NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot

Abstract—Research on depth-based human activity analysis achieved outstanding performance and demonstrated the effectiveness of 3D representation for action recognition. The existing depth-based and RGB+D-based action recognition benchmarks have a number of limitations, including the lack of large-scale training samples, realistic number of distinct class categories, diversity in camera views, varied environmental conditions, and variety of human subjects. In this work, we introduce a large-scale dataset for RGB+D human action recognition, which is collected from 106 distinct subjects and contains more than 114 thousand video samples and 8 million frames. This dataset contains 120 different action classes including daily, mutual, and health-related activities. We evaluate the performance of a series of existing 3D activity analysis methods on this dataset, and show the advantage of applying deep learning methods for 3D-based human action recognition. Furthermore, we investigate a novel one-shot 3D activity recognition problem on our dataset, and a simple yet effective Action-Part Semantic Relevance-aware (APSR) framework is proposed for this task, which yields promising results for recognition of the novel action classes. We believe the introduction of this large-scale dataset will enable the community to apply, adapt, and develop various data-hungry learning techniques for depth-based and RGB+D-based human activity understanding. [*The dataset is available at: http:// rose1.ntu.edu.sg/ Datasets/ actionRecognition.asp*.]

Index Terms—Activity Understanding, Video Analysis, 3D Action Recognition, RGB+D Vision, Deep Learning, Large-Scale Benchmark.

1 Introduction

The development of depth sensors, *e.g.*, Microsoft Kinect, Intel RealSense, and Asus Xtion, enables us to obtain effective 3D structure information of the objects and scenes [1]. This empowers the computer vision solutions to move important steps towards 3D vision, such as 3D object recognition [2], 3D scene understanding [3], and 3D activity analysis [4], *etc*. Unlike RGB video-based activity analysis [5], [6], [7], [8], 3D action recognition suffers from the lack of large-scale benchmark datasets. There are no publicly shared sources like YouTube to supply "in-the-wild" 3D video samples of a realistically various set of action classes. This limits our ability to build large-sized benchmarks to train, evaluate, and compare the strengths of different approaches, especially the recent data-hungry techniques like deep learning methods. To the best of our knowledge, all the currently available 3D action recognition benchmarks have limitations in various aspects.

First is the small number of subjects and narrow range of performers' ages. This can significantly limit the intra-class variation of the actions. The constitution of human activities depend on the gender, age, physical condition, and even cultural aspects of the subjects. As a result, variation of human subjects is crucial for building a realistic action recognition benchmark.

The second factor is the limited number of action categories. When only a small set of classes are available, each can be very distinguishable by finding a simple motion pattern or even by the appearance of an interacted object. However, when the number of classes grows, similar motion patterns and interacted objects will be shared among different classes, which makes the action recognition much more challenging.

The third limitation is the highly restricted camera views. In most of the current datasets, the samples are captured from a front view with a fixed camera viewpoint. In some others, the views are often bounded to fixed front and side views using multiple cameras at the same time.

The fourth factor is the limited variation of the collection environments (*e.g.*, backgrounds) which can also be important to achieve a sensible activity analysis dataset.

Finally and most importantly, the very limited number of video samples hinders the application of the advanced data-driven learning methods to this problem. Though several attempts have been done [9], [10], they mostly suffer from over-fitting and have to scale down the size of their learning models. Therefore, they clearly need many more samples to generalize and perform better on the testing videos.

In order to overcome these limitations, a large-scale benchmark dataset, NTU RGB+D 120 dataset, is developed for 3D human activity analysis.

The proposed dataset consists of $114,480$ RGB+D video samples that are captured from 106 distinct human subjects. We have collected RGB videos, depth sequences, skeleton data (3D locations of 25 major body joints), and infrared frames using Microsoft Kinect v2. The action samples are captured from 155 different camera viewpoints. The subjects in this dataset are in a wide range of age distribution (from

- *J. Liu, M. Perez, and A. C. Kot are with ROSE Lab, School of EEE, Nanyang Technological University, Singapore. E-mail: {jliu029, mauricio001, eackot}@ntu.edu.sg.*
- *A. Shahroudy is with Department of Electrical Engineering, Chalmers University of Technology, Sweden. E-mail: amirsh@chalmers.se.*
- *L.-Y. Duan is with National Engineering Laboratory for Video Technology, Peking University, China, and also with Peng Cheng Laboratory, China. E-mail: lingyu@pku.edu.cn.*
- *G. Wang is with Alibaba Group, China. E-mail: wanggang@ntu.edu.sg.*

TABLE 1: Comparison of the proposed NTU RGB+D 120 dataset and some of the other publicly available datasets for 3D action recognition. Our dataset provides many more video samples, action classes, human subjects, and camera views in comparison with other available datasets for RGB+D action recognition.

Datasets		#Videos	#Classes	#Subjects	#Views	Sensors	Data Modalities	Year
MSR-Action3D	[11]	567	20	10	1	N/A	D+3DJoints	2010
CAD-60	[12]	60	12	4	-	Kinect v1	RGB+D+3DJoints	2011
RGBD-HuDaAct	[13]	1,189	13	30	1	Kinect v1	RGB+D	2011
MSRDailyActivity3D	[14]	320	16	10	1	Kinect v1	RGB+D+3DJoints	2012
UT-Kinect	[15]	200	10	10	4	Kinect v1	RGB+D+3DJoints	2012
Act4^2	[16]	6,844	14	24	4	Kinect v1	RGB+D	2012
CAD-120	[17]	120	10+10	4	-	Kinect v1	RGB+D+3DJoints	2013
3D Action Pairs	[18]	360	12	10	1	Kinect v1	RGB+D+3DJoints	2013
Multiview 3D Event	[19]	3,815	8	8	3	Kinect v1	RGB+D+3DJoints	2013
Northwestern-UCLA	[20]	1,475	10	10	3	Kinect v1	RGB+D+3DJoints	2014
UWA3D Multiview	[21]	∼900	30	10	1	Kinect v1	RGB+D+3DJoints	2014
Office Activity	[22]	1,180	20	10	3	Kinect v1	RGB+D	2014
UTD-MHAD	[23]	861	27	8	1	Kinect v1+WIS	RGB+D+3DJoints+ID	2015
UWA3D Multiview II	[24]	1,075	30	10	5	Kinect v1	RGB+D+3DJoints	2015
M^{2}I	[25]	∼1,800	22	22	2	Kinect v1	RGB+D+3DJoints	2015
SYSU 3DHOI	[26]	480	12	40	1	Kinect v1	RGB+D+3DJoints	2017
NTU RGB+D 120		**114,480**	**120**	**106**	**155**	**Kinect v2**	**RGB+D+3DJoints+IR**	

It is worth mentioning that most of the other datasets were collected on a single or few backgrounds and under fixed illumination condition, while there is high variation of environmental conditions in our dataset, which uses 96 different backgrounds and contains significant illumination variation.

10 to 57) and from different cultural backgrounds (15 countries), which brings very realistic variation to the quality of actions. We also provide the ambiance inconstancy by capturing the dataset under various environmental conditions (96 different backgrounds with illumination variation).

The large amount of variation in subjects, views, and backgrounds makes it possible to have more sensible cross-subject and cross-setup evaluations for various 3D-based action analysis methods. The proposed dataset will help the community to move steps forward in 3D human activity analysis, and make it possible to apply data-hungry methods, such as deep learning techniques, for this task.

The performance of state-of-the-art 3D action recognition approaches is evaluated on our dataset, which shows the capability of applying deep models for activity analysis with the suggested cross-subject and cross-setup evaluation criteria. We also evaluate the performance of fusion across different data modalities, *e.g.*, RGB, depth, and skeleton data, for activity analysis, since they provide appearance and geometrical information respectively, and are complementary for more accurate action recognition.

In this paper, we also investigate a novel one-shot 3D action recognition problem based on the proposed dataset. An Action-Part Semantic Relevance-aware (APSR) framework is proposed to handle this task by utilizing the semantic relevance between each body part and each action class at the distributed word embedding level [27], [28]. Human actions can be represented by a combination of the movements of different body parts [29], [30], and the importance degrees of body parts' motion are not equal for recognizing different action categories. As a result, we need to put more emphasis on the more relevant body parts when recognizing an action performed by a human. In this paper, we show that the name (text description) of the novel action class can assist in the identification of the relevant body parts, and by exploiting the semantic relevance between the action's and body part's descriptions as a guidance, the relevant body parts of the novel action categories can be emphasized, and thus the one-shot recognition performance is improved.

The rest of this paper is organized as follows. Section 2 reviews the current 3D-based human action recognition approaches and benchmarks. Section 3 introduces the proposed dataset, its structure, and the defined evaluation criteria. Section 4 presents the proposed APSR framework for one-shot 3D human action recognition. Section 5 shows the experimental evaluations with our benchmark. Finally, section 6 concludes the paper.

2 Related Work

We briefly review publicly available benchmark datasets and recent methods for 3D human activity analysis in this section. Here we introduce a limited number of the most famous ones. Readers are referred to these survey papers [31], [32], [33], [34], [35], [36] for a more extensive list of the current 3D activity analysis datasets and methods.

2.1 3D Activity Analysis Datasets

After the release of the Microsoft Kinect [37], several datasets have been collected to conduct research on 3D

human action recognition and to evaluate different methods in this field.

The MSR-Action3D dataset [11] was the earliest which opened up the research in depth-based action analysis. The samples of this dataset were limited to depth sequences of gaming actions, *e.g.*, *forward punching, side boxing, forward kicking, side kicking, tennis swinging, tennis serving, golf swinging, etc*. Later, the skeleton data was added to this dataset. The skeletal information includes the 3D locations of 20 different joints at each frame. A decent number of methods have been evaluated on this benchmark, and the recent ones reported close to saturation accuracies [38], [39].

The MSR-DailyActivity dataset [14] was among the most challenging benchmarks in this field. It contains 320 samples of 16 daily activities with higher intra-class variation. The small number of samples and fixed camera viewpoints are the limitations of this dataset. Some reported results on this dataset also achieved very high accuracies [40], [41], [42], [43].

The RGBD-HuDaAct dataset [13] was one of the largest datasets. It contains RGB and depth sequences of 1189 videos of 12 human daily actions (plus one background class), with high variation in time lengths. The special characteristic of this dataset is the synced and aligned RGB and depth channels, which enables local multi-modal analysis of RBGD [1] signals.

The CAD-60 [12] and CAD-120 [17] datasets contain RGB, depth, and skeleton data of human actions. The special characteristic of these datasets is the variety of camera views. Unlike most of the other datasets, the cameras in these two datasets were not bound to front-view or side-views. However, the limited number of video samples (60 and 120) is the downside of them.

The 3D Action Pairs dataset [18] was proposed to provide multiple pairs of action classes. Each pair contains very closely related actions with differences along temporal axis, *e.g.*, *pick up/put down a box, push/pull a chair, put on/take off a hat, etc*. State-of-the-art methods [38], [43], [44] achieved perfect accuracy on this benchmark.

The Northwestern-UCLA [20] and the Multiview 3D Event [19] datasets used more than one depth sensors at the same time to collect multi-view representations of the same action, and scale up the number of samples.

The G3D [45] and PKUMMD [46] datasets were introduced for activity analysis in continuous sequences, which respectively contain 210 and 1076 videos. In each dataset, the videos were collected in the same environment.

It is worth mentioning that there are more than 40 datasets for 3D human action recognition [31]. Though each of them provided important challenges of human activity analysis, they have limitations in some aspects. TABLE 1 shows the comparison between some of the current datasets and our large-scale RGB+D action recognition dataset.

By summarizing the contributions of our dataset over the existing ones, NTU RGB+D 120 dataset has: (1) many more action classes; (2) many more video samples for each action class; (3) much more intra-class variation, *e.g.*, poses, interacted objects, ages and cultural backgrounds of the actors; (4) many more collection environments, *e.g.*, different backgrounds and illumination conditions; (5) more camera views; (6) more camera-to-subject distance variation; (7) used Kinect v2 which provides more accurate depth-maps and 3D joints, especially in a multi-camera setup, compared to the previous version of Kinect.

1. We emphasize the difference between RGBD and RGB+D terms. We suggest using RGBD when the two modalities are aligned pixel-wise, and RGB+D when the resolutions of the two are different and frames are not aligned.

TABLE 2: Comparison of the dataset version introduced in this paper and the one released in our preliminary conference paper [47]. The top three rows show a comparison of the sizes of these two versions. The bottom two rows show a comparison of the recognition accuracies when evaluating several methods on it. The methods [48], [49] are evaluated by using the suggested two evaluation criteria.

Dataset Version	Preliminary [47] (NTU RGB+D 60)	Extended (NTU RGB+D 120)
#Videos	56,880	114,480
#Classes	60	120
#Subjects	40	106
ST-LSTM [48]	69.2% 77.7%	55.7% 57.9%
MTLN [49]	79.6% 84.8%	58.4% 57.9%

This work is an extension of our previous conference paper [47]. In [47], we introduced the preliminary version of our dataset that contains 60 action classes. In this paper, we significantly extend it and build the NTU RGB+D 120 dataset that is much larger and provides much more variation of environmental conditions, subjects, and camera views, *etc*. It also provides more challenges of 3D human activity analysis. A brief comparison between these two versions is shown in TABLE 2. Besides, in this paper, we propose a new framework for one-shot 3D action recognition based on the proposed NTU RGB+D 120 dataset.

2.2 3D Action Recognition Methods

3D action recognition by hand-crafted features. Since the introduction of the first few benchmarks, such as the MSR-Action3D [11] and MSR-DailyActivity [14] datasets, a decent number of feature extraction and classifier learning methods have been proposed and evaluated based on them.

Oreifej and Liu [18] proposed to calculate the four-dimensional normals (X-Y-depth-time) from depth sequences, and accumulate them on spatio-temporal cubes as quantized histograms over 120 vertices of a regular polychoron. Yang and Tian [50] proposed supernormal vectors as aggregated dictionary-based codewords of four-dimensional normals over space-time grids. The work of [24] introduced histograms of oriented principle components of depth cloud points to extract robust features against viewpoint variations. Lu *et al.* [40] applied τ test-based binary range-sample features on depth maps and achieved robust representation against noise, scaling, camera views, and background clutter.

To have view-invariant representations of the actions, features can be extracted from the 3D body joint positions [51], [52] which are available for each frame. Evangelidis

et al. [53] divided the body into part-based joint quadruples, and encoded the configuration of each part with a succinct 6D feature vector, so called skeletal quads. To aggregate the skeletal quads, they applied Fisher vectors, and classified the samples by a linear SVM. In [54], different skeleton configurations were represented as points on a Lie group. Actions as time-series of skeletal configurations were encoded as curves on this manifold. The work of [42] utilized group sparsity-based class-specific dictionary coding with geometric constraints to extract skeleton features.

In most of the 3D action recognition scenarios, there are more than one modality of information, and combining different data modalities can help to improve the classification accuracy. Ohn-Bar and Trivedi [55] combined second order joint-angle similarity representations of skeletons with a modified two step HOG feature on spatio-temporal depth maps to build global representation of each video sample, and utilized a linear SVM to classify the actions. Wang *et al.* [56] combined Fourier temporal pyramids of skeletal information with local occupancy pattern features extracted from depth maps, and applied a data mining framework to discover the most discriminative combinations of body joints. A structured sparsity-based multi-modal feature fusion technique was introduced by [57] for action recognition in RGB+D domain. In [58], random decision forests were utilized for learning and feature pruning over a combination of depth and skeleton-based features. Hu *et al.* [41] introduced a joint heterogeneous feature learning framework by combing RGB, depth, and skeleton data for human activity recognition. The work of [38] proposed hierarchical mixed norms to fuse different features and select most informative body parts in a joint learning framework.

3D action recognition with deep networks. Recently, deep learning based-approaches have been proposed for 3D human activity analysis [9], [59], [60]. Specifically, many of them have been evaluated based on the preliminary version [47] of our dataset, or pre-trained on it for transfer learning for other tasks [43], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79].

Some approaches used recurrent neural networks (RNNs) to model the motion dynamics and context dependencies for 3D human action recognition. Du *et al.* [9] proposed a multi-layer RNN framework for 3D action recognition based on a hierarchy of skeleton-based inputs. Liu *et al.* [48] introduced a Spatio-Temporal LSTM network by modeling the context information in both temporal and spatial dimensions. Zhang *et al.* [80] added a view-adaptation scheme to the LSTM network to regulate the observation viewpoints. Luo *et al.* [61] proposed an unsupervised learning method by using a LSTM encoder-decoder framework for action recognition in RGB+D videos.

Convolutional neural networks (CNNs) have also been applied to 3D human action recognition. Wang *et al.* [81] proposed a "scene flow to action map" representation for RGB+D based action recognition with CNNs. Ke *et al.* [82] transformed the 3D skeleton data to ten feature arrays and input them to CNNs for action recognition. Rahmani *et al.* [66] designed a deep CNN model to transfer the visual appearance of human body-parts acquired from different views to a view-invariant space for depth-based activity analysis.

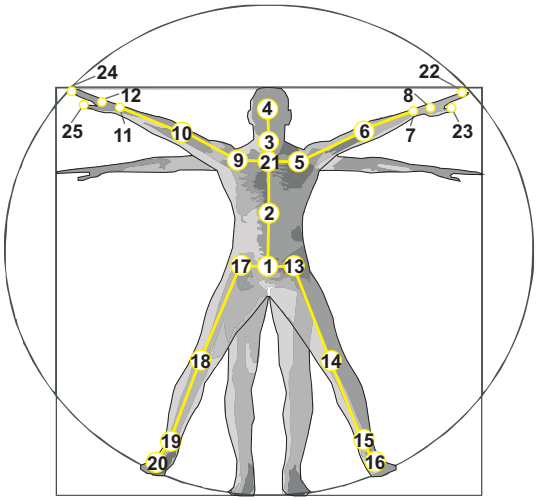

Fig. 1: Illustration of the configuration of 25 body joints in our dataset. The labels of these joints are: (1) base of spine, (2) middle of spine, (3) neck, (4) head, (5) left shoulder, (6) left elbow, (7) left wrist, (8) left hand, (9) right shoulder, (10) right elbow, (11) right wrist, (12) right hand, (13) left hip, (14) left knee, (15) left ankle, (16) left foot, (17) right hip, (18) right knee, (19) right ankle, (20) right foot, (21) spine, (22) tip of left hand, (23) left thumb, (24) tip of right hand, (25) right thumb.

Beside RNNs and CNNs, some other deep models have also been introduced for 3D human action recognition. Huang *et al.* [62] incorporated Lie group structure into a deep architecture for skeleton-based action recognition. Tang *et al.* [77] applied deep progressive reinforcement learning to distil the informative frames in the video sequences. Rahmani and Mian [83] introduced a nonlinear knowledge transfer model to transform different views of the human actions to a canonical view for action classification.

One-shot 3D Action Recognition. Plenty of advanced techniques, such as metric learning [84] and meta learning [85], [86], have been introduced for one-shot object recognition and image classification [87], [88], [89]. In the context of 3D activity analysis, there are also a few attempts on one-shot-based learning.

Fanello *et al.* [90] used 3D-HOF and Global-HOG features for one-shot action recognition in RGB+D videos. Wan *et al.* [91] extracted mixed features around the sparse keypoints that are robust to scale, rotation and partial occlusions. Konečnỳ *et al.* [92] combined HOG and HOF descriptors together with the dynamic time warping technique for one-shot RGB+D-based action recognition.

Different from these works, a simple yet effective APSR framework is introduced in this paper, which emphasizes the features of the relevant body parts by considering the semantic relevance of the action class and each body part, for one-shot 3D action recognition.

3 The NTU RGB+D 120 Dataset

In this section, we introduce the details of the proposed NTU RGB+D 120 action recognition dataset and the defined evaluation criteria.

3.1 Dataset Structure

3.1.1 Data Modalities

We use Microsoft Kinect sensors to collect our dataset. We collect four major data modalities acquired by this sensor, namely, the depth maps, the 3D joint information, the RGB frames, and the infrared (IR) sequences.

The depth maps are sequences of two dimensional depth values in millimeters. To maintain all the information, we apply lossless compression for each individual frame. The resolution of each depth frame is 512×424.

The joint information consists of 3-dimensional locations of 25 major body joints for each detected and tracked human body in the scene. The corresponding pixels on RGB frames and depth maps are also provided for each body joint. The configuration of these joints is illustrated in Fig. 1.

The RGB videos are recorded in the provided resolution of 1920×1080.

The infrared sequences are also collected and stored frame by frame at the resolution of 512×424.

3.1.2 Action Classes

We have 120 action categories in total, which are divided into three major groups, including 82 daily actions (eating, writing, sitting down, moving objects, *etc*), 12 health-related actions (blowing nose, vomiting, staggering, falling down, *etc*), and 26 mutual actions (handshaking, pushing, hitting, hugging, *etc*).

Compared to the preliminary version [47] of our dataset, the proposed NTU RGB+D 120 dataset contains many more action classes. Here we summarise the characteristics of the newly added actions compared to the actions in the preliminary version: *(1) Fine-grained hand/finger motions.* Most of the actions in the preliminary version of our dataset have significant body and hand motions, while the newly added classes in this extended version contain some actions that have fine-grained hand and finger motions, such as "make ok sign" and "snapping fingers". *(2) Fine-grained object-related individual actions.* The newly added actions include some fine-grained object-involved single-person actions, in which the body movements are not significant and the sizes of the involved objects are relatively small, such as "counting money" and "play magic cube". *(3) Object-related mutual actions.* Most of the two-person mutual actions in the preliminary version do not involve objects. In this extended version, some of the newly added mutual actions involve the interactions with objects, such as "wield knife towards other person" and "hit other person with object". *(4) Different actions with similar posture patters but with different motion speeds.* In this extended version, there are some different actions that have similar posture patterns but have different motion speeds. For example, "grab other person's stuff" is a newly added action, and its main difference compared to "touch other person's pocket (steal)" is the motion speed. *(5) Different actions with similar body motions but with different objects involved.* There are some different actions that have very similar body motions but involve different objects. For example, the motions in the newly added action "put on bag/backpack" are similar to those in the existing action "put on jacket". *(6) Different actions with similar objects involved but with different body motions.* Among the newly added actions, there are also some different classes that share the same interacted objects, such as "put on bag/backpack" and "take something out of a bag/backpack".

3.1.3 Subjects

We invited 106 distinct subjects to our dataset collection sessions. These subjects are from 15 different countries. Their ages are between 10 and 57, and heights are between $1.3m$ and $1.9m$. Fig. 9 illustrates the variety of the subjects in age, gender, and height. Each subject is assigned a consistent ID number over the entire dataset.

3.1.4 Collection Setups

We use 32 collection setups to build our dataset, and over different setups, we change the location and background. Specifically, in each setup, we use three cameras at the same time to capture three different horizontal views for the same action sample. The three cameras are set up at the same height yet from three different horizontal angles: $-45^\circ, 0^\circ, +45^\circ$. Each subject is asked to perform each action twice, once towards the left camera, and once towards the right camera. In this way, in each collection setup, we capture two front views, one left side view, one right side view, one left side 45 degrees view, and one right side 45 degrees view. The three cameras are assigned consistent camera numbers in our dataset, where camera 1 always observes the 45 degrees views, while cameras 2 and 3 observe the front and side views.

To further increase the number of camera views, over different collection setups, we also change the vertical heights of the cameras and their distances to the subjects, as reported in TABLE 3. All the camera and setup numbers are provided for each video sample in our dataset.

3.2 Benchmark Evaluations

To have standard evaluations for the methods to be tested on this benchmark, we define precise criteria for two types of action classification evaluation. For each criterion, we report the classification accuracy in percentage.

3.2.1 Cross-Subject Evaluation

For cross-subject evaluation, the 106 subjects are split into training and testing groups. Each group consists of 53 subjects. The IDs of the training subjects in this evaluation are: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38, 45, 46, 47, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, 70, 74, 78, 80, 81, 82, 83, 84, 85, 86, 89, 91, 92, 93, 94, 95, 97, 98, 100, 103. The remaining subjects are reserved for testing.

3.2.2 Cross-Setup Evaluation

For cross-setup evaluation, we pick all the samples with even collection setup IDs for training, and those with odd setup IDs for testing, *i.e.*, 16 setups are used for training, and the other 16 setups are reserved for testing.

TABLE 3: The cameras' height and distance to the subjects in each collection setup.

Setup No.	Height (m)	Distance (m)	Setup No.	Height (m)	Distance (m)
(01)	1.7	3.5	(02)	1.7	2.5
(03)	1.4	2.5	(04)	1.2	3.0
(05)	1.2	3.0	(06)	0.8	3.5
(07)	0.5	4.5	(08)	1.4	3.5
(09)	0.8	2.0	(10)	1.8	3.0
(11)	1.9	3.0	(12)	2.0	3.0
(13)	2.1	3.0	(14)	2.2	3.0
(15)	2.3	3.5	(16)	2.7	3.5
(17)	2.5	3.0	(18)	1.8	3.3
(19)	1.6	3.5	(20)	1.4	4.0
(21)	1.7	3.2	(22)	1.9	3.4
(23)	2.0	3.2	(24)	2.4	3.3
(25)	2.5	3.3	(26)	1.5	2.7
(27)	1.3	3.5	(28)	1.1	2.9
(29)	2.5	2.8	(30)	2.4	2.7
(31)	1.6	3.0	(32)	2.3	3.0

4 APSR Framework for One-Shot 3D Action Recognition

Existing works [87], [88] show that once some categories have been learned, the knowledge gained in this process can be abstracted and used to learn novel classes efficiently, even if only one learning example per new class is given (*i.e.*, via one-shot learning). Since the samples of certain categories may be difficult to collect [32], one-shot visual recognition becomes an important research branch in computer vision.

In this section, we introduce the one-shot 3D action recognition scenario based on our proposed dataset, and show how a large auxiliary set could be used to assist the one-shot recognition for the novel classes. Specifically, an Action-Part Semantic Relevance-aware (APSR) framework is proposed for more reliable one-shot 3D action recognition.

4.1 One-Shot Recognition on NTU RGB+D 120

We define the one-shot 3D action recognition scenario on the proposed NTU RGB+D 120 dataset as follows.

We split the full dataset into two parts: the auxiliary set and the evaluation set. There are no overlaps of classes between these two sets. The auxiliary set contains multiple action classes and samples, and these samples can be used for learning (*e.g.*, learning a feature generation network). The evaluation set consists of the novel action classes for one-shot recognition evaluation, and one sample from each novel class is picked as the exemplar, while the remaining samples are used to test the recognition performance.

4.2 APSR Framework

Previous works [30], [93] have shown that the importance degrees of the features from different body parts are not the same in analyzing different actions. For example, the features extracted from the leg joints are more relevant in recognizing the action "kicking", compared to those from other body parts. Therefore, it is intuitive to identify the body parts that are more relevant to the action performed in a video sequence, and correspondingly emphasize their features to achieve reliable recognition performance.

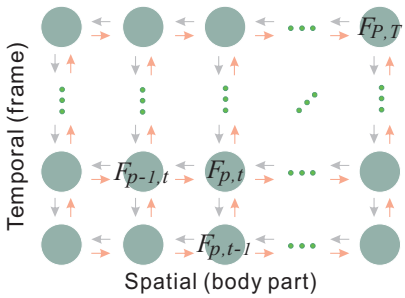

Fig. 2: Illustration of the body part feature generation network.

However, in our one-shot recognition scenario, identifying the relevant body parts of the novel actions is difficult, as learning to generalize beyond the single specific exemplar is often very challenging [88].

In this paper, a simple yet effective APSR framework, which can be used to generalize to the novel action categories, is introduced for one-shot recognition. The APSR framework emphasizes the relevant body parts for each new class of actions by considering the semantic relevance between the action and each body part based on their descriptions. Specifically, we design a network to generate the features of each body part, and then perform weighted pooling over these features with the guidance of relevance scores in the word embedding space. Finally, the pooling result is used to perform one-shot recognition.

Feature Generation Network. The feature generation network is adopted from the 2D Spatio-Temporal LSTM (ST-LSTM) [94] designed for 3D activity analysis, which models the context and dependency information in both temporal dimension (over different frames) and spatial dimension (over different body parts). Readers are referred to [94] for more details about the mechanism of ST-LSTM.

The original ST-LSTM [94] models the context information via a single pass. To produce a more discriminative set of features for each part, we introduce bidirectional context passing (similar to bidirectional LSTM [95]) for our feature generation network, as illustrated in Fig. 2.

The input to our feature generation network is an action sample (here we input its skeleton data for efficiency), and the outputs are the features of all body parts at each frame. Concretely, at the unit (p, t) of this network, the input is the 3D coordinates of the skeletal joint of the body part (p) in the frame (t), and the output is the feature ($F_{p,t}$) representing this body part at this frame. By incorporating the spatio-temporal context information into each part, the obtained feature set $\mathbf{F} = \{F_{p,t} \mid p \in \{1, ..., P\}, t \in \{1, ..., T\}\}$ is powerful for representing each body part (p) at each frame (t) in the performing of the action instance, where P is the number of body parts, and T is the number of frames used for each video sample.

Semantic Relevance. The proposed method is inspired by the recent works on word embedding in natural language processing [27], [28], [96]. In these works, each word is mapped into a continuous embedding space, and two semantically relevant words will have large cosine similarity in this space [96], [97], [98]. By pre-training on a massive natural language corpus, these models demonstrate their

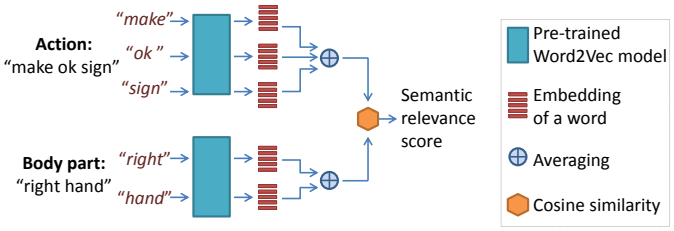

Fig. 3: Illustration of estimating the semantic relevance score between the novel action's text description (name) and the body part's text description (name). Here we take the action "make ok sign" with the body part "right hand" as an example. Each word in a text description is fed to the pre-trained Word2Vec model to produce its embedding (a 300-dimensional vector), and the representation of a text description is obtained by averaging the embedding of all the words in it. Finally, the semantic relevance is estimated by calculating the cosine similarity between the two representations.

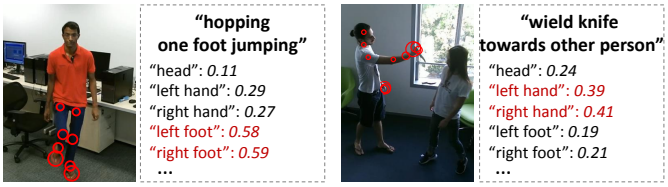

Fig. 4: Examples of semantic relevance scores between action's name and each body part's name. In the 1st and 3rd columns, the body parts (joints) with larger scores are labeled with red circles. In the 2nd and 4th columns, we show the scores of several body parts. These scores are obtained by using pre-trained word embedding model [28]. Semantically, "right foot" and "left foot" are very relevant to "hopping, one foot jumping", while "right hand" and "left hand" are relevant to "wield knife towards other person".

superior ability in preserving the semantic relations among different words, and thus have been successfully transferred to different tasks, such as document classification [99], image classification [88], and image/video caption generation [100].

This motivates us to utilize the prior knowledge about relevant body parts for recognizing new action classes, and the semantic relevance (cosine similarity) between the novel action's name and each body part's name in the embedding space can be used as prior information.

Specifically, the powerful Word2Vec model [28] that is pre-trained on a large corpus is used in our method. When given a new action, we estimate the relevance score (cosine similarity) between this action and each body part based on their text descriptions (*e.g.*, "make ok sign" versus "right hand"), by using the pre-trained Word2Vec model. For a description consisting of multiple words, its representation is obtained by averaging the embedding of all the words of it. If the estimated relevance score is negative, we reset it to zero. The method of estimating the semantic relevance score is illustrated in Fig. 3.

As shown in Fig. 4, the relevant body parts of the novel actions can be reliably indicated by using this method.

After we obtain the semantic relevance score ($r_{c,p}$) between the action class (c) and each body part (p), we normalize the score as: $s_{c,p} = r_{c,p} / \sum_{u=1}^{P} r_{c,u}$ to ensure that the total score of all body parts for this action class is 1 after normalization. Therefore, the action-part relevance score set ($\mathbf{S}_c$) of the action class (c) is obtained as: $\mathbf{S}_c = \{s_{c,p} \mid p \in \{1, ..., P\}\}$, which will be used as prior information for the feature generation network training and one-shot recognition evaluation.

Training. In our framework, we train the feature generation network on the auxiliary set that does not contain samples from the novel action categories. To train this network, at each unit (p, t) (see Fig. 2), we feed the produced body part feature $F_{p,t}$ to the softmax classifier for action classification, similar to [94]. This indicates that at each unit, a prediction of the action class is produced based on the body part feature ($F_{p,t}$). We train the feature generation network with the classifiers in an end-to-end manner by using the following loss function:

$$L = \sum_{p=1}^{P} \sum_{t=1}^{T} s_{c,p} l(c, \hat{c}_{p,t}) \tag{1}$$

where $l(c, \hat{c}_{p,t})$ is the negative log-likelihood loss measuring the difference between the true class label, c, of the sample and the prediction result, $\hat{c}_{p,t}$, at the unit (p, t). The semantic relevance score $s_{c,p}$ is used here as the weight of the classification loss at the units $(p, \cdot)$ that correspond to the body part p, *i.e.*, more relevant parts are given larger loss weights. This drives the network to learn more discriminative features on the more relevant body parts of each action class.

Evaluation. After training the network on the auxiliary set, we feed each sample from the evaluation set to the feature generation network to produce a feature set ($\mathbf{F}$) for this sample, which represents the features of different body parts at each frame. Note that during evaluation, we remove the classifiers and only use the produced features, since the classes for evaluation are not contained in the auxiliary training set.

We perform weighted pooling over the obtained features ($\mathbf{F}$) of each sample to generate an aggregated representation. Let $f(a, \mathbf{b})$ denote the aggregated representation of the sample a when using $\mathbf{b}$ as the weights for feature pooling. Then the aggregated representation of an exemplar (Ω) of a novel class is calculated as:

$$f(\Omega, \mathbf{S}_{c_\Omega}) = \sum_{p=1}^{P} \sum_{t=1}^{T} s_{c_\Omega, p} F_{p,t} \tag{2}$$

where c_Ω denotes the class label of the exemplar Ω, and $\mathbf{S}_{c_\Omega}$ is the action-part relevance score set of the action class c_Ω. Here the semantic relevance scores are used as weights to guide the aggregated representation of the sample Ω and to emphasize the features from more relevant body parts.

To test a sample (i), we calculate the distance between this sample (i) and an exemplar (Ω) as:

$$D(i, \Omega) = D_{\cos}\Big(f(i, \mathbf{S}_{c_\Omega}), f(\Omega, \mathbf{S}_{c_\Omega})\Big) \tag{3}$$

where $D_{\cos}$ is the cosine distance between the two representations (vectors). Note that when calculating the distance of i to the exemplar Ω, we generate its aggregated representation as $f(i, \mathbf{S}_{c_\Omega})$, which indicates the weighted pooling on the feature set of i is based on the relevance score set of the exemplar's class c_Ω.

For each testing sample, we calculate its distances to all the exemplars from all action categories by using Eq. (3), and then perform classification using the nearest neighbour classifier, as in [101].

5 Experiments

In this section, some state-of-the-art methods that were designed for 3D action recognition are evaluated based on the suggested cross-subject and cross-setup criteria. Then the action recognition performances achieved by adopting different data modalities are compared. The performance of the proposed APSR framework for one-shot 3D action recognition is also evaluated.

5.1 Experimental Evaluations of 3D Action Recognition

5.1.1 Evaluation of state-of-the-art methods

Twelve state-of-the-art 3D action recognition methods are evaluated on our dataset, namely, the Part-Aware LSTM [47], the Soft RNN [102], the Dynamic Skeleton [26], the Spatio-Temporal LSTM [48], the Internal Feature Fusion [94], the GCA-LSTM [30], the Multi-Task Learning Network [49], the FSNet [103], [108], the Skeleton Visualization [104], the Two-Stream Attention LSTM [105], the Multi-Task CNN with RotClips [106], and the Body Pose Evolution Map method [107]. Using the cross-subject and cross-setup evaluation criteria, the results of these methods are reported in TABLE 4.

Among these approaches, the method in [26] uses the Fourier temporal pyramid [56] and the hand-crafted features for action classification. The methods in [30], [47], [48], [94], [102], [105] all use RNN/LSTM for 3D action recognition, and the approaches in [49], [103], [104], [106], [107] all use convolutional networks for spatial and temporal modeling. Specifically, the evaluations of [104], [107] are performed by using the efficient MobileNet [109] as the base model, and skeleton data as input. All the evaluated implementations are from the original authors of the corresponding papers.

5.1.2 Evaluation of using different data modalities

Since multiple data modalities are provided in our dataset, we also evaluate the performance of using different data modalities (*e.g.*, RGB, depth, and skeleton data) as input for action recognition, and report the results in TABLE 5.

In this table, the accuracy on RGB video is achieved by learning visual and motion (optical flow) features at each frame of the RGB video by training the VGG model [110], and performing classification by averaging the classification results of all frames, similarly to [111]. The accuracy on depth video is obtained by using the similar method as the RGB video, and training the VGG model based on the depth data. The accuracy on 3D skeleton data is achieved by using the method in [48].

We observe that when using the RGB or depth video as input, the performance of the cross-setup evaluation is weaker than that of the cross-subject one. The performance disparity can be justified as: (1) In the cross-setup evaluation scenario, the heights and distances of the cameras are not the same over different setups. This indicates camera viewpoints are different, and thus the appearance of the actions can be significantly different. (2) The background is also changing across different setups. However, when using the RGB or depth frames as input, the methods are more prone to learn from the appearance or view-dependent motion patterns.

By using the 3D skeleton data as input, the method performs better in the cross-setup evaluation. This is partially because the method using 3D skeleton data is stronger to generalize among different views, since the 3D skeletal representation is view-invariant in essence. However, it is still prone to errors of the body tracker.

We also evaluate the performance of fusing multiple data modalities for action recognition. The results in TABLE 5 show that compared to the method of using a single modality, fusing multiple modalities bring a performance improvement, since they contain complementary and discriminative appearance and geometrical information, which is useful for activity analysis.

5.1.3 Evaluation of using different sizes of training set

In the aforementioned experiments, the evaluated deep learning models, such as Part-Aware LSTM [47] and FSNet [103], are trained with the defined large training set from our NTU RGB+D 120 dataset. Here we call this training set as "full training set".

To evaluate the effect of the training set size on the recognition performance, we use different ratios of training samples from the full training set for network training. We then evaluate the action recognition performance on the same testing set based on the cross-setup evaluation criterion.

We take six methods [30], [47], [48], [94], [103], [105] as examples, and show their results achieved by using different sizes of training set in Fig. 5. The results show that when more samples are used for network training, the action recognition accuracies of all these methods increase obviously. For example, when using 20% of the samples from the full training set for training, the accuracy of FSNet is 40.6%, while the accuracy reaches 62.4% when the full training set is used for network training.

We also evaluate the performance of using different sizes of training data for action recognition with different data modalities, and show the results in Fig. 6. The results in this figure also show the benefit of using more data for network training to achieve better action recognition performance.

5.1.4 Detailed analysis according to data modalities

Here we analyze the results obtained by using different data modalities in detail. The skeleton data modality is evaluated using the Spatio-Temporal LSTM [48]. The RGB and depth data modalities are both evaluated using the two-stream framework [108]. Modality fusion is performed by fusing the results of the three modalities.

TABLE 4: The results of different methods, which are designed for 3D human activity analysis, using the cross-subject and cross-setup evaluation criteria on the NTU RGB+D 120 dataset.

Method		Cross-Subject Accuracy	Cross-Setup Accuracy
Part-Aware LSTM	[47]	25.5%	26.3%
Soft RNN	[102]	36.3%	44.9%
Dynamic Skeleton	[26]	50.8%	54.7%
Spatio-Temporal LSTM	[48]	55.7%	57.9%
Internal Feature Fusion	[94]	58.2%	60.9%
GCA-LSTM	[30]	58.3%	59.2%
Multi-Task Learning Network	[49]	58.4%	57.9%
FSNet	[103]	59.9%	62.4%
Skeleton Visualization (Single Stream)	[104]	60.3%	63.2%
Two-Stream Attention LSTM	[105]	61.2%	63.3%
Multi-Task CNN with RotClips	[106]	62.2%	61.8%
Body Pose Evolution Map	[107]	64.6%	66.9%

TABLE 5: Evaluation of using different data modalities (RGB, depth, and 3D skeleton data) for action recognition on the NTU RGB+D 120 dataset.

Data Modality	Cross-Subject Accuracy	Cross-Setup Accuracy
RGB Video	58.5%	54.8%
Depth Video	48.7%	40.1%
3D Skeleton Sequence	55.7%	57.9%
RGB Video + Depth Video	61.9%	59.2%
RGB Video + 3D Skeleton Sequence	61.2%	63.1%
Depth Video + 3D Skeleton Sequence	59.2%	61.2%
RGB Video + Depth Video + 3D Skeleton Sequence	64.0%	66.1%

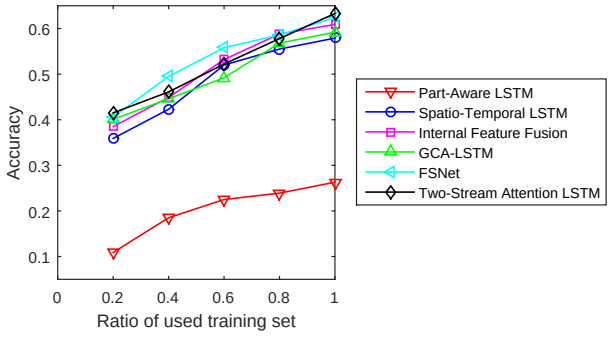

Fig. 5: Evaluation of using different sizes of training set for action recognition with different methods. In this figure, ratio 1.0 means the full training set is used for network training.

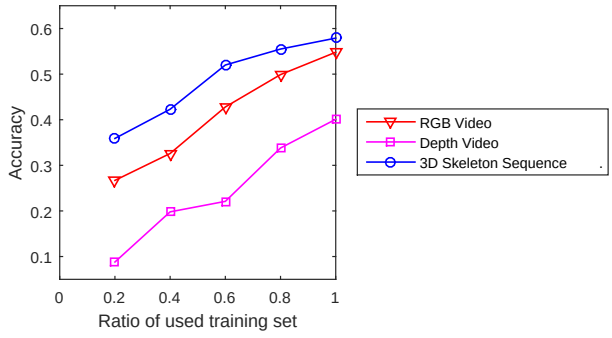

Fig. 6: Evaluation of using different sizes of training set for action recognition with different data modalities.

We first plot the confusion matrices of different data modalities. Specifically, we show the confusion matrix of the RGB modality as an example in Fig. 7.

We then perform action-wise analysis for different data modalities. Considering the large number of action categories, for each data modality, we analyze the action classes that have high recognition accuracies (top 10 accurate classes), and the actions that are easily misclassified to other classes (top 10 confused action pairs). We show the results in TABLE 6.

Based on the results in TABLE 6, we find that the actions that have significant motions and discriminative posture

TABLE 6: Action recognition results of different data modalities on the NTU RGB+D 120 dataset.

Data Modality	Top 10 accurate actions	Top 10 confused (misclassified) action pairs
RGB	1. walk apart from each other 2. walk towards each other 3. hopping 4. carry things with other person 5. arm swings 6. staggering 7. pick up things 8. put on jacket 9. hugging other person 10. move heavy objects	1. take off a shoe→put on a shoe 2. kick backward→side kick 3. rub two hands together→clapping 4. reading→writing 5. clapping→rub two hands together 6. vomiting condition→bow 7. ball up paper→fold paper 8. open a box→fold paper 9. both hands up→cheer up 10. yawn→blow nose
Depth	1. carry things with other person 2. walk apart from each other 3. move heavy objects 4. hugging other person 5. kick backward 6. walk towards each other 7. hopping 8. arm swings 9. staggering 10. open a box	1. take off a shoe→put on a shoe 2. reading→writing 3. playing with tablet→writing 4. bow→vomiting condition 5. both hands up→stretch oneself 6. put on jacket→put on bag/backpack 7. vomiting condition→bow 8. cheer up→both hands up 9. rub two hands together→clapping 10. take off bag/backpack→take off jacket
Skeleton	1. walk apart from each other 2. standing up 3. walk towards each other 4. hugging other person 5. arm swings 6. squat down 7. sitting down 8. pushing other person 9. arm circles 10. kick backward	1. put on a shoe→take off a shoe 2. hit other person with object→wield knife towards other person 3. make ok sign→make victory sign 4. thumb up→make victory sign 5. put on jacket→put on bag/backpack 6. touch other person's pocket (steal)→grab other person's stuff 7. make victory sign→make ok sign 8. play magic cube→counting money 9. take a photo of other person→shoot at other person with a gun 10. handshaking→giving something to other person
RGB + Depth + Skeleton	1. walk apart from each other 2. carry things with other person 3. hugging other person 4. walk towards each other 5. standing up 6. hopping 7. squat down 8. move heavy objects 9. arm swings 10. put on jacket	1. take off a shoe→put on a shoe 2. vomiting condition→bow 3. both hands up→stretch oneself 4. clapping→rub two hands together 5. yawn→blow nose 6. put on a shoe→take off a shoe 7. hush (say quite)→blow nose 8. rub two hands together→clapping 9. reading→writing 10. stretch oneself→both hands up

(1) Top 10 accurate actions denote the actions that have the top 10 recognition accuracies.
(2) Top 10 confused (misclassified) action pairs denote the action pairs that have the top 10 confusion rates (misclassification percentages).
(3) $A \rightarrow B$ denotes a confused action pair, where some samples of class A are misclassified to class B.

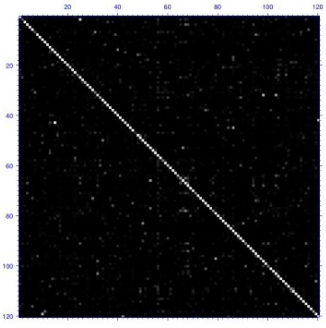

Fig. 7: Confusion matrix of the RGB data modality.

patterns could be more accurately recognized. For example, the actions "walk apart from each other" and "walk towards each other", which have very discriminative and significant motions, are both in the top 10 accurate actions when using any of the three data modalities as input.

We also observe that when using skeleton data as input, the actions that involve interactions with objects may be easily misclassified. For example, "play magic cube" is often confused with "counting money", and "handshaking" can also be misclassified to "giving something to other person", as shown in TABLE 6. This is possibly because the actions in each pair have similar human motion patterns, and the perception of the existences of the objects and their appearances is important for accurately recognizing these actions. However, the appearance information of the objects is ignored by the recognition model when using skeleton information only. In contrast, when we perform action recognition based on the RGB or depth data that captures the object information, many object-related actions could be accurately recognized. For example, "carry things with other person" and "move heavy objects" are both in the top 10 accurate actions of the RGB and depth data modalities.

Although RGB and depth data modalities both have good ability in representing the object-related actions, the actions with object involved may be misclassified when the same (or similar) objects are shared by different actions, where the objects may even mislead the classification. For example, when using RGB data as input, "ball up paper" tend to be misclassified to "fold paper", and "reading" and "writing" can also be confused, as shown in TABLE 6. An interesting observation is that for RGB data modality, many

samples of "open a box" are misclassified to "fold paper", while "open a box" is in the top 10 accurate actions for the depth data modality. This performance disparity could be explained as: in the RGB image, the appearances of the box and the paper can be similar, but the depth data can well represent the 3D shape information of the objects, thus depth data is more powerful in distinguishing the box from paper. As a result, the action "open a box" can be more accurately recognized by using depth data than by using RGB data. We also observe that "put on jacket", which is classified well with RGB data, is easily confused with "put on bag/backpack" with depth data. This may be because the jacket and the backpack can be more easily distinguished from their color and texture information, than from their 3D shape information.

As shown in TABLE 6, "kick backward" is in the top 10 accurate actions of both skeleton and depth data modalities, while "kick backward" is easily confused to "side kick" when using RGB data. A possible reason is that both the depth data modality and the skeleton data modality provide the 3D structure information that implies the 3D direction information of the body part motions. However, such 3D information is not provided when using the RGB data.

We also observe that many actions that contain fine-grained hand gestures and finger motions, such as "make ok sign", "make victory sign", and "thumb up", are easily misclassified when using the skeleton data only, as depicted in TABLE 6. The performance limitation of the skeleton data in handing these actions is possibly due to that only three joints are provided for each hand in the skeleton data, and besides, the skeleton data provided by Kinect's tracker algorithm is not perfect and can be noisy sometimes. This indicates that the skeleton data has difficulties in representing the very detailed hand and finger motions. Therefore, the actions with fine-grained hand/finger motions could be easily misclassified when using the skeleton data only.

In TABLE 6, we also find that there are several *tough* action pairs that are easily confused for all the data modalities, such as "take off a shoe" and "put on a shoe". A possible explantation is that the human motion and object appearance information in these actions are both very similar, and thus they are difficult to be accurately distinguished.

In summary, the 2D appearance information (*e.g.*, color and texture) and the 3D shape and structure information provided by different data modalities can all affect the recognition performance of different types of actions.

In TABLE 6, beside presenting the top 10 accurate actions and the top 10 confused action pairs for different data modalities, we also show the results of fusing the three modalities. We observe that when fusing these modalities which provide complementary appearance and geometric information, the recognition performance is improved. For example, when using RGB, depth, or skeleton data only, the recognition accuracies of "walk apart from each other" are 94%, 93%, and 92%, respectively. When fusing the three modalities, the accuracy reaches 99%. We also find that the confusion rates of the *tough* action pairs can also drop when fusing the three data modalities. For the very *tough* action pair: "take off a shoe" and "put on a shoe", the confusion rates for different single modalities are 52%, 65%, and 39%, respectively, and the confusion rate decreases to 32% with modality fusion.

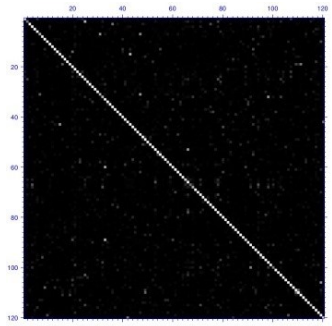

Fig. 8: Confusion matrix of Internal Feature Fusion [94].

5.1.5 Detailed analysis according to methods

We also analyze the experimental results of different 3D action recognition methods on our dataset in detail. We take five state-of-the-art methods as examples for analysis, namely Internal Feature Fusion [94], GCA-LSTM [30], Multi-Task Learning Network [49], FSNet [103], and Multi-Task CNN with RotClips [106].

We first plot the confusion matrices of these methods. The confusion matrix of the method, Internal Feature Fusion, is shown in Fig. 8 as an example.

We then perform action-wise analysis for these methods. Specifically, we perform detailed analysis for the top 10 accurate actions and the top 10 confused action pairs of each method (see TABLE 7), considering the large number of action classes.

Among these methods, the method Internal Feature Fusion performs action recognition by fusing the 3D skeleton-based geometric features and the RGB-based appearance features, and the other four approaches all use 3D skeleton data as input for 3D action recognition.

In TABLE 7, we observe that the top 10 confused action pairs of GCA-LSTM, Multi-Task Learning Network, FSNet, and Multi-Task CNN with RotClips all contain many object-related actions (such as "put on jacket" and "play magic cube") and fine-grained hand/finger motion-based actions (such as "make victory sign" and "thumb up"). This is possibly owing to that all of these four approaches perform action recognition based on the 3D skeleton data that is not able to represent the object information and the fine-grained finger motions well. Therefore, these approaches have difficulties in dealing with the object-related and fine-grained hand/finger motion-based activities. In contrast, we observe that there are many object-related actions (such as "put on jacket", "bounce ball", and "carry things with other person") in the top 10 accurate actions of the method Internal Feature Fusion, which uses both 3D skeleton data and RGB data as input for action recognition.

In this table, there are also some actions (*e.g.*, "take off a shoe" and "put on a shoe") that are very similar in motions and appearances, and are difficult to be distinguished well by all the five methods.

We also observe that the actions "walk apart from each other" and "walk towards each other" that have significant motions and discriminative posture patterns are in the top 10 accurate actions of all the methods.

TABLE 7: Action recognition results of different methods on the NTU RGB+D 120 dataset.

Method	Top 10 accurate actions	Top 10 confused (misclassified) action pairs
Internal Feature Fusion [94]	1. walk apart from each other 2. walk towards each other 3. carry things with other person 4. hugging other person 5. standing up 6. kick backward 7. take off jacket 8. arm swings 9. put on jacket 10. bounce ball	1. put on a shoe→take off a shoe 2. blow nose→hush (say quite) 3. clapping→rub two hands together 4. take off a shoe→put on a shoe 5. rub two hands together→clapping 6. stretch oneself→both hands up 7. yawn→blow nose 8. bow→vomiting condition 9. grab other person's stuff→touch other person's pocket (steal) 10. vomiting condition→sneeze/cough
GCA-LSTM [30]	1. walk apart from each other 2. standing up 3. walk towards each other 4. hugging other person 5. high five 6. handshaking 7. arm swings 8. sitting down 9. arm circles 10. squat down	1. put on a shoe→take off a shoe 2. play magic cube→counting money 3. make victory sign→make ok sign 4. hit other person with object→wield knife towards other person 5. take something out of a bag/backpack→put something into a bag/backpack 6. kick backward→hopping 7. staggering→kick backward 8. put on jacket→put on bag/backpack 9. giving something to other person→exchange things with other person 10. grab other person's stuff→touch other person's pocket (steal)
Multi-Task Learning Network [49]	1. cross toe touch 2. walk apart from each other 3. walk towards each other 4. arm swings 5. kick backward 6. cheers and drink 7. squat down 8. arm circles 9. running on the spot 10. cheer up	1. giving something to other person→exchange things with other person 2. point finger at other person→shoot at other person with a gun 3. slapping other person→hit other person with object 4. rub two hands together→apply cream on hand back 5. cutting paper using scissors→staple book 6. yawn→hush (say quite) 7. take off glasses→take off headphone 8. make ok sign→thumb up 9. take off a shoe→put on a shoe 10. make ok sign→make victory sign
FSNet [103]	1. standing up 2. arm circles 3. walk apart from each other 4. kick backward 5. arm swings 6. cross toe touch 7. grab other person's stuff 8. cheer up 9. running on the spot 10. walk towards each other	1. put on a shoe→take off a shoe 2. make victory sign→make ok sign 3. put on jacket→put on bag/backpack 4. hit other person with object→wield knife towards other person 5. counting money→play magic cube 6. rub two hands together→clapping 7. pushing other person→slapping other person 8. pat on back of other person→hit other person with object 9. kicking other person→step on foot of other person 10. hit other person with body→support other person with hand
Multi-Task CNN with RotClips [106]	1. cross toe touch 2. walk apart from each other 3. running on the spot 4. arm circles 5. walk towards each other 6. squat down 7. arm swings 8. kick backward 9. cheer up 10. playing rock-paper-scissors	1. slapping other person→hit other person with object 2. point finger at other person→shoot at other person with a gun 3. make victory sign→make ok sign 4. tear up paper→open a box 5. staple book→cutting paper using scissors 6. yawn→hush (say quite) 7. giving something to other person→exchange things with other person 8. playing with tablet→play magic cube 9. take off glasses→take off headphone 10. take off a shoe→put on a shoe

In TABLE 7, we observe the top 10 confused action pairs of FSNet contain some two-person mutual action pairs without object involved, such as "hit other person with body" with "support other person with hand", and "kicking other person" with "step on foot of other person", while these actions can be classified relatively reliably by the other approaches. A possible explanation of the performance limitation of FSNet in handling these mutual actions is that in FSNet, the features of the two persons are extracted separately, and these features are then simply averaged for action recognition. This indicates the interaction patterns between the two persons are not well represented, and thus the mutual actions may be easily misclassified by FSNet.

We also observe that the action "kick backward" is easily confused with "hopping" by the method GCA-LSTM, while "kick backward" is in the top 10 accurate actions of all the other four methods. A possible reason is that GCA-LSTM normalizes the skeleton data for the single-person actions by rotating the skeleton to the frontal view and translating the body center to the origin in each frame at the pre-processing stage. After such normalization, "hopping" could be similar to "kick backward", since the vertical movements of the body center in "hopping" are ignored by the method GCA-LSTM.

In TABLE 7, "grab other person's stuff" is in the top 10 accurate actions of the method FSNet, while "grab other person's stuff" can be easily confused with "touch other person's pocket (steal)" by Internal Feature Fusion and GCA-LSTM. A possible explanation of the performance disparity is that FSNet may be able to learn the motion speed information better than the other two methods. The human postures of these two actions are quite similar and their main difference is the motion speed, *i.e.*, the motions in "touch other person's pocket (steal)" are very slow, while the motions in "grab other person's stuff" are much faster. Both Internal Feature Fusion and GCA-LSTM use recurrent

TABLE 8: The results of one-shot 3D action recognition on the NTU RGB+D 120 dataset.

Method	Evaluation Accuracy
Average Pooling [94]	42.9%
Fully Connected [30]	42.1%
Attention Network [30]	41.0%
APSR	**45.3**%

models to learn the temporal dynamics of the actions based on the sampled 20 frames from each action sequence. This implies the speed information of the actions may be ignored by them. In contrast, FSNet uses a temporal convolutional model to learn the temporal context information over all the frames of each action sample, and thus is able to better learn the motion speed information, which is an important cue to distinguish "grab other person's stuff" from "touch other person's pocket (steal)".

5.2 Experimental Evaluations of One-Shot Recognition

We evaluate the one-shot recognition performance on our dataset. In our experiments, the auxiliary set for feature generation network training contains 100 action classes, and the evaluation set for one-shot recognition evaluation contains the remaining 20 classes.

We compare the following methods for one-shot 3D action recognition:

(1) Average Pooling. This method is similar to the approach in [94]. To adapt [94] for one-shot recognition, we use ST-LSTM [94] as the feature generation network to learn the features of the body parts at each frame, and during evaluation, the features of all body parts at all frames are aggregated with average pooling. The distance between the testing sample and each exemplar is calculated using the average pooling representation of each video.

(2) Fully Connected. In this method, the feature generation network is constructed by adding a fully connected layer above the ST-LSTM model. Concretely, the outputs from all spatio-temporal units of the ST-LSTM are concatenated and fed to a fully connected layer to generate a global representation for the input video [30]. During training, the ST-LSTM and the fully connected layer is trained in an end-to-end fashion. During evaluation, the distance is calculated using the global representation of each video.

(3) Attention Network. This method is similar to the above "Fully Connected" method, except that an attention mechanism [30] is added to the feature generation network, *i.e.*, the attention scores of different joints are automatically learned in this method.

(4) APSR. This is the proposed Action-Part Semantic Relevance-aware (APSR) framework, which assigns different scores to different joints (*i.e.*, weighted pooling), by considering the sematic relevance between the novel action's name and body part's name for one-shot recognition.

The comparison results of these approaches are shown in TABLE 8. The proposed APSR framework achieves the best results, which indicates the generalization capability of the proposed method on the novel action categories. We also observe that the performance of "Attention Network" [30], which learns to assign weights to different joints with the attention mechanism, is even weaker than that of "Average Pooling" [94]. A possible explanation is that the attention ability is trained on the auxiliary set that does not contain the novel actions, thus when handling the novel actions, its performance is even worse than directly performing average pooling. This further demonstrates the superiority of the introduced APSR framework.

TABLE 9: The results of using different sizes of auxiliary training set to learn the feature generation network, for one-shot recognition on the novel action classes.

Auxiliary Training Set		Evaluation Accuracy
#Training Samples	#Training Classes	
19,000	20	29.1%
38,000	40	34.8%
57,000	60	39.2%
76,000	80	42.8%
95,000	100	45.3%

In the aforementioned experiments, the feature generation network of our APSR framework is trained on a large auxiliary set containing about 100 thousand videos. We also try downsizing the auxiliary training set, and evaluate the one-shot 3D action recognition performance on the same evaluation set. The results in TABLE 9 show that the one-shot recognition accuracy drops and the generalization capability to novel classes is weakened, when using fewer classes and samples for learning the feature generation network. This also implies the demand for a large dataset, and it is in line with our motivation for proposing the NTU RGB+D 120 dataset.

5.3 Discussions

The introduction of this very large-scale and challenging dataset with high variability in different aspects (*e.g.*, subjects, environments, camera views, and action categories) will facilitate the users to apply, adapt, develop, and evaluate various learning-based techniques for the future research on human activity analysis. Below we discuss some of the potential research problems and techniques that could be investigated by taking advantage of our dataset:

(1) Activity analysis with different data modalities. Four different data modalities are provided by our dataset, namely, depth videos, 3D skeleton data, RGB videos, and infrared sequences. Different modalities have different structures of the data, and have different application advantages. Therefore, users can utilize our dataset to investigate the algorithms for depth-based, skeleton-based, RGB-based, or infrared-based action recognition.

(2) Heterogeneous feature fusion analysis. The provided data modalities contain complementary appearance and 3D geometrical information for human activity analysis. Thus users can take advantage of our dataset to identify the strengths of respective modalities, and further investigate

various fusion techniques for the heterogeneous features [43] extracted from different data modalities.

(3) Deep network pre-training. Most of the existing datasets for RGB+D action recognition are relative small, thus the deep models evaluated on them often suffer from over-fitting issues. Since the proposed dataset has a large number of samples with diversity in various factors, it can also be employed for network pre-training. By initializing the network parameters on our proposed large-scale dataset, the deep models are expected to be able to generalize better on other relative small datasets for 3D activity analysis, as analyzed in [69].

(4) Cross-subject activity analysis. In the proposed dataset, the 106 human subjects are in a wide range of age and height distribution, and are from different cultural backgrounds. These factors bring realistic variation to the quality of actions, and make it possible to have more sensible cross-subject evaluations for the 3D activity analysis methods. This also encourages the community to develop action recognition algorithms that are robust for different subjects.

(5) Cross-environment activity analysis. Our dataset is collected under different environmental conditions that use 96 different backgrounds with significant illumination variation. This enables the users to perform cross-environment activity analysis. The different collection environments can also facilitate the analysis of the algorithms' robustness against the variation in backgrounds and illuminations.

(6) Cross-view activity analysis. The proposed dataset is collected with 155 camera views, which facilitates the cross-view activity analysis and encourages the users to develop action recognition approaches that are robust against view variation for the practical applications.

(7) Cross-modal transfer learning. Learning representations from a large labeled modality for transfer learning for the smaller-scale new modalities has attracted a lot of research attention and has been applied to different tasks recently [112]. The proposed large dataset that provides different data modalities could be utilized for the research on cross-modal transfer learning.

(8) Mutual activity analysis. Human-human interaction analysis is also an important branch of human activity analysis. The proposed dataset contains 25 thousand two-person mutual action videos that correspond to 26 different two-person interaction classes. This facilitates the users to investigate and develop various approaches for handling the task of mutual action recognition.

(9) Real-time skeleton-based early action recognition. 3D skeleton data has shown its advantages in real-time early action recognition due to its succinct and high level representation, as analyzed in [64]. This indicates our large dataset can also be used for the research on real-time early action recognition.

(10) One-shot 3D activity analysis. Our large-scale dataset can also be used to learn a discriminative representation model for one-shot 3D activity analysis for novel action classes.

6 Conclusion

A large-scale RGB+D action recognition dataset is introduced in this paper. Our dataset includes 114,480 video samples collected from 120 action classes in highly variant camera settings. Compared to the current datasets for this task, our dataset is larger in orders and contains much more variety in different aspects. The large scale of the collected data facilitates us to apply data-driven learning methods to this problem and achieve promising performance. We also propose an APSR framework for one-shot 3D action recognition. The provided experimental results show the availability of large-scale data enables the data-driven learning frameworks to achieve promising results.

References

[1] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, 2013.

[2] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3d object recognition in cluttered scenes with local surface features: a survey," *TPAMI*, 2014.

[3] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *IJCV*, 2015.

[4] J. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *PR Letters*, 2014.

[5] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011.

[6] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv*, 2012.

[7] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

[8] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv*, 2017.

[9] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015.

[10] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," in *THMS*, 2015.

[11] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPR Workshops*, 2010.

[12] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images," in *AAAI Workshops*, 2011.

[13] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *ICCV Workshops*, 2011.

[14] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.

[15] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *CVPR Workshops*, 2012.

[16] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *ECCV Workshops*, 2012.

[17] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *IJRR*, 2013.

[18] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, 2013.

[19] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *ICCV*, 2013.

[20] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *CVPR*, 2014.

[21] 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.

[22] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in *ACM MM*, 2014.

[23] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*, 2015.

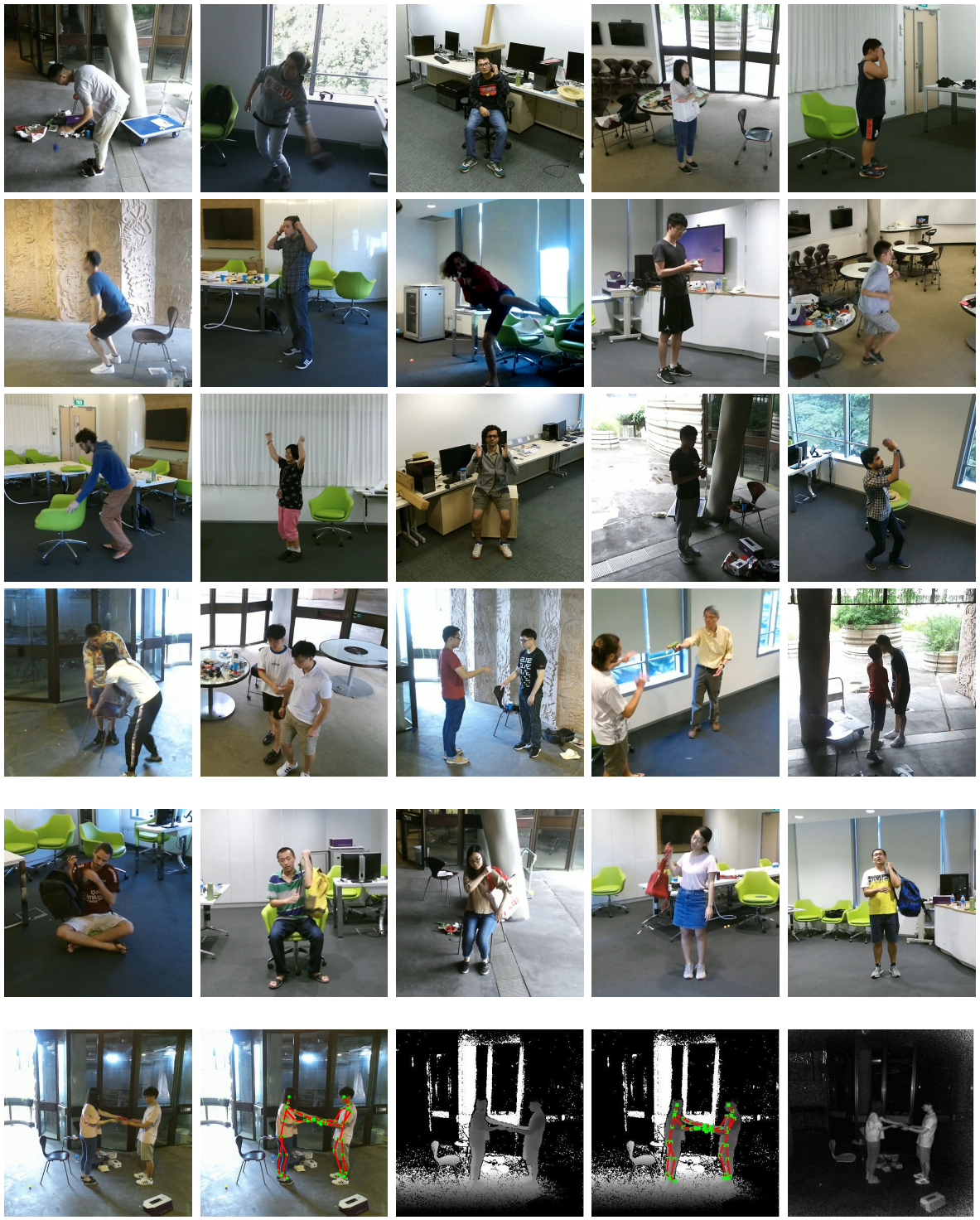

Fig. 9: Sample frames of the NTU RGB+D 120 dataset. The first four rows show the variety in human subjects, camera views, and environmental conditions. The fifth row depicts the intra-class variation of the performances. The last row illustrates the RGB, RGB+joints, depth, depth+joints, and IR modalities of a sample frame.

[24] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *TPAMI*, 2016.

[25] N. Xu, A. Liu, W. Nie, Y. Wong, F. Li, and Y. Su, "Multi-modal & multi-view & interactive benchmark dataset for human action recognition," in *ACM MM*, 2015.

[26] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2186–2200, 2017.

[27] H. Zamani and W. B. Croft, "Relevance-based word embedding," in *SIGIR*, 2017.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[29] 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*, 2013.

[30] 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.

[31] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," *PR*, 2016.

[32] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *CVIU*, 2018.

[33] L. L. Presti and M. La Cascia, "3d skeleton-based human action classification: A survey," *PR*, 2016.

[34] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *PR Letters*, 2013.

[35] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: A review," *CVIU*, 2017.

[36] R. Lun and W. Zhao, "A survey of applications and human motion recognition with microsoft kinect," *IJPRAI*, 2015.

[37] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, 2012.

[38] A. Shahroudy, T. T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *TPAMI*, 2016.

[39] A. B. Tanfous, H. Drira, and B. B. Amor, "Coding kendall's shape trajectories for 3d action recognition," in *CVPR*, 2018.

[40] C. Lu, J. Jia, and C.-K. Tang, "Range-sample depth feature for action recognition," in *CVPR*, 2014.

[41] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *CVPR*, 2015.

[42] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *ICCV*, 2013.

[43] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in rgb+d videos," *TPAMI*, 2018.

[44] Y. Kong and Y. Fu, "Bilinear heterogeneous information machine for rgb-d action recognition," in *CVPR*, 2015.

[45] V. Bloom, D. Makris, and V. Argyriou, "G3d: A gaming action dataset and real time action recognition evaluation framework," in *CVPR Workshops*, 2012.

[46] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv*, 2017.

[47] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016.

[48] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *ECCV*, 2016.

[49] Q. Ke, M. Bennamoun, S. An., F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *CVPR*, 2017.

[50] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *CVPR*, 2014.

[51] J. Shotton, A. W. 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.

[52] J. Liu, H. Ding, A. Shahroudy, L.-Y. Duan, X. Jiang, G. Wang, and A. C. Kot, "Feature boosting network for 3d pose estimation," *TPAMI*, 2019.

[53] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *ICPR*, 2014.

[54] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *CVPR*, 2014.

[55] E. Ohn-Bar and M. Trivedi, "Joint angles similarities and hog^2 for action recognition," in *CVPR Workshops*, 2013.

[56] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *TPAMI*, 2014.

[57] A. Shahroudy, G. Wang, and T.-T. Ng, "Multi-modal feature fusion for action recognition in rgb-d sequences," in *ISCCSP*, 2014.

[58] 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.

[59] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *ICCV*, 2015.

[60] 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.

[61] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," in *CVPR*, 2017.

[62] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," in *CVPR*, 2017.

[63] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *ICCV*, 2017.

[64] Q. Ke, J. Liu, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Global regularizer and temporal-aware cross-entropy for skeleton-based early action recognition," in *ACCV*, 2018.

[65] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *CVPR Workshops*, 2017.

[66] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *ICCV*, 2017.

[67] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *CVPR*, 2018.

[68] J. Cavazza, P. Morerio, and V. Murino, "When kernel methods meet feature learning: Log-covariance network for action recognition from skeletal data," in *CVPR Workshops*, 2017.

[69] F. Baradel, C. Wolf, and J. Mille, "Pose-conditioned spatio-temporal attention for human action recognition," *arXiv:1703.10106*, 2017.

[70] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," *CVPR*, 2018.

[71] J. Wang, A. Cherian, F. Porikli, and S. Gould, "Video representation learning using discriminative pooling," in *CVPR*, 2018.

[72] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang, "Deep bilinear learning for rgb-d action recognition," in *ECCV*, 2018.

[73] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer lstm networks," *TMM*, 2018.

[74] B. Zhang, J. Han, Z. Huang, J. Yang, and X. Zeng, "A real-time and hardware-efficient processor for skeleton-based action recognition with lightweight convolutional neural network," *TCS-II*, 2019.

[75] Q. Ke, J. Liu, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Computer vision for human–machine interaction," in *Computer Vision for Assistive Healthcare*, 2018.

[76] M. Liu, C. Chen, and H. Liu, "3d action recognition using data visualization and convolutional neural networks," in *ICME*, 2017.

[77] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *CVPR*, 2018.

[78] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *CVPR*, 2017.

[79] Z. Shi and T.-K. Kim, "Learning and refining of privileged information-based rnns for action recognition from depth sequences," in *CVPR*, 2017.

[80] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *ICCV*, 2017.

[81] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for rgb-d

based action recognition with convolutional neural networks," in *CVPR*, 2017.
[82] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "Skeletonnet: Mining deep part features for 3-d action recognition," *SPL*, 2017.
[83] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *CVPR*, 2015.
[84] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML*, 2015.
[85] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017.
[86] H. Yang, X. He, and F. Porikli, "One-shot action localization by learning sequence matching network," in *CVPR*, 2018.
[87] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *TPAMI*, 2006.
[88] P. Wang, L. Liu, C. Shen, Z. Huang, A. van den Hengel, and H. T. Shen, "Multi-attention network for one shot learning," in *CVPR*, 2017.
[89] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *NIPS*, 2016.
[90] S. R. Fanello, I. Gori, G. Metta, and F. Odone, "One-shot learning for real-time action recognition," in *ICPRIA*, 2013.
[91] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *TPAMI*, 2016.
[92] J. Konečnỳ and M. Hagara, "One-shot-learning gesture recognition using hog-hof features," *JMLR*, 2014.
[93] H. Chen, G. Wang, J.-H. Xue, and L. He, "A novel hierarchical framework for human action recognition," *PR*, 2016.
[94] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *TPAMI*, 2017.
[95] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *ASRU*, 2013.
[96] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.
[97] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *PRL*, 2016.
[98] C. Li, J. Cao, Z. Huang, L. Zhu, and H. T. Shen, "Leveraging weak semantic relevance for complex video event classification," in *ICCV*, 2017.
[99] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *ACL*, 2016.
[100] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.
[101] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017.
[102] J.-F. Hu, W. Zheng, L. Ma, G. Wang, J. Lai, and J. Zhang, "Early action prediction by soft regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[103] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *TPAMI*, 2018.
[104] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *PR*, 2017.
[105] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *TIP*, 2018.
[106] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *TIP*, 2018.
[107] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *CVPR*, 2018.
[108] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Ssnet: Scale selection network for online 3d action prediction," in *CVPR*, 2018.
[109] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv*, 2017.
[110] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.
[111] K. Simonyan and A. Zisserman., "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.
[112] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *CVPR*, 2016.