gradients (hodg),
- (iii) joint position differences (jpd),
- (iv) joint movement volume features (jmvt).

We follow [86] and evaluate the performance of the 10 variants of HDG in our experiments.

HPM+TM. We followed [14] and set the number of Fourier Pyramid levels to 3 and the number of low frequency Fourier coefficients to 4 for all datasets. We used the human pose model trained by Rahmani and Mian [14] to extract view-invariant features from each depth sequence. We also compared the recognition accuracies of this algorithm given Average Pooling (AP) versus Temporal Modelling (TM) used for extraction of CNN features.

P-LSTM. We applied the same normalization preprocessing step as in [18] for the skeletal representation. In our experiments, the number of video segments and the number of hidden neurons for 1-layer RNN, 2-layer RNN, 1-layer LSTM, 2-layer LSTM, 1-layer P-LSTM and 2-layer P-LSTM were all set to 8 and 50, respectively. We also evaluated the performance of different numbers of video segments and different numbers of hidden neurons in P-LSTM. The learning rate and the number of epochs in our experiments were set to 0.01 and 300, respectively.

Clips+CNN+MTLN. The learning rate was set to 0.001 and the batch size was set to 100 for MTLN. We selected four different experimental settings from [21] to compare the performance of recognition: Frames+CNN, Clips+CNN+Pooling, Clips+CNN+Concatenation, and Clips+CNN+MTLN.

IndRNN. We used the Adam optimizer with the initial learning rate 2×10^{-4} and applied the decay by 10 once the evaluation accuracy did not increase. For cross-subject and cross-view experiments, the dropout rates were set to 0.25 and 0.1, respectively.

ST-GCN. For the convolution operations, we used the optimal partition strategy according to the ablation study in [26]. As different datasets have different numbers of body joints (see Table II), we reconstructed the spatio-temporal skeleton graphs. For NTU RGB+D dataset, we used the same experimental settings as described in [26] (e.g., we work with up to two human subjects per sequence). For the remaining 5 datasets, we used a different setting as only one performing subject was present per video.

Moreover, we performed extra experiments for IndRNN and ST-GCN: instead of using 3D skeleton sequences as inputs, we used jpd features which redefine the 3D skeleton joint positions by translating them to be centred at the torso (or 'spine') joint.

C. Evaluation Measure

The recognition accuracy A_c of an algorithm for any given action class c is defined as the proportion of correct class c labels returned by the algorithm:

$$A_c = \# \text{correct_class}_c \text{_labels} / \# \text{actual_class}_c \text{_labels}. \quad (1)$$

To show the recognition accuracies of an algorithm for all the action classes, a confusion matrix is often used. The overall performance of an algorithm on a given dataset is evaluated using the average recognition accuracy \bar{A} defined as: $\bar{A} = \frac{1}{C} \sum_{c=1}^C A_c$, where C is the total number of action classes in a given dataset.

To show the overall performance of each algorithm on M datasets, we first rank the performance of each algorithm from 1 to 5 (a lower rank value represents a better performance) based on the recognition accuracy so each algorithm has its own rank value r_i given the i^{th} dataset. We then compute the Average Rank (AVRank) as follows:

$$\text{AVRank} = \frac{1}{M} \sum_{i=1}^M r_i. \quad (2)$$

D. Optimisation of Hyperparameters for HDG

There are 3 hyperparameters in the HDG algorithm. The first hyperparameter is the number of subvolumes, which we set to the same value as in [12]. The second and third hyperparameters, which are the number of trees N_{trees} used in training and the threshold θ used in feature pruning, were optimised during our experiments (Table III). As the length of the combined features in the HDG algorithm is large (e.g., the length of the HDG-all features for the MSRAction3D dataset is 13,250), we trained one RDF to select the feature components that have high importance values. This helped to increase the processing speed without compromising the recognition accuracy.

We evaluated the effect of hyperparameters N_{trees} and θ on the HDG algorithm for different combinations of individual

HDG features using the MSRAAction3D dataset (for single-view) and the UWA3D Multiview Activity II dataset (for cross-view). Table III shows the optimal values for N_{trees} and θ obtained from the grid search. The corresponding dimensions of different HDG combined features before and after pruning are also indicated. Compared to other individual features of HDG, j_{pd} and j_{mv} are small-sized features, thus when used alone in both datasets, their dimensions are not reduced by much during feature pruning. However, when either or both of them are combined with other individual features in cross-view action recognition, their importance values are significantly higher. This allows a large reduction in feature dimension after pruning (see the last 4 rows of the table). Our experiments in the next section also confirm that skeleton-based features deal better with the view-invariance than depth-based features.

The optimal values of N_{trees} and θ shown in Table III were used to prune the HDG features for all datasets. As different datasets have different numbers of body joints (see Table II), the dimensions of these HDG features after pruning across datasets are not the same.

V. EXPERIMENTAL RESULTS

A. MSRACTION3D, 3D Action Pairs, CAD-60, and UWA3D Activity Datasets

Table IV summarizes the results for the single-view action recognition on these datasets. The algorithms with the highest recognition accuracy for the handcrafted feature and deep learning feature categories are highlighted in bold.

Among the methods that use handcrafted features, SCK+DCK outperformed all other methods on all the datasets. The use of RBF kernels to capture higher-order statistics of the data and complexity of action dynamics demonstrates its effectiveness for action recognition. For the 3D Action Pairs and UWA3D Activity datasets, HON4D and HOPC are, respectively, the second top performers. The poorer performance of HDG was due to noise in the depth sequences which affected the HDG-hod and HDG-hodg features, even though the human subjects were successfully segmented. In general, HDG-hodg outperformed HDG-hod, and HDG-jmv outperformed HDG-jpd. We also found that concatenating more individual features in HDG (see Table IV, row HDG-hod+hodg+jpd to row HDG-all features) helped improve action recognition. We note that our results for HDG are different from those reported in [12] because we used $1 \times 1 \times 5 = 5$ cells [86] instead of $2 \times 2 \times 5 = 20$ cells to store the joint motion features.

For deep learning methods, the 1-layer P-LSTM method (8 video segments, 50 hidden neurons) outperformed others on the UWA3D Activity dataset. In general, a 1-layer LSTM with more hidden neurons performed better than a 1-layer LSTM with fewer neurons, and P-LSTM performed better than the traditional LSTM and RNN (see the last column in Table IV). The 2-layer P-LSTM (20 video segments, 50 hidden neurons) achieved the best recognition accuracy on the MSRAAction3D dataset and HPM+TM outperformed on the 3D Action Pairs dataset. Comparing the last 5 rows of Table IV for different variants of 2-layer P-LSTM shows that having more video

segments and/or hidden neurons does not guarantee better performance. The reasons are: (i) more video segments have less averaging effect and so it is likely that noisy video frames with unreliable skeletal information would be used for feature representation; (ii) having too many hidden neurons would cause overfitting in the training process.

Using the j_{pd} features instead of the raw 3D joint coordinates boosts the recognition accuracies for both IndRNN and ST-GCN on almost all the datasets. For the CAD-60 and MSRAAction3D datasets, the 4-layer IndRNN and 6-layer IndRNN using j_{pd} features are the top two performers. Compared to using the raw 3D joint coordinates, the improvement due to the use of j_{pd} is 4.56% for IndRNN (6 layers) on the MSRAAction3D dataset and 32.97% for IndRNN (4 layers) on the CAD-60 dataset.

The last column of Table IV computed using Eq. (2) shows that SCK+DCK obtains average rank 1 score, followed closely by the 6-layer IndRNN with j_{pd} with average rank 2 score. The ST-GCN method uses a more complex architecture having 9 layers of spatio-temporal graph convolutional operators, which require a large dataset for training. As all the datasets in Table IV are quite small, its poor performance (average rank 4.75 score) is not unexpected.

B. NTU RGB+D Dataset

Table V summarizes the evaluation results for cross-subject and cross-view action recognition on the NTU RGB+D dataset, where methods are grouped into the handcrafted feature and deep learning feature categories. Top performing methods are highlighted in bold.

Among the methods using handcrafted features, SCK+DCK performed the best for both cross-subject action recognition and cross-view action recognition. Similarly to results on the four datasets shown in Table IV, combining more individual features in HDG resulted in higher recognition accuracy for both the cross-subject and cross-view action recognition.

Among deep learning methods, ST-GCN and the 6-layer IndRNN combined with the j_{pd} features outperformed other deep learning methods in both cross-subject and cross-view action recognition. In particular, with j_{pd} , ST-GCN became the top performer (achieving 83.36%) for cross-subject action recognition and IndRNN (6 layers) achieved the highest accuracy (89.0%) for cross-view action recognition. Compared to using the raw 3D skeleton joint coordinates, using the j_{pd} features helps improve the recognition accuracies of both IndRNN and ST-GCN. For example, with j_{pd} features, both the top-1 and top-5 accuracies of ST-GCN increased by 1.79% and 0.61% in the cross-subject experiment.

For the other deep learning methods, the recognition accuracies of 2-layer RNN, 2-layer LSTM and 2-layer P-LSTM are higher than those of 1-layer RNN, 1-layer LSTM and 1-layer P-LSTM. Similar to the results in Table IV, P-LSTM performed better than the traditional LSTM and RNN. HPM+AP and HPM+TM, on the other hand, did not perform so well. The reason is that both of these methods were trained given only 339 representative human poses from a human pose dictionary whereas the 60 action classes of NTU RGB+D dataset include many more human poses of higher complexity.

TABLE III: Optimal hyperparameter values and feature dimensions before and after pruning for the HDG combined features.

Combination of individual features	MSRAction3D				UWA3D Multiview Activity II			
	dimensions before pruning	optimal θ $\times 10^{-3}$	optimal N_{trees}	dimensions after pruning	dimensions before pruning	optimal θ $\times 10^{-3}$	optimal N_{trees}	dimensions after pruning
HDG-hod	2,500	1.5	100	1,442	2,500	2.7	60	1,231
HDG-hodg	10,000	20.9	200	550	10,000	2.4	60	3,891
HDG-jpd	150	11.5	80	148	150	48.1	120	142
HDG-jmv	600	3.4	140	571	450	3.1	60	449
HDG-hod+hodg	12,500	17.0	180	786	12,500	4.7	140	3,987
HDG-jpd+jmv	750	2.0	80	690	600	6.8	100	581
HDG-hod+hodg+jpd	12,650	8.2	160	2,221	12,650	29.4	80	239
HDG-hod+hodg+jmv	13,100	7.4	180	2,189	12,950	25.0	100	135
HDG-hodg+jpd+jmv	10,750	8.3	180	1,711	10,600	15.2	140	456
HDG-all features	13,250	13.3	120	1,013	13,100	19.0	100	300

TABLE IV: Comparison of average cross-subject action recognition accuracies (percentage) for the four single-view datasets (*i.e.*, $M = 4$ in Eq. (2)). Each block of rows shows the performance of one method and its variants. The best algorithm for each dataset is highlighted in bold. The last column of the table shows the average rank of the best performing algorithm in each block. The final rank values are computed using Eq. (2), where top performing methods have smaller rank values. Other poorer performing methods in the same block are not considered for their rank values, so their final ranks are marked as ‘–’.

	Method	MSRAction3D	3D Action Pairs	CAD-60	UWA3D Activity	AVRank
Hand-crafted features	HON4D [11] (Depth)	82.15	96.00	72.70	48.89	3.25
	HOPC [13] (Depth)	85.49	92.44	47.55	60.58	3.25
	LARP-SO-logarithm map [19] (Skel.)	88.69	92.96	69.12	51.96	–
	LARP-SO-unwrapping while rolling [19] (Skel.)	88.47	94.09	69.12	53.05	–
	LARP-SO-FTP [19] (Skel.)	89.40	94.67	76.96	50.41	2.50
	HDG-hod [86] (Depth)	66.22	81.20	26.47	44.35	–
	HDG-hodg [86] (Depth)	70.34	90.98	50.98	54.23	–
	HDG-jpd [86] (Skel.)	55.54	53.78	46.08	40.88	–
	HDG-jmv [86] (Skel.)	62.40	84.87	41.18	51.02	–
	HDG-hod+hodg [86] (Depth)	71.81	90.96	51.96	55.17	–
Deep learning features	HDG-jpd+jmv [86] (Skel.)	65.57	84.93	49.02	55.57	–
	HDG-hod+hodg+jpd [86] (Depth + Skel.)	72.06	90.72	51.47	56.41	–
	HDG-hod+hodg+jmv [86] (Depth + Skel.)	75.41	92.27	49.51	58.80	–
	HDG-hodg+jpd+jmv [86] (Depth + Skel.)	75.00	92.28	52.94	59.82	–
	HDG-all features [86] (Depth + Skel.)	75.45	92.13	51.96	60.33	4.00
	SCK+DCK [23] (Skel.)	89.47	96.00	89.22	61.52	1.00
	Frames+CNN [21] (Skel.)	60.73	73.71	58.82	46.47	–
	Clips+CNN+Pooling [21] (Skel.)	67.64	74.86	58.82	46.47	–
	Clips+CNN+Concatenation [21] (Skel.)	71.27	78.29	61.76	53.85	–
	Clips+CNN+MTLN [21] (Skel.)	73.82	79.43	67.65	54.81	2.25
Deep learning features	HPM+AP [14] (Depth)	56.73	56.11	44.12	42.32	–
	HPM+TM [14] (Depth)	72.00	98.33	44.12	54.78	2.50
	1-layer RNN [18] (Skel.)	18.02	32.76	54.90	14.27	–
	2-layer RNN [18] (Skel.)	27.80	56.13	54.91	35.36	–
	1-layer LSTM [18] (Skel.)	62.26	67.14	61.77	50.81	–
	2-layer LSTM [18] (Skel.)	65.33	73.72	63.24	46.78	–
	1-layer P-LSTM [18] (Skel.)	70.50	70.86	61.76	55.16	–
	2-layer P-LSTM [18] (Skel.)	69.35	72.00	67.65	50.81	2.75
	Our implementation with modified hyperparam. values:					
	1-layer LSTM (8 segments, 100 hidden neurons) (Skel.)	64.75	73.14	58.82	52.58	–
Deep learning features	2-layer P-LSTM (10 segments, 50 hidden neurons) (Skel.)	66.09	75.43	67.65	50.00	–
	2-layer P-LSTM (10 segments, 100 hidden neurons) (Skel.)	67.43	71.43	54.41	50.32	–
	2-layer P-LSTM (20 segments, 50 hidden neurons) (Skel.)	73.18	71.43	52.94	49.68	–
	2-layer P-LSTM (20 segments, 100 hidden neurons) (Skel.)	70.50	71.43	58.82	53.55	–
	IndRNN (4 layers) [27] (Skel.)	71.50	90.05	51.72	44.63	–
	IndRNN (6 layers) [27] (Skel.)	72.91	89.53	57.03	42.66	–
	Our improved results:					
	IndRNN (4 layers, with jpd) (Skel.)	76.34	82.66	84.69	52.09	–
	IndRNN (6 layers, with jpd) (Skel.)	77.47	86.88	80.16	51.34	2.00
	ST-GCN* [26] (Skel.)	27.64 (69.09)	20.00 (77.14)	23.53 (70.59)	22.12 (45.83)	–
Deep learning features	Our improved results:					
	ST-GCN* (with jpd) (Skel.)	18.18 (64.00)	54.16 (96.57)	26.47 (67.65)	36.54 (70.51)	4.75

*For ST-GCN, the numbers inside the parentheses denote the top-5 accuracy.

C. UWA3D Multiview Activity II Dataset

The ten algorithms were compared on the UWA3D Multiview Activity II dataset using cross-view action recognition. Table VI summarizes the results. The top 2 action recognition algorithms are highlighted in bold in each column for the handcrafted feature and deep learning feature categories.

For the methods using handcrafted features, the HDG-all

features performed the best for cross-view action recognition (the last column in Table VI) followed by other HDG variants that use skeleton-only methods. HDG-jpd+jmv was the second best performer followed by SCK+DCK, HDG-jmv, and LARP-SO-FTP, which all performed better than depth-based features such as HON4D, HDG-hod, HDG-hodg and HDG-hod+hodg. According to

TABLE V: Comparison of average recognition accuracies (percentage) for both cross-subject and cross-view action recognition on the NTU RGB+D Dataset.

	Method	Cross-subject	Cross-view
Hand-crafted features	HON4D [11] (Depth)	30.6	7.3
	HOPC [13] (Depth)	40.3	30.6
	LARP-SO-FTP [19] (Skel.)	52.1	53.4
	HDG-hod [86] (Depth)	20.1	13.5
	HDG-hodg [86] (Depth)	23.0	25.2
	HDG-jpd [86] (Skel.)	27.8	35.9
	HDG-jmv [86] (Skel.)	38.1	50.0
	HDG-hod+hdg [86] (Depth)	24.6	26.5
	HDG-jpd+jmv [86] (Skel.)	39.7	51.9
	HDG-hod+hdg+jpd [86] (Depth + Skel.)	29.4	38.8
	HDG-hod+hdg+jmv [86] (Depth + Skel.)	39.0	57.0
	HDG-hodg+jpd+jmv [86] (Depth + Skel.)	41.2	57.2
	HDG-all features [86] (Depth + Skel.)	43.3	58.2
	SCK+DCK [23] (Skel.)	72.8	74.1
Deep learning features	Frames+CNN [21] (Skel.)	75.7	79.6
	Clips+CNN+Concatenation [21] (Skel.)	77.1	81.1
	Clips+CNN+Pooling [21] (Skel.)	76.4	80.5
	Clips+CNN+MTLN [21] (Skel.)	79.6	84.8
	Clips+CNN+MTLN [‡] [21] (Skel.)	79.54	84.70
	HPM+AP [13] (Depth)	40.2	42.2
	HPM+TM [13] (Depth)	50.1	53.4
	1-layer RNN [18] (Skel.)	56.0	60.2
	2-layer RNN [18] (Skel.)	56.3	64.1
	1-layer LSTM [18] (Skel.)	59.1	66.8
	2-layer LSTM [18] (Skel.)	60.7	67.3
	1-layer P-LSTM [18] (Skel.)	62.1	69.4
	2-layer P-LSTM [18] (Skel.)	62.9	70.3
	2-layer P-LSTM [‡] [18] (Skel.)	63.02	70.39
	IndRNN (4 layers) [27] (Skel.)	78.6	83.8
	IndRNN (6 layers) [27] (Skel.)	81.8	88.0
	Our improved results: IndRNN (4 layers, with jpd)(Skel.)	79.5	84.5
	IndRNN (6 layers, with jpd)(Skel.)	83.0	89.0
ST-GCN* [26] (Skel.)	81.57 (96.85)	88.76 (98.83)	
	Our improved results: ST-GCN* (with jpd) (Skel.)	83.36 (97.46)	88.84 (98.87)

[‡]Our implementations for reproducing original authors' experiment results.

*For ST-GCN, the numbers inside the parentheses denote the top-5 accuracy.

the table, using one or both skeleton-based features (HDG-jpd and/or HDG-jmv) in HDG improved its results.

For deep learning methods, HPM+TM and HPM+AP achieved the highest results. Although Clips+CNN+MTLN performed better than other 'Clips' variants, it did not perform as good as HPM+TM and HPM+AP, due to the limited number of video samples in the dataset. P-LSTM performed better than the traditional LSTM and RNN. We noticed that stacking more layers for IndRNN or using jpd features for both IndRNN and ST-GCN did not help increase the recognition accuracy. This is due to the lack of representative training videos (compared to the results on the NTU RGB+D dataset in cross-view action recognition).

Fig. 3 shows a confusion matrix for the HDG-all representation on the UWA3D Multiview Activity II dataset when V_3 and V_4 views were used for training and V_1 was used for testing. According to the figure, the algorithm was confused by a few actions which have similar motion trajectories. For example, *one hand waving* is similar to *two hand waving* and *two hand punching*; *walking* is very similar to *irregular walking*; *bending* and *putting down* have very similar appearance; *sneezing* and *coughing* are very similar actions.

VI. DISCUSSIONS

A. Single-view versus cross-view

Table VII summarizes the performance of each algorithm for both single-view action recognition and cross-view action recognition. One can observe from the table that some algorithms such as HON4D, LARP-SO-FTP, Clips+CNN+MTLN,

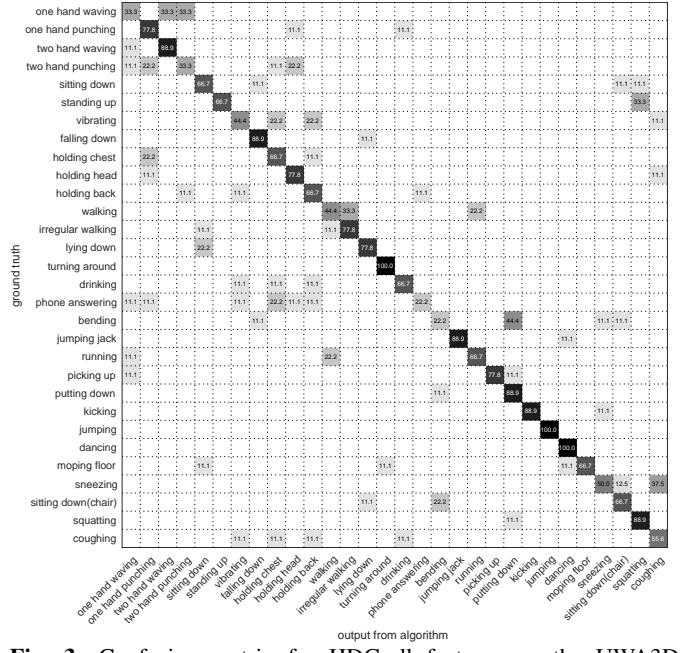


Fig. 3: Confusion matrix for HDG-all features on the UWA3D Multiview Activity II dataset when V_3 and V_4 were used for training and V_1 was used for testing. For each action class along the diagonal, the darker is the colour, the higher is the recognition accuracy.

P-LSTM, and SCK+DCK performed better for single-view action recognition but performed slightly worse in cross-view action recognition. For cross-subject action recognition, SCK+DCK outperformed all other algorithms; for cross-view action recognition, HDG-all, SCK+DCK and our improved IndRNN with jpd features all performed well (having the same average rank values).

Fig. 4 summarizes the average accuracy of all the algorithms grouped under the handcrafted and deep learning feature categories. On average, we found that the recognition accuracy for cross-view action recognition is lower than cross-subject action recognition using handcrafted features, with the poorest performance from depth-based features. Combining both depth- and skeleton-based features together helps improve cross-view action recognition and gives a similar performance for cross-subject action recognition. Overall, the performance on cross-view recognition is lower than cross-subject recognition, except for the deep learning depth-based feature methods. It should be noted that there are only two methods (HPM+AP and HPM+TM) falling into this category and they performed well on the UWA3D Multiview Activity II dataset and reasonably well on the NTU RGB+D dataset.

B. Influence of camera views in cross-view evaluation

To analyze the influence of using different combinations of camera views in training and testing, we classified all the algorithms in Table VI into 3 feature types: depth-based, skeleton-based, and depth+skeleton-based representations. Scatter plots showing the performance of all the algorithms on these three feature types are given in Fig. 5, where the blue dots and red crosses denote, respectively, algorithms using handcrafted features and deep learning features.

TABLE VI: Comparison of average recognition accuracies (percentage) for cross-view action recognition on the UWA3D Multiview Activity II dataset.

	Training view	$V_1 \& V_2$		$V_1 \& V_3$		$V_1 \& V_4$		$V_2 \& V_3$		$V_2 \& V_4$		$V_3 \& V_4$		Average
	Testing view	V_3	V_4	V_2	V_4	V_2	V_3	V_1	V_4	V_1	V_3	V_1	V_2	
Hand-crafted features	HON4D [11] (Depth)	31.1	23.0	21.9	10.0	36.6	32.6	47.0	22.7	36.6	16.5	41.4	26.8	28.9
	HOPC [13] (Depth)	25.7	20.6	16.2	12.0	21.1	29.5	38.3	13.9	29.7	7.8	41.3	18.4	22.9
	Holistic HOPC* [13] (Depth)	32.3	25.2	27.4	17.0	38.6	38.8	42.9	25.9	36.1	27.0	42.2	28.5	31.8
	Local HOPC+STK-D* [13] (Depth)	52.7	51.8	59.0	57.5	42.8	44.2	58.1	38.4	63.2	43.8	66.3	48.0	52.2
	LARP-SO-logarithm map [19] (Skel.)	48.2	47.4	45.5	44.9	46.3	52.7	62.2	46.3	57.7	45.8	61.3	40.3	49.9
	LARP-SO-unwrapping while rolling [19] (Skel.)	50.4	45.7	44.0	44.5	40.8	49.6	57.4	44.4	57.6	47.4	59.2	40.8	48.5
	LARP-SO-FTP [19] (Skel.)	54.9	50.0	54.9	48.1	56.0	66.5	57.2	62.5	54.0	68.9	43.6	56.0	
	HDG-hod [86] (Depth)	22.5	17.4	12.5	10.0	19.6	20.4	26.7	13.0	18.7	10.0	27.9	17.2	18.0
	HDG-hodg [86] (Depth)	26.9	34.2	20.3	18.6	34.7	26.7	41.0	29.2	29.4	11.8	40.7	28.8	28.5
	HDG-jpd [86] (Skel.)	36.3	32.4	31.8	35.5	34.4	38.4	44.2	30.0	44.5	33.7	44.4	34.0	36.6
Deep learning features	HDG-jmv [86] (Skel.)	57.2	59.3	59.3	54.3	56.8	50.6	63.4	52.4	65.7	53.7	67.7	56.9	58.1
	HDG-hod+hdg [86] (Depth)	26.6	33.6	17.9	19.3	34.4	26.2	40.5	27.6	28.6	11.6	38.4	29.0	27.8
	HDG-jpd+jmv [86] (Skel.)	61.0	61.8	59.3	56.0	60.0	57.4	68.8	54.2	71.1	57.2	69.7	59.0	61.3
	HDG-hod+hdg+jpd [86] (Depth + Skel.)	31.0	43.5	25.7	21.4	45.9	31.1	53.2	35.7	38.0	11.6	49.7	38.3	35.4
	HDG-hod+hdg+jmv [86] (Depth + Skel.)	59.0	62.2	58.1	52.0	62.5	57.1	66.0	54.2	67.7	52.7	70.3	61.1	60.2
	HDG-hodg+jpd+jmv [86] (Depth + Skel.)	58.2	61.8	54.8	47.6	63.5	58.7	69.0	52.3	64.9	47.1	67.2	59.4	58.7
	HDG-all features [86] (Depth + Skel.)	60.9	64.3	57.9	54.6	62.6	59.2	68.9	55.8	69.8	55.2	71.8	62.6	61.9
	SCK+DCK [23] (Skel.)	56.0	61.8	50.0	61.8	54.1	56.7	71.4	56.9	72.9	49.6	71.7	44.0	58.9
	Frames+CNN [21] (Skel.)	30.4	29.8	27.8	23.8	31.7	27.2	39.2	27.8	31.4	23.2	38.8	28.6	30.0
	Clips+CNN+Pooling [21] (Skel.)	31.6	33.3	30.6	27.8	35.3	29.6	44.3	31.3	35.3	23.2	43.1	31.0	33.0
Deep learning features	Clips+CNN+Concatenation [21] (Skel.)	33.2	36.9	32.5	29.8	36.9	31.2	46.3	31.0	39.2	23.6	45.5	31.7	34.8
	Clips+CNN+MTLN [21] (Skel.)	36.4	38.9	34.1	30.6	37.7	33.2	46.7	31.3	38.8	25.6	49.8	33.7	36.4
	HPM+AP [14] (Depth)	68.3	51.7	60.2	62.2	38.7	50.0	58.7	37.8	70.6	61.6	74.0	55.6	57.5
	HPM+TM [14] (Depth)	81.7	76.4	74.1	78.7	57.9	69.4	75.8	62.9	81.4	79.9	83.3	73.7	74.6
	1-layer RNN [18] (Skel.)	11.4	11.1	10.0	14.9	11.9	13.3	11.1	15.2	10.3	12.2	9.8	12.4	12.0
	2-layer RNN [18] (Skel.)	23.4	21.6	27.3	22.7	26.3	20.3	21.7	19.5	20.7	24.8	27.0	21.5	23.1
	1-layer LSTM [18] (Skel.)	28.9	12.8	19.2	15.5	20.0	33.1	48.2	16.5	43.3	15.5	38.7	16.4	25.7
	2-layer LSTM [18] (Skel.)	29.7	16.3	20.4	12.8	26.8	30.5	53.0	11.0	37.0	17.1	47.4	20.0	26.8
	1-layer P-LSTM [18] (Skel.)	26.8	23.5	22.4	22.7	24.4	31.3	49.6	20.3	38.3	19.5	45.9	17.6	28.5
	2-layer P-LSTM [18] (Skel.)	30.1	24.7	25.2	21.5	24.4	37.4	51.0	22.7	43.1	21.1	47.8	17.2	30.5
Our implementation with modified hyperparam. values:	2-layer P-LSTM (10 segments, 50 hidden neurons) (Skel.)	23.5	24.3	22.4	25.5	25.6	30.9	48.6	21.1	41.5	19.5	43.9	15.6	28.5
	2-layer P-LSTM (10 segments, 100 hidden neurons) (Skel.)	21.1	22.7	24.8	21.1	22.4	33.7	47.8	21.1	45.1	22.0	49.4	17.6	29.1
	2-layer P-LSTM (20 segments, 50 hidden neurons) (Skel.)	25.2	18.3	24.8	21.5	26.4	30.5	49.0	19.1	41.5	24.4	43.5	14.8	28.3
	2-layer P-LSTM (20 segments, 100 hidden neurons) (Skel.)	27.6	24.3	24.8	21.9	28.4	32.9	44.3	24.3	44.3	20.3	45.9	15.6	29.6
	IndRNN (4 layers) [27] (Skel.)	34.3	53.8	35.2	42.5	39.1	38.9	49.2	42.5	46.0	27.1	48.6	30.9	40.7
	IndRNN (6 layers) [27] (Skel.)	30.7	47.2	32.2	36.0	38.8	35.4	44.5	37.9	40.6	23.9	39.2	25.2	36.0
	IndRNN (4 layers, with jpd) (Skel.)	33.5	40.2	26.9	40.9	30.4	41.1	50.7	36.7	46.1	24.6	49.0	23.0	36.9
	IndRNN (6 layers, with jpd) (Skel.)	29.4	36.3	23.9	38.2	26.0	36.8	45.4	33.0	41.2	20.5	44.7	18.6	32.8
	ST-GCN [26] (Skel.)	36.4	26.2	20.6	30.2	33.7	22.4	43.1	26.6	16.9	12.8	26.3	36.5	27.6
	ST-GCN (with jpd) (Skel.)	29.6	21.8	15.5	19.1	17.1	18.8	35.3	14.3	31.0	14.8	13.3	15.5	20.5

*This result is obtained from [13] for comparison.

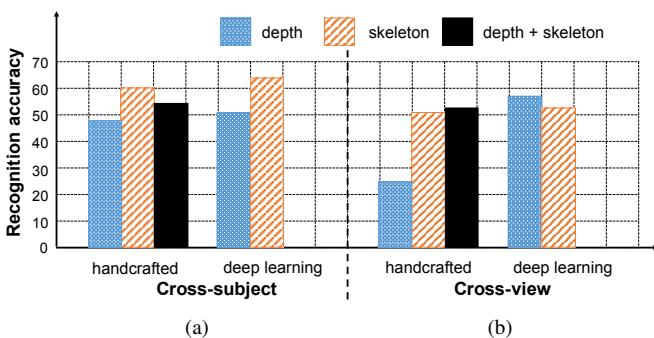


Fig. 4: The average recognition accuracy (in percentage) of methods using handcrafted and deep learning features for cross-subject and cross-view recognition. Numbers of methods using handcrafted (i) depth-based features: 7; (ii) skeleton-based features: 7; (iii) depth+skeleton-based features: 4. Numbers of methods using deep learning (i) depth-based features: 2; (ii) skeleton-based features: 20; (iii) depth+skeleton-based features: 0 (see Tables IV–VI).

According to plots in Fig. 5 the recognition accuracy was high when the left view V_2 and right view V_3 were used for training and front view V_1 for testing (see axis tick mark $V_2V_3-V_1$). This is obvious as the viewing angle of the front view V_1 is between V_2 and V_3 , as shown in Fig. 2. However, it is surprising that the recognition accuracy was even slightly higher for $V_2V_4-V_1$ and $V_3V_4-V_1$ (see Figs. 5(b) and 5(c)).

TABLE VII: The average recognition accuracies / [AVRank] of all the ten algorithms for both cross-subject and cross-view action recognition.

Algorithms	Cross-subject [†]	Cross-view [‡]
HON4D [11] (Depth)	52.8 / [3.6]	18.1 / [5.0]
HDG-all [86] (Depth + Skel.)	56.6 / [3.8]	60.1 / [1.5]
HOPC [13] (Depth)	55.9 / [3.4]	41.4 / [4.0]
LARP-SO-FTP [19] (Skel.)	65.0 / [2.4]	54.7 / [3.0]
SCK+DCK [23] (Skel.)	78.7 / [1.0]	66.5 / [1.5]
HPM+TM [14] (Depth)	58.7 / [3.0]	64.0 / [3.0]
Clips+CNN+MTLN [21] (Skel.)	74.3 / [2.4]	60.6 / [3.0]
P-LSTM [18] (Skel.)	64.0 / [3.0]	50.4 / [4.0]
IndRNN [27] (Skel.)	72.6 / [2.4]	62.1 / [2.0]
IndRNN with jpd (Skel.)	77.6 / [2.0]	60.8 / [1.5]
ST-GCN [26] (Skel.)	52.4 / [4.2]	58.2 / [3.5]
ST-GCN with jpd (Skel.)	57.7 / [4.0]	54.7 / [3.0]

[†]For cross-subject action recognition, the performance of each algorithm is computed by averaging all the recognition accuracies / rank values on the MSRA3D, 3D Action Pairs, CAD-60, UWA3D Activity and NTU RGB+D datasets. [‡]For cross-view action recognition, they are computed over the UWA3D Multiview Activity II and NTU RGB+D datasets.

One can also notice that for depth-based methods which use handcrafted features, the recognition accuracy dropped the most when the left view V_2 and top view V_4 were used for training, and the right view V_3 was used for testing. The

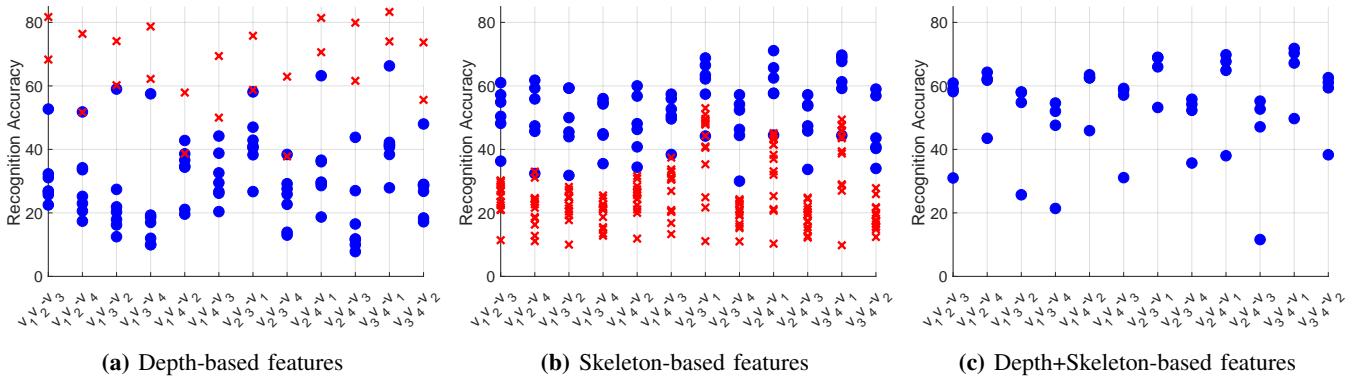


Fig. 5: Scatter plots showing the performance of cross-view action recognition on the UWA3D Multiview Activity II dataset. The blue dots and red crosses, respectively, represent methods using handcrafted features and deep learning features. On the horizontal axis of each plot, we use the notation $V_i V_j - V_k$ to denote views i and j being used for training and view k being used for testing.

main reason for this behavior is that the visual appearances of an action in V_2 and V_3 are different, and it is very difficult to find the same view-invariant features from these views. Other combinations of views, such as $V_1 V_3 - V_4$, $V_1 V_3 - V_2$, and $V_1 V_2 - V_4$, also led to a lower performance (see the distribution of the blue dots in Fig. 5(a)).

C. Depth-based features versus skeleton-based features

Cross-subject action recognition. It seems that for the cross-subject action recognition, skeleton-based features outperformed depth-based features for both the handcrafted and deep learning feature categories (Fig. 4(a)). However, it is surprising that adding the depth-based features to skeleton-based features actually led to a slight decrease in the action recognition accuracy. The main reason is the background clutter and noise (for instance, see Fig. 1) in the depth sequences, making the depth-based features less representative for robust action recognition.

On the other hand, some skeleton-based algorithms such as LARP-SO-FTP and its variants performed well in human action recognition because they only extracted reliable features from those human body joints which have high confidence values. The only exception is the HOPC algorithm which uses depth-based features and which performed better than skeleton-based features such as HDG-jpd+jmv in the cross-subject action recognition (see the average performance in the last column of Table IV). One must note that the HOPC algorithm is different from other depth-based algorithms as it treats the depth images as a 3D pointcloud. Such an approach allows the HOPC algorithm to estimate the orientations of the local surface patches on the body of the human subject to more robustly deal with viewpoint changes.

Cross-view action recognition. In Fig. 5(b), for the cross-view action recognition, methods using handcrafted features (blue dots) performed better than those using deep learning features (red crosses) for skeleton-based features. For depth-based features (Fig. 5(a)), no clear conclusion can be drawn for the handcrafted versus deep learning features.

Comparing the handcrafted features (blue dots) in both Figs. 5(a) and 5(b), we can see that algorithms using skeleton-based features produced better results than depth-based fea-

tures. This is evident from Fig. 5(b) as the blue dots are well above the red crosses, whereas in Fig. 5(a) the blue dots are more spread out and they occupy mainly the lower region of the plot. Our pruning process of the HDG algorithm also proved that skeleton-based features are more robust than the depth-based algorithms in the cross-view action recognition.

For the plot in Fig. 5(c), the algorithms in our experiments did not use deep learning to compute depth+skeleton features, thus no comparison on the performance between using handcrafted and deep learning features is possible.

D. Handcrafted features versus deep learning features

As expected, our experiments confirm that deep learning methods (*e.g.*, Clips+CNN+MTLN, IndRNN, and ST-GCN) performed better on large datasets such as the NTU RGB+D dataset, which has more than 56,000 video sequences, but achieved a lower recognition accuracy on the other smaller datasets. The main reason for this behavior is the reliance of deep learning methods on large amount of data for training to avoid overfitting (in contrast to handcrafted methods). However, most existing action datasets have small numbers of performing subjects and action classes, and very few camera views due to high cost of collecting/annotating data.

We also found that the performance of most handcrafted methods is highly dependent on the datasets. For example, HON4D and HOPC performed better only on some specific datasets, such as MSRACTION3D and 3D Action Pairs, but achieved lower performance on the NTU RGB+D dataset. This means that features that have been handcrafted to work on one dataset may not be transferable to another dataset. Moreover, as handcrafted methods prevent overfitting well, they may lack the capacity to learn from big data. With larger benchmark datasets becoming available to the research community, future research trend is more likely to shift to using deep learning features with efforts being put into tuning their hyperparameters.

E. ‘Quo Vadis, action recognition?’

Based on the methods evaluated in this paper, we see the following trends in human action recognition. Handcrafted features progressed from global descriptors (*e.g.*, HON4D

in 2013) to local descriptors (*e.g.*, HOPC in 2016) and combinations of global and local feature representation (*e.g.*, HDG in 2014). More recent works focus on designing robust 3D human joint-based representations (*e.g.*, LARP-SO-FTP and SCK+DCK in 2016). Moreover, researchers have adopted features extracted from skeleton sequences as they are easier to process and analyze than depth videos. We also note that deep learning representations are evolving from basic neural networks (*e.g.*, traditional RNN and LSTM) to adapted and/or dedicated networks that rely on a pre-trained network (*e.g.*, HPM+AP, HPM+TM, and Clips+CNN+MTLN). Researchers also modify basic neural network models to improve their learning capability for action recognition (*e.g.*, P-LSTM and IndRNN). Recently, new models such as ST-GCN were devised to robustly model the human actions in large datasets. Given the surge of large datasets and powerful GPU machines, the newest methods learn representations in an end-to-end manner.

VII. CONCLUSION

We have presented, analyzed and compared ten state-of-the-art algorithms in 3D action recognition on six benchmark datasets. These algorithms cover the use of handcrafted and deep learning features computed from depth and skeleton video sequences. We believe that our comparison results will be useful for future researchers interested in designing robust representations for recognizing human actions from videos captured by the Kinect camera and/or similar action sequence capturing sensors.

In our study, we found that skeleton-based features are more robust than depth-based features for both cross-subject action recognition and cross-view action recognition. Handcrafted features performed better than deep learning features given smaller datasets. However, deep learning methods achieved very good results if trained on large datasets. While accuracy as high as 90% has been achieved by some algorithms on cross-subject action recognition, the average accuracy on cross-view action recognition is much lower.

From our literature review and comprehensive experiments, we see that many research papers on human action recognition have already attempted to tackle the challenging issues mentioned in the introduction of the paper. Examples include using a skeleton model or treating depth maps as 3D point-cloud to overcome the changes of viewpoint; using Fourier temporal pyramid or dynamic time warping to deal with different speeds of execution of actions; using depth videos to overcome changes in lighting conditions and eliminate visual appearances that may be irrelevant to the actions performed. While these papers all report promising action recognition results, new and more robust action recognition algorithms are still required, given that there is a pressing demand for using these algorithms in real and new environments.

ACKNOWLEDGMENTS

We would like to thank the authors of HON4D [11], HOPC [13], LARP-SO [19], HPM+TM [14], IndRNN [27] and ST-GCN [26] for making their codes publicly available.

We thank the ROSE Lab of Nanyang Technological University (NTU), Singapore, for making the NTU RGB+D dataset freely accessible. We would like to thank Data61/CSIRO for providing the computing cluster for running our experiments.

REFERENCES

- [1] A. F. Bobick and J. W. Davis, "The Recognition of Human Movement Using Temporal Templates," *TPAMI*, vol. 23, no. 3, pp. 257–267, March 2001.
- [2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," *ICCV*, pp. 1–8, 2005.
- [3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *ICCV*, pp. 1–8, 2005.
- [4] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal Shape and Flow Correlation for Action Recognition," *Proc. 7th Int. Workshop on Visual Surveillance*, 2007.
- [5] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," *CVPR*, pp. 1–8, 2008.
- [6] J. Liu and M. Shah, "Learning Human Actions via Information Maximization," *CVPR*, pp. 1–9, 2008.
- [7] M. Breganzo, S. Gong, and T. Xiang, "Recognising Actions as Clouds of Space-Time Interest Points," *CVPR*, pp. 1948–1955, 2009.
- [8] J. Liu, J. Luo, and M. Shah, "Recognizing Realistic Actions from Videos in the Wild," *CVPR*, pp. 1–8, 2009.
- [9] W. Li, Z. Zhang, and Z. Liu, "Action Recognition Based on A Bag of 3D Points," in *CVPR*, 2010, pp. 9–14.
- [10] X. Yang and Y. Tian, "EigenJoints-based Action Recognition Using Naive-Bayes-Nearset-Neighbor," in *CVPR*, 2012, pp. 14–19.
- [11] O. Oreifej and Z. Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," in *CVPR*, 2013, pp. 716–723.
- [12] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real Time Action Recognition Using Histograms of Depth Gradients and Random Decision Forests," in *WACV*, 2014, pp. 626–633.
- [13] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Histogram of Oriented Principal Components for Cross-View Action Recognition," *TPAMI*, pp. 2430–2443, December 2016.
- [14] H. Rahmani and A. Mian, "3D Action Recognition from Novel Viewpoints," *CVPR*, pp. 1–12, 2016.
- [15] T. Mallick, P. P. Das, and A. K. Majumdar, "Characterizations of Noise in Kinect Depth Images: A Review," *IEEE SEN*, vol. 14, no. 6, pp. 1731–1740, June 2014.
- [16] 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 *CVPR*, 2011, pp. 1297–1304.
- [17] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," *CVPR*, pp. 588–595, 2014.
- [18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," *CVPR*, pp. 1010–1019, 2016.
- [19] R. Vemulapalli and R. Chellappa, "Rolling Rotations for Recognizing Human Actions from 3D Skeletal Data," *CVPR*, pp. 4471–4479, 2016.
- [20] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "SkeletonNet: Mining Deep Part Features for 3D Action Recognition," *IEEE SPL*, vol. 24, no. 6, pp. 731–735, 2017.
- [21] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A New Representation of Skeleton Sequences for 3D Action Recognition," in *CVPR*, June 2017.
- [22] H. Rahmani and M. Bennamoun, "Learning Action Recognition Model From Depth and Skeleton Videos," in *ICCV*, 2017, pp. 5832–5841.
- [23] P. Koniusz, A. Cherian, and F. Porikli, "Tensor Representations via Kernel Linearization for Action Recognition from 3D Skeletons," *ECCV*, pp. 1–19, 2016.
- [24] J. Zhang, H. P. H. Shum, J. Han, and L. Shao, "Action Recognition From Arbitrary Views Using Transferable Dictionary Learning," in *TIP*, 2018, pp. 4709–4723.
- [25] H. Rahmani, A. Mian, and M. Shah, "Learning a Deep Model for Human Action Recognition from Novel Viewpoints," *TPAMI*, pp. 1–14, 2015.
- [26] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," in *AAAI*, 2018.
- [27] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN," *CVPR*, pp. 5457–5466, 2018.
- [28] C. Li, Z. Cui, W. Zheng, C. Xu, and J. Yang, "Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition," in *AAAI*, 2018, pp. 3482–3489.
- [29] H. Wang and L. Wang, "Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks," in *CVPR*, 2017, pp. 1–10.
- [30] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep Progressive Reinforcement Learning for Skeleton-based Action Recognition," in *CVPR*, 2018, pp. 1–10.
- [31] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global Context-Aware Attention LSTM Networks for 3D Action Recognition," in *CVPR*, 2017, pp. 1–10.
- [32] Z. Huang, C. Wan, T. Probst, and L. V. Gool, "Deep Learning on Lie Groups for Skeleton-based Action Recognition," in *CVPR*, 2017, pp. 6099–6108.
- [33] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble Deep Learning for Skeleton-based Action Recognition using Temporal Sliding LSTM networks," in *ICCV*, 2017, pp. 1012–1020.
- [34] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning," in *ECCV*, 2018, pp. 1–16.
- [35] H. Wang and L. Wang, "Beyond Joints: Learning Representations From Primitive Geometries for Skeleton-Based Action Recognition and Detection," in *TIP*, 2018, pp. 4382–4394.

- [36] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates," in *TPAMI*, 2018, pp. 1–14.
- [37] A. B. Tanfous, H. Drira, and B. B. Amor, "Coding Kendall's Shape Trajectories for 3D Action Recognition," in *CVPR*, 2018, pp. 2840–2849.
- [38] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised Representation Learning with Long-Term Dynamics for Skeleton Based Action Recognition," in *AAAI*, 2018, pp. 2644–2651.
- [39] K. P. Fishkin, M. Philipose, and A. Rea, "Hands-On RFID: Wireless Wearables for Detecting Use of Objects," *IEEE ISWC*, 2005.
- [40] M. R. Hodges and M. E. Pollack, "An 'Object-Use Fingerprint': The Use of Electronic Sensors for Human Identification," *UbiComp*, 2007.
- [41] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall, "Recognizing Daily Activities with RFID-Based Sensors," *UbiComp*, 2009.
- [42] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human Activity Detection from RGBD Images," *PAIR*, 2011.
- [43] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured Human Activity Detection from RGBD Images," in *ICRA*, 2012.
- [44] H. S. Koppula, R. Gupta, and A. Saxena, "Learning Human Activities and Object Affordances from RGB-D Videos," *IJRR*, Jan 2013.
- [45] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning Rich Features from RGB-D Images for Object Detection and Segmentation," *ECCV*, pp. 1–16, 2014.
- [46] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic Image Networks for Action Recognition," *CVPR*, pp. 3034–3042, 2016.
- [47] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal Multipart Learning for Action Recognition in Depth Videos," *TPAMI*, vol. 38, no. 10, pp. 2123–2129, 2016.
- [48] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal Multiplier Networks for Video Action Recognition," *CVPR*, pp. 4768–4777, 2017.
- [49] A. Cherian, N. Rai, K. Sikka, and G. Sharma, "AdaScan: Adaptive Scan Pooling in Deep Convolutional Neural Networks for Human Action Recognition in Videos," *CVPR*, pp. 3376–3385, 2017.
- [50] A. Cherian, P. Koniusz, and S. Gould, "Higher-order Pooling of CNN Features via Kernel Linearization for Action Recognition," *WACV*, pp. 130–138, 2017.
- [51] A. Cherian, B. Fernando, M. Harandi, and S. Gould, "Generalized Rank Pooling for Activity Recognition," *CVPR*, pp. 3222–3231, 2017.
- [52] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos," *TPAMI*, pp. 1045–1058, 2018.
- [53] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *CVPR*, 2017, pp. 1–10.
- [54] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *CVPR*, 2016, pp. 1–9.
- [55] J.-F. Hu, W.-S. Zheng, J. Pan1, J. Lai, and J. Zhang, "Deep Bilinear Learning for RGB-D Action Recognition," in *ECCV*, 2018, pp. 1–17.
- [56] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," *CVPR*, pp. 1290–1297, 2012.
- [57] L. Xia and J. K. Aggarwal, "Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera," in *CVPR*, 2013, pp. 2834–2841.
- [58] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition," in *ECCV*, 2014, pp. 742–757.
- [59] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Action Classification with Locality-constrained Linear Coding," in *ICPR*, 2014, pp. 3511–3516.
- [60] X. Yang and Y. Tian, "Super Normal Vector for Activity Recognition using Depth Sequences," in *CVPR*, 2014, pp. 804–811.
- [61] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Deep Convolutional Neural Networks for Action Recognition Using Depth Map Sequences," in *CVPR*, 2015, pp. 1–8.
- [62] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action Recognition From Depth Maps Using Deep Convolutional Neural Networks," *IEEE T HUM-MACH SYST*, pp. 498–509, 2016.
- [63] H. Rahmani, D. Q. Huynh, A. Mahmood, and A. Mian, "Discriminative Human Action Classification using Locality-constrained Linear Coding," *Pattern Recognit. Lett.*, pp. 62–71, 2016.
- [64] C. Zhang, Y. Tian, X. Guo, and J. Liu, "DAAL: Deep Activation-based Attribute Learning for Action Recognition in Depth Videos," *CVIU*, pp. 37–49, 2017.
- [65] Z. Shi and T.-K. Kim, "Learning and Refining of Privileged Information-based RNNs for Action Recognition from Depth Sequences," in *CVPR*, 2017, pp. 3461–3470.
- [66] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints," *CVPRW*, 2012.
- [67] Y. Du, W. Wang, and L. Wang, "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition," in *CVPR*, 2015, pp. 1110–1118.
- [68] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential Recurrent Neural Networks for Action Recognition," *CVPR*, pp. 1–9, 2015.
- [69] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks," *AAAI*, 2016.
- [70] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition," *ECCV*, pp. 816–833, 2016.
- [71] B. B. Amor, J. Su, and A. Srivastava, "Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories," *TPAMI*, vol. 38, no. 1, pp. 1–13, 2016.
- [72] N. E. D. Elmaldany, Y. He, and L. Guan, "Information Fusion for Human Action Recognition via Biset/Multiset Globality Locality Preserving Canonical Correlation Analysis," in *TIP*, 2018, pp. 5275–5287.
- [73] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," in *CVPR*, 2019, pp. 1227–1236.
- [74] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition," in *CVPR*, 2019, pp. 3595–3603.
- [75] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition with Directed Graph Neural Networks," in *CVPR*, 2019, pp. 7912–7921.
- [76] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in *CVPR*, 2019, pp. 12027–12035.
- [77] Y. Tas and P. Koniusz, "CNN-based Action Recognition and Supervised Domain Adaptation on 3D Body Skeletons via Kernel Feature Maps," in *BMVC*, 2018.
- [78] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, pp. 1735–1780, 1997.
- [79] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact Bilinear Pooling," *CVPR*, pp. 1–10, 2016.
- [80] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *ICLR*, pp. 1–14, 2015.
- [81] R. Caruana, *Multitask Learning*, ser. 1. Springer, 1997, vol. 28, no. 41–75.
- [82] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral Networks and Deep Locally Connected Networks on Graphs," in *ICLR*, 2014, pp. 1–14.
- [83] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," in *NIPS*, 2016, pp. 1–9.
- [84] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *ICLR*, 2017, pp. 1–14.
- [85] P. Koniusz, L. Wang, and A. Cherian, "Tensor Representations for Action Recognition," *CoRR*, 2019.
- [86] L. Wang, "Analysis and Evaluation of Kinect-based Action Recognition Algorithms," Master's thesis, School of the Computer Science and Software Engineering, The University of Western Australia, Nov 2017.



Lei Wang received the M.E. degree in software engineering from The University of Western Australia (UWA), Australia, in 2018. Since then, he has worked as a Computer Vision Researcher at iCetana Pty Ltd. He was a Visiting Researcher in Machine Learning Research Group at Data61/CSIRO, and he is a Visiting Researcher with the Department of Computer Science and Software Engineering, UWA. He is also pursuing a Ph.D. degree under the supervision of Dr. Piotr Koniusz and A/Prof Du Q. Huynh at the Australian National University (ANU) and Data61/CSIRO. His research interests include human action recognition in videos, machine learning and computer vision.



Du Q. Huynh received the PhD degree in computer vision from The University of Western Australia, Crawley, Australia, in 1994. Since then, she has worked for the Australian Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP) and Murdoch University. She is currently an associate professor with the Department of Computer Science and Software Engineering, The University of Western Australia. Her research interests include visual target tracking, video image processing, machine learning, and pattern recognition. She has previously worked on shape from motion, multiple view geometry, and 3D reconstruction.



Piotr Koniusz A Senior Researcher in Machine Learning Research Group at Data61/CSIRO, formerly known as NICTA, and a Senior Honorary Lecturer at Australian National University (ANU). Previously, he worked as a postdoctoral researcher in the team LEAR, INRIA, France. He received his BSc degree in Telecommunications and Software Engineering in 2004 from the Warsaw University of Technology, Poland, and completed his PhD degree in Computer Vision in 2013 at CVSSP, University of Surrey, UK. His interests include visual categorization, spectral learning on graphs and tensors, kernel methods and deep learning.