# A Benchmark Dataset and Comparison Study for Multi-modal Human Action Analytics



JIAYING LIU, SIJIE SONG, CHUNHUI LIU, YANGHAO LI, and YUEYU HU,
Institute of Computer Science and Technology, Peking University, China

Large-scale benchmarks provide a solid foundation for the development of action analytics. Most of the previous activity benchmarks focus on analyzing actions in RGB videos. There is a lack of large-scale and high-quality benchmarks for multi-modal action analytics. In this article, we introduce PKU Multi-Modal Dataset (PKU-MMD), a new large-scale benchmark for multi-modal human action analytics. It consists of about 28,000 action instances and 6.2 million frames in total and provides high-quality multi-modal data sources, including RGB, depth, infrared radiation (IR), and skeletons. To make PKU-MMD more practical, our dataset comprises two subsets under different settings for action understanding, namely Part I and Part II. Part I contains 1,076 untrimmed video sequences with 51 action classes performed by 66 subjects, while Part II contains 1,009 untrimmed video sequences with 41 action classes performed by 13 subjects. Compared to Part I, Part II is more challenging due to short action intervals, concurrent actions and heavy occlusion. PKU-MMD can be leveraged in two scenarios: action recognition with trimmed video clips and action detection with untrimmed video sequences. For each scenario, we provide benchmark performance on both subsets by conducting different methods with different modalities under two evaluation protocols, respectively. Experimental results show that PKU-MMD is a significant challenge to many state-of-the-art methods. We further illustrate that the features learned on PKU-MMD can be well transferred to other datasets. We believe this large-scale dataset will boost the research in the field of action analytics for the community.

CCS Concepts: • **Computing methodologies → Activity recognition and understanding**; *Supervised learning by classification*;

Additional Key Words and Phrases: Benchmark, multi-modal, action detection, action recognition

**ACM Reference format:**
Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. 2020. A Benchmark Dataset and Comparison Study for Multi-modal Human Action Analytics. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 2, Article 41 (May 2020), 24 pages.
https://doi.org/10.1145/3365212

This work was supported by National Natural Science Foundation of China under contract No. 61772043, Beijing Natural Science Foundation under contract No. 4192025, Microsoft Research Asia (FY19-Research-Sponsorship-115) and Peking University Tencent Rhino Bird Innovation Fund.

Authors' addresses: J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, Institute of Computer Science and Technology, Peking University, Zhongguancun North Street 128#, Haidian, Beijing, China; emails: {liujiaying, ssj940920, liuchunhui, lyttonhao, huyy}@pku.edu.cn.


1551-6857/2020/05-ART41 $15.00
https://doi.org/10.1145/3365212

## 1 INTRODUCTION

Action analytics in videos is an intensively studied research area, with broad applications in human-machine interaction, video surveillance, and robotics. In recent years, the success of deep learning has made data-driven learning methods get ahead with superior performance for human action analytics. Many efforts, including advanced methods [10, 20, 49, 52, 69, 77, 83] and benchmarks [3, 27, 47, 54], have been made to boost the research in this field.

A key category of action analytics methods aims at analyzing human actions in RGB videos [10, 24, 39, 52, 69, 83]. To facilitate the research in this branch, several famous large-scale datasets have been collected [3, 47]. For action detection, ActivityNet [3] is a superior RGB video dataset gathered from Internet media like YouTube with well-annotated labels and boundaries. For action recognition, UCF-101 [54] and HMDB-51 [27] have been popular benchmarks and have served the community well for many years. However, the quantities and variations of these two datasets limit their contributions to action recognition models based on deep learning. Kinetics [5] is a successor to the previous standard datasets, which is large enough to train deep networks from scratch and challenging enough to act as a performance benchmark.

In recent years, with the advent of affordable color-depth sensing cameras, such as RealSense and Kinect [23, 81], there is an increasing amount of visual data containing multi-modalities (e.g., RGB, depth, infrared radiation (IR), skeletons, etc.). Different modalities contain modal-specific characteristics and can be utilized to adapt to different application scenarios. RGB data are the most easily accessible in our daily life and are able to provide appearance information. Compared to conventional RGB videos, depth information is invariant to color and texture changes and less sensitive to illumination variations, which can help light up the accuracy of human action analysis. Heat-sensitive IR data, however, are more accessible in night-vision cameras. It is believed that the exploitation of IR information can greatly broaden the application scenarios of human action analytics, especially in night surveillance. Three-dimensional (3D) skeletons are intrinsic high-level representations that are robust to viewpoints, illumination, and cluttered backgrounds. They are comprehensive for summarizing a series of human dynamics in the videos. Besides, the low dimension of 3D skeleton data makes it possible to achieve real-time computing.

Due to the aforementioned advantages of different modalities, many efforts have been devoted to exploring action analytics based on depth information [76], IR [25], or skeletons [11]. In the meantime, inspired by the intuition that different modalities are capable of providing complementary information, multi-modal action analytics has attracted much attention [18, 38, 51, 73]. However, due to the lack of large-scale and high-quality benchmarks, there are not sufficient data to exploit the potential of deep models. To the best of our knowledge, existing action benchmarks have limitations in the following aspects.

• **Limitation in data modalities:** As mentioned above, different modalities intuitively capture features from distinctive aspects and provide complementary information. Nevertheless, most current datasets focus mainly on one modality of action representations. A few of them contain RGB, depth, and skeleton information [8, 32, 40, 55]. However, the data qualities are not so satisfactory, i.e., the misalignment of multi-modal data and the low resolution of depth maps. Besides, most datasets do not provide IR data. However, the easy access of IR data from night-vision cameras makes IR-based action analytics in demand to facilitate practical applications.

• **Shortage in large-scale action analytics datasets:** The recent action analytics methods are mainly based on data-hungry deep learning models. Current datasets are simply not large enough to train the network from scratch, especially for modalities like depth, IR, and skeletons. There is no doubt that more configuration diversities could enlarge the intra-class difference and narrow the inter-class variations. Larger datasets usually make the issues even more challenging and will enable a new generation of action analytics algorithms.

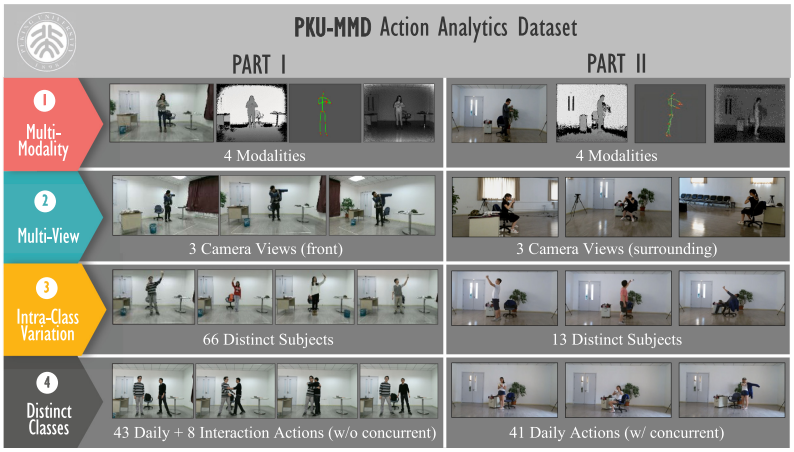



Fig. 1. Overview of the PKU-MMD Dataset. Our dataset consists of two subsets: Part I and Part II. Compared with Part I, in which the viewpoints are from front, Part II provides surrounding viewpoints, and thus there are more occlusion. Besides, the action intervals are shorter and concurrent actions are taken into account in Part II.

Though the NTU RGB+D dataset [49] is a famous large-scale multi-modal dataset, it only contains trimmed videos for action recognition. There is a lack of large-scale datasets with untrimmed videos to facilitate multi-modal action detection, which also plays a key role in action analytics. In this article, we develop a new large-scale multi-modal human activity dataset (PKU-MMD) that contains untrimmed videos and can be utilized for action recognition and detection.[1] To make our dataset more practical, we consider different settings in the data collection process, which then lead to different levels of difficulty (i.e., easy and hard) in action understanding. As shown in Figure 1, our dataset includes two subsets. Both subsets are recorded from multiple viewpoints and provide synchronous multi-modal data, including RGB frames, depth maps, IR information, and skeletons. The previous conference paper [36] mainly discusses the benchmark performance on action analytics with skeleton data from Part I, which consists of 1,076 untrimmed sequences with 51 action classes performed by 66 subjects. In this article, we extend the dataset with Part II, which contains additional 1,009 untrimmed sequences with 41 action classes performed by 13 actors. Compared to Part I, there is a larger viewpoint variation in Part II with surrounding cameras. Besides, action intervals are shorter and concurrent actions are taken into account, leading to less clear action boundaries. And the occlusion caused by viewpoints results in heavier skeleton noises. Thus, Part II is more challenging for action understanding. In experiments, we explore multi-modal action analytics with all the modalities provided. A comparison study of different action analytics methods with different modalities is given. We also demonstrate that PKU-MMD can be utilized to support action understanding on other datasets. Our contributions can be summarized as follows:

- We build the currently largest multi-modal dataset PKU-MMD for action analytics with sufficient variations in viewpoints, subjects, and action types. Our dataset consists of two subsets under different settings and with different levels of difficulty in action understanding. Besides, our dataset can be utilized for trimmed action recognition and untrimmed action detection.

[1]Our dataset can be found at https://struct002.github.io/PKUMMD/.

- We conduct extensive and systematic experiments to quantitatively compare several action recognition and detection methods. Our evaluation and analysis demonstrate that our dataset is a great challenge to many state-of-the-arts for action analytics.
- We analyze the feasibility of multi-modal action recognition and detection. We show that each modal data contributes to action analytics, while their fusion results achieve further remarkable performance.
- We show that the features learned on PKU-MMD can be well transferred and utilized to support action understanding on other datasets through a cross-dataset study.

The rest of this article is organized as follows. In Section 2, we review the development of action analytics and popular datasets. In Section 3, we introduce PKU-MMD in detail and explain the evaluation protocols. Then we show our benchmarks and further analysis on action recognition and detection in Section 4. Concluding remarks are finally given in Section 5.

## 2 RELATED WORK

In this section, we summarize the development of action analytics and then briefly introduce a series of approaches and benchmarks for multi-modal action analysis. For a more extensive conclusion of human action analytics, we refer to corresponding survey papers [4, 80].

### 2.1 Development of Action Analytics

From the topics to be addressed, action analytics can be categorized into action recognition and action detection. Action recognition aims to label a trimmed video clip, while action detection refers to not only recognizing but localizing actions within an untrimmed video. Early human action analytics mainly focuses on action recognition, in which extracting robust video representations is the key issue. Traditional methods employ hand-crafted descriptors for video representations. As a kind of low-level video representation, densely tracking points in the optical flow field with more features like Histogram of Oriented Gradient (HOG), Histogram of Flow (HOF), and Motion Boundary Histograms (MBH) [43, 59] achieve good performance in action recognition. And there are also many works on middle video representations. A spatio-temporal latent variable model is developed in Reference [46] to form clusters of trajectories. Wang et al. [65] defined motionlet as a middle-level representation, which is a spatio-temporal part with coherent appearance and motion features. In Reference [21], spatio-temporal patches are mined according to their discriminative and representative properties. To create a more dominant and compact representation, Zhu et al. [84] proposed a two-layer structure to automatically exploit a mid-level video feature. In recent years, deep neural networks have been exploited for action recognition [10, 52, 69, 73]. These approaches automatically learn robust video representations directly from raw data. Convolutional neural networks (CNN) are usually constructed to model spatial features [52, 69] and recurrent neural networks (RNN) have been utilized to handle temporal relations [10, 73].

For action detection, early methods utilize sliding-window schemes [50, 64]. These methods usually have low computational efficiency or unsatisfactory localization accuracy due to the overlapping design and unsupervised localization scheme. More recent works employ action proposal approaches [12, 22, 67], which are more efficient in retrieving temporal segments. The methods mentioned above are designed for offline action detection [50, 57, 71], which generate action boundaries after observing the entire video sequence. There are some works [17, 33, 48, 79] recognizing and locating actions on the fly before completion of the action. Hoai et al. [17] enabled early event detection by proposing a learning formulation based on a structural SVM. Leveraging the merits of Long Short-Term Memory (LSTM) network, Li et al. [33] introduced a joint classification and regression network to forecast the occurrence of start and end of actions.

In the meantime, with the development of action analytics, the video sources and application scenarios are becoming more and more diverse and challenging. Early action analytics targets home surveillance activities like *drinking* or *waving hands*. These videos are easy and cheap to capture. Thus, the analysis of these simple indoor activities is the starting point of action analytics. Due to the rapid development of Internet, video data from online media like YouTube [3, 54] are easier to access. Recently, there are also several works aiming at collecting datasets in certain fields like TV series [9], movies [28], and the Olympic Games [26]. The videos from Internet suffer from camera motion, illumination variations, background clutter, and so on. Therefore, it requires more robust feature representations to achieve high-quality action recognition and detection. In addition, the launch of color-depth sensing cameras like Microsoft Kinect broadens the diversity of video data modalities, as well as the application scenarios of human action analytics. As Kinect provides a real-time algorithm to generate information of RGB, depth, IR, and skeletons, it becomes an ideal source to support real-time algorithms and to be utilized on devices like robots or mobile phones. Thus, researchers are encouraged to develop multi-modal action analytics.

### 2.2 Multi-Modal Action Analytics

In recent years, many algorithms have been designed for action analytics with a single modality. For RGB videos, the two-stream architecture [52] is a classical structure for action recognition and has become a backbone of many other approaches [14, 69]. For depth videos, action analytics mostly relies on hand-crafted heuristics [42, 76], which typically extracts spatio-temporal features from interest points to describe the local appearance. For skeleton-based action analytics, conventional methods are designed to represent geometric relationships of body parts [58, 79]. In References [11, 19, 53, 85], recurrent neural networks are utilized to model the temporal dynamics automatically for skeletons and obtain competitive performance. More recently, graph convolutional networks [75] are employed for skeleton-based action recognition to improve expressive power and generalization capacity of deep video features.

Inspired by the fact that different modalities provide complementary information, there are some works integrating multiple modalities to leverage the compensated feature learning. As a compact human representation, poses are used as guidance to extract high-level activity information and then improve action understanding [66]. With the advance in depth cameras, 3D skeletons, depth, and IR images are more available. Several works employ a kind of modality as auxiliary data that are required in training and discarded in testing [38, 51]. Based on the assumption that RGB and skeleton data share similar high-level feature spaces, a regularized LSTM is developed in Reference [38] to enhance the feature learning from RGB sequences. Shi and Tim [51] proposed to achieve action recognition from depth sequences by learning an RNN with privileged information from skeletons. In addition, some works combine data from several modalities for both training and testing. A chained multi-stream network [86] built on a Markov chain model is developed to integrate appearance, motion and pose features. In Reference [18], Hu et al. found that features from different channels (RGB, depth) share similar hidden structures and proposed a joint model to explore the shared and specific features. The existing works illustrate that the introduction of multiple cues from different modalities effectively improves the performance of action analytics.

### 2.3 Multi-Modal Activity Datasets

We have surveyed tens of well-designed action datasets that greatly improved the study of multi-modal action analytics. A comparison with several datasets and PKU-MMD is given in Table 1.

*CMU Mocap* [8] is the early resource including skeletons for action recognition. Captured by the motion capture system, it is able to provide accurate skeletons with a variety of human actions, such as sports, human locomotions, interactions, and so on.

Table 1. Comparison of Multi-modal Datasets

| Datasets | Classes | Videos | Labeled Instances | Actions per Video | Trimmed | Views | Modalities | Temporal Label | Year |
|---|---|---|---|---|---|---|---|---|---|
| CMU Mocap [8] | 45 | 2,235 | 2,235 | 1 | T | 1 | R/S | No | 2001 |
| HDM05 [40] | 130 | 2,337 | 2,337 | 1 | T | 1 | R/S | No | 2007 |
| MSR-Action3D [32] | 20 | 567 | 567 | 1 | T | 1 | R/D/S | No | 2010 |
| CAD-60 [55] | 12 | 60 | 60 | 1 | T | — | R/D/S | No | 2011 |
| MSR-DailyActivity [61] | 16 | 320 | 320 | 1 | T | 1 | R/D/S | No | 2012 |
| ACT4 [7] | 14 | 6,844 | 6,844 | 1 | T | 4 | R/D | No | 2012 |
| UTKinect-Action [74] | 10 | 200 | 200 | 1 | T | 4 | R/D/S | No | 2012 |
| 3D Action Pairs [42] | 12 | 360 | 360 | 1 | T | 1 | R/D/S | No | 2013 |
| DML-SmartAction [1] | 12 | 932 | 932 | 1 | T | 3 | R/D | No | 2013 |
| MHAD [41] | 11 | 660 | 660 | 1 | T | 4 | R/D/S | No | 2013 |
| Multiview 3D Event [70] | 8 | 3,815 | 3,815 | 1 | T | 3 | R/D/S | No | 2013 |
| Northwestern-UCLA [62] | 10 | 1,475 | 1,475 | 1 | T | 3 | R/D/S | No | 2014 |
| UWA3D Multiview [45] | 30 | ~900 | ~900 | 1 | T | 1 | R/D/S | No | 2014 |
| Office Activity [63] | 20 | 1,180 | 1,180 | 1 | T | 3 | R/D | No | 2014 |
| UTD-MHAD [6] | 27 | 861 | 861 | 1 | T | 1 | R/D/S | No | 2015 |
| TJU Dataset [35] | 22 | 1,936 | 1,936 | 1 | T | 1 | R/D/S | No | 2015 |
| UWA3D Multiview II [44] | 30 | 1,075 | 1,075 | 1 | T | 5 | R/D/S | No | 2015 |
| SYSU 3D HOI Set [18] | 12 | 480 | 480 | 1 | T | 1 | R/D/S | No | 2015 |
| NTU RGB+D [49] | 60 | 56,880 | 56,880 | 1 | T | 80 | R/D/IR/S | No | 2016 |
| G3D [2] | 20 | 210 | 1,467 | 7 | U | 1 | R/D/S | Yes | 2012 |
| SBU Kinect interaction [78] | 8 | 21 | 300 | 14.3 | T/U | 1 | R/D/S | Yes | 2012 |
| MSRC-12 [15] | 12 | 594 | 6,244 | ~11 | U | — | S | Yes | 2012 |
| CAD-120 [56] | 20 | 120 | ~1,200 | ~8.2 | U | — | R/D/S | Yes | 2013 |
| Compostable Activities [34] | 16 | 693 | 2,529 | 3.6 | U | 1 | R/D/S | Yes | 2014 |
| Watch-n-Patch [72] | 21 | 458 | ~2,500 | 2~7 | U | — | R/D/S | Yes | 2015 |
| OAD [33] | 12 | 59 | ~700 | ~12 | U | 1 | R/D/S | Yes | 2016 |
| **PKU-MMD (Part I)** | **51** | **1,076** | **21,545** | **20.02** | T/U | 3 | **R/D/IR/S** | **Yes** | **2017** |
| **PKU-MMD (Part II)** | **41** | **1,009** | **6,952** | **6.89** | T/U | 3 | **R/D/IR/S** | **Yes** | **2018** |

R: color videos; S: skeletons; D: depth maps; IR: infrared images; T: trimmed videos; U: untrimmed videos.

*HDM05* [40] is captured by an optical marker-based technology. It provides RGB videos and skeletons, containing over 2,000 videos and 130 human actions.

*MSR Action3D Dataset* [32] is one of the earliest datasets that capture multi-modal data with Kinect devices. This dataset is composed by instances chosen in the context of interacting with game consoles like *high arm wave*, *horizontal arm wave*, *hammer*, and *hand catch*. The skeleton data are provided with 3D locations of 20 joints with 15 fps.

*CAD-60* [55] and *CAD-120* [56] are a series of multi-modal datasets for action recognition and detection, respectively. The camera views are not fixed for actors. Compared to CAD-60, CAD-120 provides extra labels of temporal locations. However, the two datasets are limited in the number of video samples.

*ACT4* [7] is a large dataset designed to facilitate practical applications in real life. The action categories in ACT4 mainly focus on the activities of daily livings.

*Multiview 3D Event* [70] and *Northwestern-UCLA* [62] datasets start to use multi-view configuration to capture videos, which is followed in the collection of many succeeding datasets.

Table 2. The Desirable Properties of the PKU-MMD Dataset

| Properties | Features |
|---|---|
| Large Scale | Extensive action categories |
| | Massive samples for each class |
| Diverse Modality | Three camera views |
| | Sufficient action variations |
| | Multi-modality (RGB, depth, IR, etc.) |
| Wide Application | Trimmed clips for recognition |
| | Continuous videos for detection |
| | Inner analysis of context-related actions |

*SYSU 3D HOI Set* [18] focuses on human-object interactions. The involved motions and the appearance of objects are highly similar. More inter-subject variations are observed due to more participants, making the dataset more challenging.

*NTU RGB+D Dataset* [49] is a state-of-the-art large-scale benchmark for action recognition with sufficient data modalities. It illustrates a series of evaluation protocols and provides valuable experience for large-scale data collection.

*G3D* [2] is designed for real-time action detection in gaming containing synchronized videos. As the earliest activity detection dataset, most sequences of G3D contain multiple actions in a controlled indoor environment with a fixed camera, and a typical setup for gesture-based gaming.

*MSRC-12* [15] is a gesture/action dataset only with 3D skeleton data captured by a Kinect sensor. The dataset comprises 594 sequences collecting from 30 people performing 12 gestures. It contains information about when a particular gesture should be detected.

*Watch-n-Patch* [72] and *Compostable Activities* [34] are the datasets consisting of the continuous sequences to learn high-level action co-occurrence and temporal relations. They consist of moderate number of action instances. The dataset is recorded in different environments under different views.

*OAD* [33] is a new dataset targeting online action detection and forecasting. Fifty-nine videos describing daily activities are captured by Kinect v2 devices. This dataset defines several criteria for online action detection.

However, as the quick development of deep learning-based action analytics, these datasets are not able to satisfy the demand of data-driven algorithms. Therefore, we collect PKU-MMD dataset to overcome their drawbacks from the following perspectives in Table 2.

## 3 THE PKU-MMD DATASET

In this section, we first describe the details of PKU-MMD and then define the evaluation protocols on the dataset.

### 3.1 Overview of the Dataset

PKU-MMD is our new large-scale dataset focusing on multi-modal action analytics, including action recognition and action detection. The dataset is captured via the Kinect v2 sensors from multiple viewpoints with recording ratio set as 25 fps, which collect color images, depth maps, IR sequences, and human skeleton joints synchronously. RGB videos are recorded in the provided resolution of $1920 \times 1080$. Depth maps are sequences of 2D depth values in millimeters. To maintain all the information, we apply lossless compression for each individual frame. The resolution of each depth frame is $512 \times 424$. IR sequences are also collected and stored frame by frame in the resolution of $512 \times 424$. Skeleton information consists of 3D locations of 25 major body joints as in

Table 3. A Detailed List about 51 Action Categories in the PKU-MMD Dataset

| Taxonomy | Detailed Actions[$] | | |
|---|---|---|---|
| Health related | touch head (headache) | touch neck (neckache) | touch back (backache) |
| | touch chest (stomachache/heart pain) | | |
| Home related | brush teeth | comb hair | wipe face |
| | drink water | eat meal/snack | |
| Dressing related | put on glasses | put on jacket | put on a hat/cap |
| | take off glasses | take off jacket | take off a hat/cap |
| Interaction with people | handshake* | push other person* | hug other person* |
| | kick other person* | punch/slap other person* | pat on back of other person* |
| | point finger at the other person* | give something to other person* | |
| Interaction with items | drop | write | read |
| | pick up | take a selfie | tear up paper |
| | type on a keyboard | play with phone/tablet | check time (from watch) |
| | use a fan (with hand or paper) | make a phone call/answer phone | point to something with finger |
| | put something inside pocket* | take out something from pocket* | |
| Human locomotion | bow | clap | throw |
| | salute | fall | hop (one foot jumping) |
| | sit down | stand up | cheer up |
| | jump up | kick something | hand waving |
| | rub two hands together | cross hands in front (say stop) | |

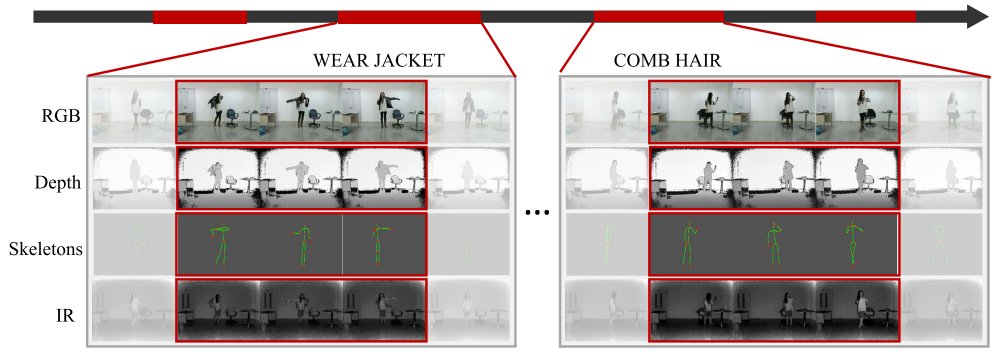
[$] The actions not shown in Part II are marked with *.

Reference [49] for each detected and tracked human body in the scene. Since some skeletal joints are not available or untracked due to occlusion, we further provide the confidence of each joint as an appendix (i.e., 0 for untracked joints, 1 for noisy joints, and 2 for good joints). For action types, our dataset covers health-related actions, home-related actions, dressing-related actions, and so on. Table 3 illustrates more details on action categories.

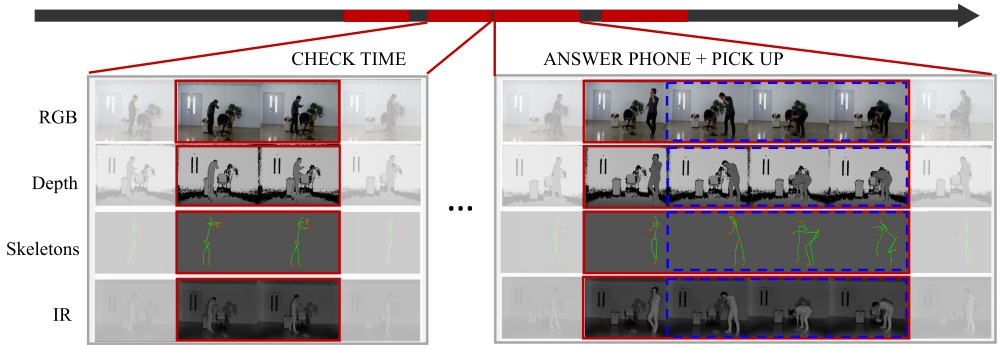
Overall, the scale of PKU-MMD is 2,085 untrimmed sequences with approximately 6.2 million frames. The videos are about 4,000 minutes in total with over 28,000 temporally localized action clips. More specifically, it consists of two subsets under different settings and with different levels of difficulty in action understanding. Figure 2 gives some sample videos from PKU-MMD. The untrimmed videos can be utilized for action detection. Meanwhile, we can get trimmed clips according to annotations of action localization and use them for action recognition.

Part I: In this subset, we collect 1,076 long continuous action sequences, each of which lasts about 3 to 4 minutes and contains approximately 20 action instances. We invite 66 distinct subjects for Part I. Each subject takes part in 4 daily action videos and 2 interactive action videos. There are 51 action classes, all of which are shown in Table 3. We use three cameras at the fixed angles and heights to capture different horizontal views synchronously. The horizontal angles of each camera are $-45^\circ$, $0^\circ$, and $+45^\circ$ as shown in Figure 3(a), with a height of 120cm.

Part II: In this subset, we collect 1,009 untrimmed videos, each of which lasts about 1 to 2 minutes and contains about 7 action instances. For this subset, we invite 13 subjects and each subject takes part in 4 daily action videos. Part II shares 41 action labels with Part I, as shown in Table 3. In addition, we consider action relations in the untrimmed videos. For example, we design an action sequence of *touching head*, *falling down*, and *touching chest* to imitate the scene in medical care and another of *reading*, *checking time* and *answering phone* as the office scene. We set surrounding cameras around subjects as shown in Figure 3(b). The horizontal angles of each camera are $-120^\circ$, $0^\circ$, and $+120^\circ$ with the height of 120 cm. Compared with Part I, this subset is relatively challenging in the following aspects.

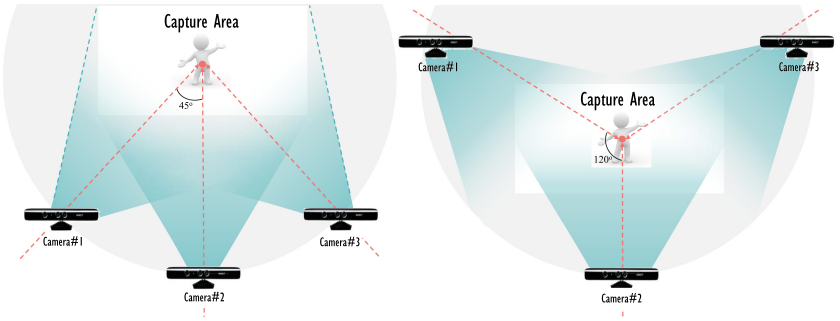



(a) A sample video from *Part* I.



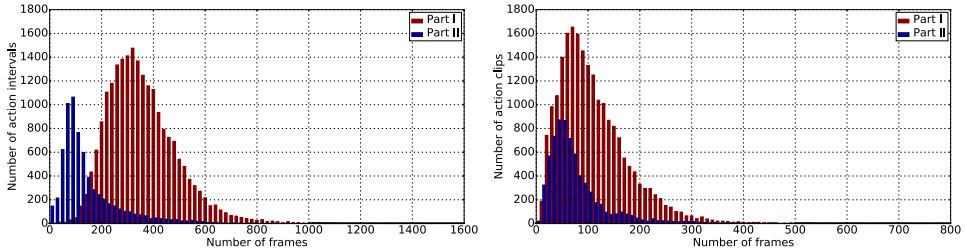

(b) A sample video from *Part* II.

Fig. 2. Sample videos from Part I and Part II from PKU-MMD, respectively. Part II is more challenging for action understanding due to dense actions with shorter intervals, concurrent actions (e.g., answering phone in the red box and picking up in the blue box), and heavy occlusion.




(a) *Part* I (b) *Part* II

Fig. 3. Camera settings for Part I and Part II of the PKU-MMD dataset, respectively. The cameras in Part I are set in the front of the subject, while in Part II, we have the subject surrounded by cameras.




(a) Action interval duration for *Part* I and *Part* II, respectively.

(b) Action clip duration for *Part* I and *Part* II, respectively.

Fig. 4. Statistics for action interval and action clip duration of PKU-MMD.

• **Short action intervals.** As shown in Figure 2, there is long idle time between action clips in the video from Part I (Figure 2(a)), but action clips in Part II are close to each other (Figure 2(b)), leading to quite short action intervals. In Figure 4(a), we statistically show the number of frames between adjacent action clips for Part I and Part II, respectively. We can see that in Part I, most action intervals are about 350 frames, while in Part II, most action intervals about 100 frames. The shorter action intervals result in less clear action boundaries, making it challenging to locate the start and end points for each action accurately. Besides, the duration of most actions in Part I is about 100 frames and that in Part II is about 50 frames according to Figure 4(b).

• **Concurrent actions.** It is common for people to perform multiple actions at the same time in realistic scenarios. Thus, in Part II, we take concurrent actions into account. For example, the man is answering phone and picking up in the meanwhile as shown in Figure 2(b). It is harder for the classifiers to learn patterns for a specific action from the mixture of several actions. However, in Part I, each action clip only contains a single action and it is much easier to train an action classifier from such data.

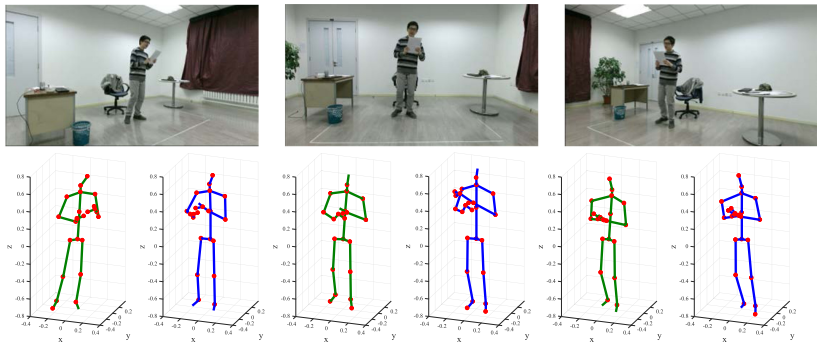
• **Heavy occlusion.** As shown in Figure 3, the camera settings for Part I and Part II are quite different. The viewpoints in Part II lead to heavier occlusion, because the cameras on each side always capture human actions from back. It will especially influence the quality of skeleton data. We show sample frames and corresponding skeletons below each RGB image in Figure 5. Note that for better visualization, we rotate the skeletons to the front, which is to fix the $X$-axis to be parallel to the vector from "left shoulder" to the "right shoulder." The original skeletons and rotated skeletons are shown in green and blue, respectively. It is observed that skeletons in Part II suffer from more noises due to heavier occlusion, as in the right and left figures in Figure 5(b).

### 3.2 Evaluation Protocols

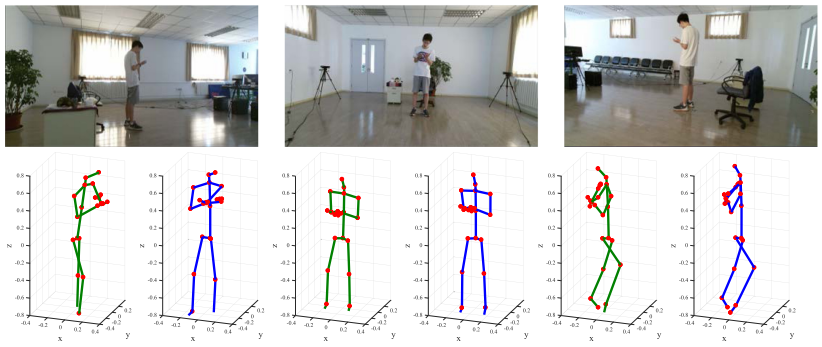
We now introduce standard evaluation protocols for all the reported results on our benchmark. We first illustrate the data split settings, and then the evaluation criteria for action recognition and detection are given, respectively.

*3.2.1 Dataset Splits.* In our benchmark, we suggest two data splits (i.e., cross-subject and cross-view) for the scenarios of action recognition and detection, respectively. The summary of data splits for Part I and Part II is given in Table 4. For simplicity, we use "CS" for cross-subject and "CV" for cross-view.

**Cross-Subject Evaluation:** Cross-subject evaluation aims to test the ability to handle intra-class variations among different actors. For Part I, 57 subjects are chosen to be training samples and 9 for testing. In action recognition, there are 19,114 action clips for training and 2,730 for testing, respectively. In action detection, there are 944 untrimmed video samples for training and 132 for



(a) Sample frames from *Part* I



(b) Sample frames from *Part* II

Fig. 5. Sample frames from different viewpoints in Part I and Part II, respectively. Below each RGB frame are its corresponding skeletons. The viewpoints in Part II lead to heavier occlusion and thus introduce more skeleton noises. We show the original skeletons in green and rotated skeletons in blue.

Table 4. Data Splits for Part I and Part II, Respectively

| Splits | Attributes | Part I | | Part II | |
|---|---|---|---|---|---|
| | | Train | Test | Train | Test |
| Cross-subject | #Subjects | 57 | 9 | 10 | 3 |
| | #Action clips | 19,114 | 2,730 | 5,339 | 1,613 |
| | #Untrimmed videos | 944 | 132 | 775 | 234 |
| Cross-view | View | #1, #3 | #2 | #1, #2 | #3 |
| | #Action clips | 14,545 | 7,299 | 4,622 | 2,330 |
| | #Untrimmed videos | 717 | 359 | 671 | 338 |

testing, respectively. For Part II, 10 subjects are chosen to be training samples and 3 for testing. In action recognition, there are 5,339 action clips for training and 1,613 for testing, respectively. In action detection, there are 775 untrimmed sequences for training and 234 for testing, respectively.

**Cross-View Evaluation:** Cross-view evaluation aims to test the robustness in terms of transformation (e.g., translation, rotation). For Part I, the videos from Camera #1 and #3 are chosen as the training set, and those from Camera #2 are as the testing set. In action recognition, there are 14,545 and 7,299 action clips for training and testing, respectively. In action detection, there are 717 untrimmed videos for training and 359 for testing. For Part II, the videos from Camera #1

and #2 are chosen as the training set, and those from Camera #3 are as the testing set. In action recognition, there are 4,622 and 2,330 action clips for training and testing, respectively. In action detection, there are 671 untrimmed videos for training and 338 for testing.

*3.2.2 Evaluation Criteria.* For all the evaluation on PKU-MMD, the action recognition results are reported with the classification accuracy in percentage. In the following, we explain the criterion for action detection.

An action proposal generated in the process of action detection is defined as positive, when the overlapping ratio between the proposed interval $I$ and the ground-truth interval $I^*$ exceeds a threshold $\theta$, which is given as

$$\frac{|I \cap I^*|}{|I \cup I^*|} > \theta, \tag{1}$$

where $I \cap I^*$ denotes the intersection of the predicted and ground-truth intervals and $I \cup I^*$ denotes their union. We then use mean average precision (mAP) to evaluate the performance of action detection.

Mean average precision is a common evaluation protocal using the information of confidence for ranked detection results. Its definition is based on the interpolated average precision [13], which is able to remove jiggles on the precision-recall curve. The interpolated precision $p_{interp}$ at a certain recall level $r$ is formulated as

$$p_{interp}(r, \theta) = \max_{r' \geq r} p(r', \theta), \tag{2}$$

where $p(r, \theta)$ is the precision-recall function under threshold $\theta$. Then the mean average precision is defined by

$$\text{mAP}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{a_n} \sum_{k=1}^{a_n} p_{interp}(r_{nk}, \theta), \tag{3}$$

where $N$ is the total number of action classes, for each class with type id of $n$, there are $a_n$ action occurrences and $r_{nk}$ is the recall result of the $k$th ranked detections.

## 4 COMPARISON STUDY

In this section, we present the comparison study through a series of evaluations on our benchmark to show differences of different modalities and action analytics methods. To show the characteristics for each subset more clearly, we separately evaluate two subsets, while they can be combined together in practice if needed. Our experiments are in terms of action recognition and action detection. We first describe the video representations adopted in our experiments. Then the employed action analytics methods for each task are introduced. We carry out preliminary analysis on the recognition and detection performance on each modality, as well as multi-modal fusion results. The analysis also serves to illustrate the feasibility and challenges in multi-modal action analytics and call on new explorations.

The modalities involved in our experiments include the following: RGB (R), optical flow (OF), depth (D), infrared radiation (IR), and skeletons (S). We adopt dense flow in Reference [69] to calculate optical flow from RGB images for its efficiency.

### 4.1 Action Recognition

Here we show the performance of Part I and Part II on action recognition, the goal of which is to assign a label of a well-trimmed video clip.

Table 5. The Number of Neurons in LSTM and BLSTM Networks

| | LSTM | | | BLSTM | | | | |
|---|---|---|---|---|---|---|---|---|
| | $l_1$ | $l_2$ | $l_3$ | $bl_1$ | $fc_1$ | $bl_2$ | $fc_2$ | $bl_3$ |
| S | 100 | 110 | 200 | $100 \times 2$ | 100 | $110 \times 2$ | 110 | $200 \times 2$ |
| R/F/D/IR | 200 | 220 | 400 | $100 \times 2$ | 100 | $110 \times 2$ | 110 | $200 \times 2$ |

*4.1.1 Recognition Methods.* To evaluate the performance of different modalities, we adopt the following methods. We do not perform temporal downsampling for all modalities in action recognition. Note that STA-LSTM [53], TPN [19], VA-LSTM [82], and ST-GCN [75] are only for skeletons.

**TSN:** Temporal Segment Network [69] is an efficient and remarkable approach for RGB-based action recognition. Based on the BN-Inception network, it enables robust feature learning with frames from each temporal segment. We leverage TSN on the evaluation of RGB, optical flow, depth, and IR images.

**STA-LSTM:** Spatio-temporal attention LSTM network [53] consists of a spatial attention model to automatically select discriminative joints and a temporal attention model to allocate importance to different frames.

**TPN:** Temporal perceptive network [19] embeds a convolutional subnetwork to enhance the feature extraction from local temporal dynamics. This method effectively improves the accuracy in large-scale action recognition on skeleton data.

**VA-LSTM:** View-adaptive network [82] aims to transform skeletons adaptively towards a suitable observation viewpoints. The model is more generalizable to multi-view skeleton data.

**ST-GCN:** Spatio-temporal graph convolutional network in Reference [75] learns the spatial and temporal patterns with automatically with greater expressive power and stronger generalizable capability, and shows impressive performance on action recognition with skeleton data.

**LSTM/BLSTM:** The merits of LSTM layers allow the network to exploit the history information and model the temporal dynamics efficiently. Here we adopt LSTM (with unidirectional recurrent layers denoted as $l$) and BLSTM (with bidirectional recurrent layers denoted as $bl$) to achieve action recognition on each modal data. Our LSTM network is composed of three layers, i.e., $l_1 - l_2 - l_3$. Similarly, the structure of our BLSTM network is $bl_1 - fc_1 - bl_2 - fc_2 - bl_3$, where $fc$ represents the fully connected layer. The detailed configurations of LSTM/BLSTM networks are shown in Table 5, which gives the number of neurons for each layer. Motivated by Reference [10], we extract deep features from convolutional layers for RGB, optical flow, depth, and IR data and feed them to LSTM/BLSTM networks. Specifically, each frame is finally represented by a vector of dimension 1,024 from the layer *global pool* of BN-Inception network [69], which is well trained for each corresponding modality. Then the features are fed into the following recurrent layers.[2]

*4.1.2 PKU-MMD Recognition Benchmarks.* We evaluate each method with corresponding modal data, including single-modal recognition and multi-modal recognition.

**Single-Modal Recognition:** Table 6 shows the results for single-modal recognition with different methods on Part I and Part II, respectively. We get some similar observations in both parts. For the modalities of RGB, optical flow, depth, and IR, we can see that based on the deep features from TSN [69], LSTM and BLSTM effectively improve the recognition results in most cases due to better temporal modelling. For skeleton data, STA-LSTM [53] and TPN [19] achieve competitive recognition accuracy, and ST-GCN [75] achieves the best performance in most cases. For a specific classifier, the recognition accuracy from optical flow is the highest. It is because the actions in

[2]For more detailed structures, please refer to our website: https://struct002.github.io/PKUMMD/.

Table 6. Recognition Accuracy (%) of Different Modalities in PKU-MMD and NTU Using Different Methods

| Partition Setting | Cross-subject (Part I) | | | | | Cross-view (Part I) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | R | OF | D | IR | S | R | OF | D | IR | S |
| TSN [69] | 79.0 | 87.9 | 79.0 | 73.2 | – | 84.1 | 90.3 | 76.9 | 68.0 | – |
| STA-LSTM [53] | – | – | – | – | **87.2** | – | – | – | – | 90.8 |
| TPN [19] | – | – | – | – | 85.7 | – | – | – | – | 93.7 |
| VA-LSTM [82] | – | – | – | – | 84.1 | – | – | – | – | 92.5 |
| ST-GCN [75] | – | – | – | – | 84.1 | – | – | – | – | 92.0 |
| LSTM | 84.0 | 90.0 | 85.1 | 80.4 | 86.7 | 88.6 | 92.1 | 81.2 | 80.1 | 94.0 |
| BLSTM | **84.8** | **90.1** | **86.2** | **80.6** | 86.4 | **88.6** | **93.1** | **83.7** | **80.1** | **94.6** |
| Partition Setting | Cross-subject (Part II) | | | | | Cross-view (Part II) | | | | |
| Methods | R | OF | D | IR | S | R | OF | D | IR | S |
| TSN [69] | 57.2 | 70.7 | 51.3 | 51.3 | – | 56.3 | 74.3 | 42.4 | 41.9 | – |
| STA-LSTM [53] | – | – | – | – | 44.3 | – | – | – | – | 28.6 |
| TPN [19] | – | – | – | – | 46.9 | – | – | – | – | 29.7 |
| VA-LSTM [82] | – | – | – | – | **50.0** | – | – | – | – | 34.5 |
| ST-GCN [75] | – | – | – | – | 48.2 | – | – | – | – | **35.8** |
| LSTM | 60.1 | 70.4 | 47.5 | 52.9 | 44.3 | 59.3 | 72.6 | 49.6 | 43.3 | 28.1 |
| BLSTM | **66.4** | **71.9** | **56.5** | **58.0** | 44.4 | **63.8** | **74.4** | **51.5** | **49.2** | 26.3 |
| Partition Setting | Cross-subject (NTU) | | | | | Cross-view (NTU) | | | | |
| Methods | R | OF | D | IR | S | R | OF | D | IR | S |
| TSN [69] | 74.3 | 85.2 | 70.5 | 67.8 | – | 76.4 | 87.2 | 63.6 | 60.0 | – |
| STA-LSTM [53] | – | – | – | – | 73.4 | – | – | – | – | 81.2 |
| TPN [19] | – | – | – | – | 75.3 | – | – | – | – | 84.0 |
| VA-LSTM [82] | – | – | – | – | 79.4 | – | – | – | – | 87.6 |
| ST-GCN [75] | – | – | – | – | **81.5** | – | – | – | – | **88.3** |
| LSTM | **81.8** | **87.6** | **82.8** | **77.9** | 71.9 | **88.5** | **93.3** | **79.0** | **73.3** | 82.0 |
| BLSTM | 80.5 | 87.4 | 82.7 | 77.1 | 71.4 | 86.4 | 91.5 | 78.1 | 72.0 | 81.9 |

PKU-MMD are motion related and optical flow provides pixel-level motion vectors. Meanwhile, we obtain the best results from BLSTM for almost all modalities, since it is able to utilize both history and future frames.

It is also noticeable that the results of Part II are much inferior than those of Part I. We further conduct comprehensive experiments under different configurations to analyze the difficulties of each subset. Table 6 presents the cross-subject results with Part I and Part II, respectively. Table 7 further shows the cross-view results on RGB and skeletons in different train/test splits. It is observed that action recognition performance with Part II is much worse than Part I under all the settings, illustrating that Part II is more challenging compared to Part I. On the one hand, it is more difficult for classifiers to learn patterns for a specific label due to concurrent actions. On the other hand, the occlusion caused by viewpoints represents a real-world challenge, especially the noises in skeletons. It is reflected in the performance with skeleton data in Part II (see Table 6 and Table 7), which is only about 26–50% in terms of accuracy.

To further illustrate the characteristics of PKU-MMD dataset on action recognition, we also conduct experiments on the well-known NTU dataset [49] with state-of-the-art methods. The results in Table 6 show that for Part I in PKU-MMD, action recognition performance on RGB and optical flow are comparable with that on the NTU dataset, while the performance on depth, infrared

Table 7. Recognition Accuracy (%) for Different Cross-view Splits with RGB and Skeletons from our PKU-MMD Dataset

| Dataset | Part I | | | | | | Part II | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Train/Test Split | #1#2/#3 | | #1#3/#2 | | #2#3/#1 | | #1#2/#3 | | #1#3/#2 | | #2#3/#1 | |
| Methods | R | S | R | S | R | S | R | S | R | S | R | S |
| TSN [69] | 81.2 | – | 84.1 | – | 82.5 | – | 56.3 | – | 46.0 | – | 50.4 | – |
| STA-LSTM [53] | – | 86.6 | – | 90.8 | – | 83.3 | – | 28.6 | – | 48.8 | – | **33.7** |
| TPN [19] | – | 86.4 | – | 93.7 | – | 86.2 | – | 29.7 | – | **49.5** | – | **33.7** |
| VA-LSTM [82] | – | 81.4 | – | 92.5 | – | 82.7 | – | 34.5 | – | 40.3 | – | 29.7 |
| ST-GCN [75] | – | **87.9** | – | 92.0 | – | **88.1** | – | **35.8** | – | 30.8 | – | 30.2 |
| LSTM | 86.1 | 83.9 | **88.6** | 94.0 | 86.6 | 85.6 | 59.3 | 28.1 | 51.6 | 40.3 | 52.7 | 29.0 |
| BLSTM | **89.0** | 86.4 | **88.6** | **94.6** | **90.2** | 86.5 | **63.8** | 26.3 | **52.6** | 46.5 | **53.5** | 30.7 |

images, and skeletons are higher than that on the NTU dataset. It indicates that the difficulty in action recognition on Part I is comparable with NTU, but we have depth, infrared images, and skeletons in higher quality, which can benefit the exploration in 3D reconstruction or other related topics. However, the action recognition performance with Part II is much inferior to that on NTU, which present new challenges in the task of action analysis. Besides, we also notice that, different from PKU-MMD, LSTM and BLSTM achieve comparable performance on NTU, which is mainly caused by shorter video length (about 300 frames/video) in NTU.

**Multi-Modal Recognition:** With the results from BLSTM for each modality, we combine the multi-modal data by average fusion to utilize complementary information. That is, the probability $p^*$ for being the $c$th class from video $\mathbf{V}$ can be formulated as

$$p^*(c|\mathbf{V}) = \frac{1}{|\Omega|} \sum_{m \in \Omega} p^m(c|\mathbf{V}^m), \tag{4}$$

where the superscript $m$ indicates which modality the score is from, $\Omega$ is the set consisting of the modalities taken into account, and $\mathbf{V}^m$ denotes the video representation of the corresponding modality of video $\mathbf{V}$. The results can be seen in Table 8. Compared with single-modal recognition, each modality is able to contribute to improve the recognition performance.

### 4.2 Action Detection

*4.2.1 Detection Methods.* In action detection, we aim to not only recognize but also localize the actions in the untrimmed video sequence. Here we introduce several approaches for action detection. As action recognition, the features for RGB, optical flow, depth and IR data are extracted from the *global-pool* layer of BN-Inception [69].

**Sliding Window-based Methods (SW-X):** Action detection can be achieved through recognizing and integrating sliding windows. The classifier is independent to the sliding window scheme. In our experiments, we use STA-LSTM [53], TPN [19], LSTM, and BLSTM to classify the sliding windows. The configurations of LSTM/BLSTM are the same as Table 5. For the sliding window-based approaches, we utilize temporal downsampling with a stride as 5. We then split the long sequences into action windows with the size as 10. Each window is recognized with different classifiers. Adjacent windows that share the same action label are linked to get the detection results. Note that STA-LSTM [53], TPN [19] are only employed for skeleton data.

**Joint Classification Regression RNN (JCRRNN):** Li et al. [33] proposed a Joint Classification Regression RNN that implements frame-level real-time action detection. Though the network

Table 8. Recognition Accuracy (%) (BLSTM) with Multi-modal Fusion

| Modalities | Part I | | Part II | |
|---|---|---|---|---|
| | CS | CV | CS | CV |
| R + OF | 91.5 | 95.1 | 74.2 | **74.8** |
| R + D | 88.4 | 90.8 | 65.1 | 61.9 |
| R + IR | 85.2 | 89.2 | 66.0 | 63.0 |
| R + S | 90.2 | 96.2 | 63.6 | 61.3 |
| OF + S | 92.9 | 97.1 | 71.3 | 72.9 |
| OF + IR | 90.5 | 93.2 | 72.7 | 70.9 |
| OF + D | 92.5 | 94.2 | 73.6 | 71.5 |
| S + D | 91.2 | 95.8 | 58.0 | 49.2 |
| S + IR | 89.1 | 95.5 | 58.0 | 47.9 |
| IR + D | 87.1 | 86.9 | 61.9 | 53.0 |
| R + OF + D | 92.3 | 95.0 | 73.0 | 74.1 |
| R + OF + IR | 90.9 | 94.7 | 73.2 | 73.7 |
| R + OF + S | 93.3 | 97.3 | 73.5 | 73.2 |
| R + OF + D + S | **94.4** | **97.5** | 73.1 | 72.7 |
| R + OF + IR + S | 93.0 | 96.9 | **74.5** | 72.5 |
| R + OF + D + IR + S | 93.7 | 96.8 | 73.3 | 70.7 |

is designed for skeleton data, we can also feed the deep features of other modalities to get the detection results.

**Untrimmed Net:** Wang et al. [68] developed a joint action recognition and detection framework. The model can be optimized in an end-to-end manner and has shown superior performance on famous THUMOS14 [64] and ActivityNet [3].

In Table 9, we also include the results from papers [29–31, 37, 60] that cite our conference version [36] and use Part I to evaluate their methods. The results are taken from the corresponding papers directly. Note that not all modalities under all settings are taken into account in these methods.

*4.2.2 PKU-MMD Detection Benchmarks.* In the detection task, we evaluate each method of single-modal detection and multi-modal detection on Part I and Part II, respectively.

**Single-Modal Detection:** Table 9 and Table 10 show detection results in terms of mAP on Part I and Part II, respectively. The results from the two subsets are quite different. Compared with mAP results of sliding window-based methods on Part I (Table 9), we obtain much lower mAP on Part II (Table 10) due to poor recognition performance for each window. Meanwhile, the concurrent actions in Part II can cause missing detections easily (Figure 6(a)). And short intervals lead to many false positives, especially at the connections between adjacent actions (Figure 6(b)), which further reduces the precision of action proposals and then mAP results. Though JCRRNN [33] outperforms SW-BLSTM greatly in Part I, SW-BLSTM achieves higher results than JCRRNN in Part II. It is mainly because JCRRNN fails to regress the start and end points of an action when action boundaries are less clear. Besides, the performance with UntrimmedNet [68] is also far from being satisfactory, since UntrimmedNet is based on frame-level action classification, the occlusion and concurrent actions bring challenges to get a well-trained classifier. Overall, the results in Table 10 illustrate that Part II constitutes a great challenge to the state of the art and is more difficult than Part I for action detection.

**Multi-Modal Detection:** We evaluate the capability of detecting actions in the multi-modal scenario with SW-BLSTM. We try different combination of different modalities, and the mAP

Table 9. MAP Results (%) of Different Modalities Using Different Methods on Part I

| Partition Setting | | Cross-subject (Part I) | | | | | Cross-view (Part I) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | $\theta$ | R | OF | D | IR | S | R | OF | D | IR | S |
| SW-STA-LSTM [53] | 0.1 | – | – | – | – | 47.5 | – | – | – | – | 48.0 |
| | 0.5 | – | – | – | – | 25.4 | – | – | – | – | 27.8 |
| SW-TPN [19] | 0.1 | – | – | – | – | 60.2 | – | – | – | – | 71.0 |
| | 0.5 | – | – | – | – | 30.4 | – | – | – | – | 40.0 |
| SW-LSTM | 0.1 | 59.7 | 66.6 | 65.2 | 53.2 | 68.4 | 63.0 | 64.3 | 58.0 | 48.2 | 76.4 |
| | 0.5 | 30.4 | 25.4 | 34.8 | 25.6 | 38.2 | 30.5 | 20.1 | 26.3 | 19.5 | 44.9 |
| SW-BLSTM | 0.1 | 62.1 | 66.7 | 65.8 | 54.2 | 69.0 | 65.3 | 63.6 | 58.9 | 49.4 | 77.5 |
| | 0.5 | 33.3 | 24.4 | 35.8 | 25.2 | 36.3 | 31.4 | 22.8 | 26.6 | 19.6 | 44.2 |
| JCRRNN [33] | 0.1 | 71.5 | 81.6 | 73.4 | **61.2** | 52.2 | **76.9** | **87.3** | **74.0** | **57.9** | 53.9 |
| | 0.5 | 53.8 | 66.8 | 57.9 | **42.8** | 35.5 | **61.5** | **74.2** | **57.6** | **40.2** | 38.0 |
| CNN+Motion+Trans [30] | 0.1 | – | – | – | – | **92.2** | – | – | – | – | **95.8** |
| | 0.5 | – | – | – | – | 90.4 | – | – | – | – | 93.7 |
| Trans RNN [60] | 0.1 | – | – | – | – | 84.2 | – | – | – | – | 93.5 |
| | 0.5 | – | – | – | – | 74.3 | – | – | – | – | 86.7 |
| Skeleton Boxes [29] | 0.1 | – | – | – | – | 61.3 | – | – | – | – | 94.5 |
| | 0.5 | – | – | – | – | 54.8 | – | – | – | – | 94.2 |
| HCN [31] | 0.5 | – | – | – | – | **92.6** | – | – | – | – | **94.2** |
| Graph Distillation [37] | 0.1 | **88.0** | **82.6** | **87.2** | – | 85.7 | – | – | – | – | – |
| | 0.5 | **80.1** | **74.7** | **79.2** | – | 78.4 | – | – | – | – | – |

Results of References [29–31, 37, 60] are from their papers.

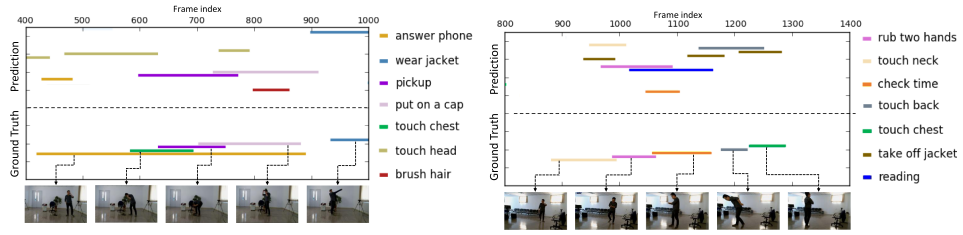
Table 10. MAP Results (%) of Different Modalities Using Different Methods on Part II

| Partition Setting | | Cross-subject (Part II) | | | | | Cross-view (Part II) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | $\theta$ | R | OF | D | IR | S | R | OF | D | IR | S |
| SW-STA-LSTM [53] | 0.1 | – | – | – | – | 5.7 | – | – | – | – | 4.7 |
| | 0.5 | – | – | – | – | 2.2 | – | – | – | – | 2.2 |
| SW-TPN [19] | 0.1 | – | – | – | – | 11.7 | – | – | – | – | 4.2 |
| | 0.5 | – | – | – | – | 3.3 | – | – | – | – | 1.5 |
| SW-LSTM | 0.1 | 20.1 | **31.0** | 10.8 | 14.8 | **11.2** | 20.3 | 27.4 | 15.9 | 13.5 | 4.2 |
| | 0.5 | 6.8 | 13.7 | 2.9 | 4.6 | 3.2 | **6.9** | 10.0 | 5.2 | 4.2 | 1.5 |
| SW-BLSTM | 0.1 | **23.2** | 30.9 | **18.7** | **19.5** | 7.0 | **21.3** | **28.2** | **17.4** | **16.4** | **7.5** |
| | 0.5 | **8.3** | **11.9** | **5.8** | **6.3** | **3.7** | 6.8 | **10.3** | **5.5** | **5.1** | **4.0** |
| JCRRNN [33] | 0.1 | 14.3 | 17.3 | 8.6 | 9.8 | 2.3 | 11.7 | 18.8 | 9.8 | 6.4 | 1.0 |
| | 0.5 | 5.9 | 8.6 | 3.0 | 3.7 | 0.5 | 4.6 | 10.4 | 3.2 | 1.6 | 0.1 |
| UntrimmedNet [68] | 0.1 | 7.1 | 5.9 | 5.3 | 4.7 | – | 7.7 | 5.1 | 5.4 | 5.6 | – |
| | 0.5 | 1.9 | 2.7 | 1.6 | 1.3 | – | 1.7 | 2.5 | 1.8 | 1.6 | – |

results under IoU $\theta = 0.5$ are shown in Table 11. The introduction of additional modalities effectively improves mAP results, compared to single-modal detection.

### 4.3 Cross-dataset Study

In this subsection, we perform a cross-dataset study to show that PKU-MMD can be utilized to support other datasets on action analytics. The experiments are conducted with skeleton data




(a) Concurrent actions cause missing detection (*i.e.*, *touching chest*).

(b) Short action intervals cause false positives.

Fig. 6. Visualization of action detection results on Part II (BLSTM with RGB videos under the cross-subject split).

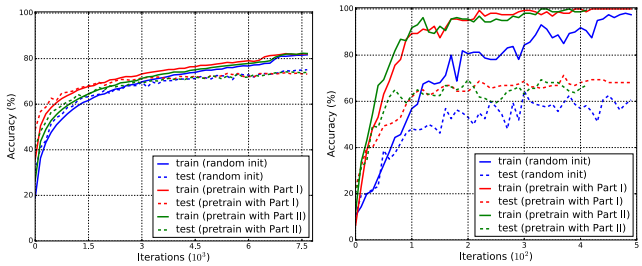
Table 11. Fusion Results (SW-BLSTM) for Action Detection in mAP ($\theta = 0.5$) (%)

| Modalities | Part I | | Part II | |
|---|---|---|---|---|
| | CS | CV | CS | CV |
| R + OF | 38.0 | 38.9 | 18.6 | 21.7 |
| R + D | 42.0 | 35.0 | 11.1 | 13.4 |
| R + IR | 35.6 | 31.6 | 12.6 | 12.4 |
| R + S | 45.0 | 48.6 | 13.1 | 13.9 |
| OF + S | 41.4 | 49.5 | **19.3** | **24.6** |
| OF + IR | 33.2 | 30.4 | 16.2 | 16.9 |
| OF + D | 39.4 | 34.2 | 14.6 | 18.5 |
| S + D | **60.7** | 45.1 | 9.1 | 8.6 |
| S + IR | 39.1 | 40.8 | 9.6 | 8.1 |
| IR + D | 36.3 | 28.3 | 10.0 | 9.6 |
| R + OF + D | 45.8 | 42.2 | 18.0 | 21.3 |
| R + OF + IR | 40.8 | 40.0 | 18.6 | 20.7 |
| R + OF + S | 48.3 | **54.1** | 18.4 | 22.0 |
| R + OF + D + S | 51.6 | 53.1 | 18.2 | 21.2 |
| R + OF + IR + S | 48.1 | 51.4 | 19.1 | 20.9 |
| R + OF + D + IR + S | 49.6 | 49.8 | 16.5 | 19.0 |

for action recognition. We pretrain LSTM configured as in Table 5 with PKU-MMD Part I and Part II and then fine-tune the network on the large NTU dataset [49] and small MSR Daily Activity dataset [61], respectively. We follow Reference [49] to split the NTU dataset with cross-subject and cross-view protocols and follow Reference [61] to use samples from the first five subjects as training and the rest as testing for the MSR dataset. The results are given in Table 12. The results on the NTU dataset with pretraining are comparable to training from scratch with random initialization, but pretraining speeds up the convergence early in training, as shown in Figure 7. It is consistent with the conclusion in Reference [16] that training from scratch can be comparable with pretraining counterparts when there are sufficient training data. On the small-scale MSR dataset, however, PKU-MMD pretraining effectively improves action recognition performance. We observe that pretraining with Part I brings more improvement than Part II. It is mainly because the larger number of samples in Part I leads to a more generalizable model. The results in Table 12 and Figure 7 illustrate that the features learned on PKU-MMD can be well transferred to the MSR dataset.

Table 12. Action Recognition Accuracy (%) on Testing Data of NTU and MSR Datasets

| Acc. (%) | NTU (CS) | NTU (CV) | MSR |
|---|---|---|---|
| random init | 71.9 | **82.0** | 68.1 |
| pretrain with Part I | **72.1** | 81.6 | **72.5** |
| pretrain with Part II | 71.9 | 81.9 | 69.4 |




(a) NTU (CV) (b) MSR

Fig. 7. Learning curves of action recognition accuracy on the training and testing sets, respectively.

Therefore, PKU-MMD can help speed up convergence for the large-scale dataset and compensate for the lack of training data when the dataset is small.

### 4.4 Discussions

We provide further discussions based on the experiment results to address multi-modal action analytics.

**What are the characteristics for each modal data?** We investigate this question through the analysis from Figure 8, which shows the average precision in action recognition with BLSTM (Part I) on each action class for each modality. We find that optical flow usually dominate the recognition performance and get the highest accuracy on most of the action categories, which is consistent with the results in Table 6. It is mainly because the optical flow provides pixel-level human motions, increasing the discriminations of action. With appearance information, RGB is good at distinguishing object-related actions, such as *reading*. Skeletons are able to recognize actions involved obvious human motion, such as *bowing*. Another advantage of skeleton data is that their training time is much less than other modalities due to the low dimension. Depth data are able to well recognize most human actions in our PKU-MMD dataset. An interesting point is that the performance with depth information sometimes can be even better than that with RGB data under the cross subject setting, which can be observed from Table 6 and Table 9. It is probably because the appearance information is hidden by depth information, leading to a more unified distribution between training and testing data, and then the model can well fit the testing samples. For IR images, however, we found that the data noises lead to degradation in performance compared with other modalities, indicating that the exploitation on action analytics from IR information is in demand.

**More modalities, better performance?** In general, the more modalities are involved, the better performance we obtain, since complementary information can be utilized to compensate missing features in the single modal data. However, with simply an average fusion scheme, we found that the introduction of IR always reduces the performance. This is probably because IR images do

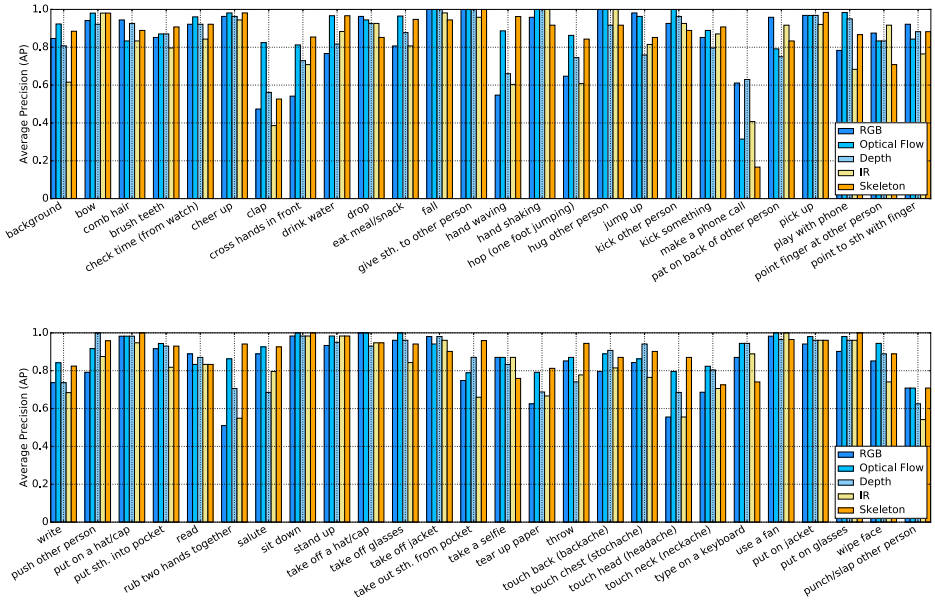



Fig. 8. Average precision of BLSTM on each action class for each modality in action recognition (Part I, cross-subject).

not contain additional information to the combination of RGB and depth in the daylight environment. And employing all the available modalities does not necessarily give the most satisfactory results, illustrating that it is challenging to fully explore the complementary information of different modalities. Nevertheless, we still believe there is much potential to boost the performance of action analytics with multi-modal data.

## 5 CONCLUSION

In this article, we release a large-scale multi-modal benchmark (PKU-MMD) for human action analytics. To make our dataset more practical, we record two subsets under different settings and thus with different levels of difficulty in action understanding. Compared with easy Part I, Part II is more challenging due to short action intervals, concurrent actions, and heavy occlusion. We introduce two applications for PKU-MMD: trimmed action recognition and untrimmed action detection on both subsets, respectively. To give a comparison study, we review several existing methods proposed for action analytics. Extensive experiments are conducted to evaluate each method on our benchmark. We further analyze the performance from different modalities as well as their fusion results. Our studies show that the multi-modal action analytics is far from mature compared with RGB-based action analytics. And our dataset brings new challenges to state-of-the-art methods. We hope our benchmark facilitate further research and serve the community in the field of human action analytics.

## REFERENCES

[1] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos, and Victor C. M. Leung. 2013. Non-intrusive human activity monitoring in a smart home environment. In *Proceedings of the IEEE International Conference on E-health Networking, Application & Services (Healthcom'13)*. 606–610.

[2] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. 2012. G3D: A gaming action dataset and real time action recognition evaluation framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 7–12.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[4] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao. 2017. RGB-D datasets using microsoft kinect or similar sensors: A survey. *Multimedia Tools Appl.* 76, 3 (2017), 4313–4355.

[5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4724–4733.

[6] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proceedings of the IEEE International Conference on Image Processing*. 168–172.

[7] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. 2012. Human daily action analysis with multi-view and color-depth data. In *Proceedings of the European Conference on Computer Vision*. 52–61.

[8] CMU. 2003. CMU Graphics Lab Motion Capture Database. Retrieved from http://mocap.cs.cmu.edu/.

[9] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. Online action detection. In *Proceedings of the European Conference on Computer Vision*. 269–284.

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2625–2634.

[11] Yong Du, Yun Fu, and Liang Wang. 2016. Representation learning of temporal dynamics for skeleton-based action recognition. *IEEE Trans. Image Process.* 25, 7 (2016), 3010–3022.

[12] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. DAPs: Deep action proposals for action understanding. In *Proceedings of the European Conference on Computer Vision*. 768–784.

[13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* 88, 2 (2010), 303–338.

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1933–1941.

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

[16] Kaiming He, Ross B. Girshick, and Piotr Dollár. 2019. Rethinking imagenet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*. 4918–4927.

[17] Minh Hoai and Fernando De la Torre. 2014. Max-margin early event detectors. *Int. J. Comput. Vis.* 107, 2 (2014), 191–202.

[18] J. F. Hu, W. S. Zheng, J. H. Lai, and J Zhang. 2016. Jointly learning heterogeneous features for RGB-D activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 11 (2016), 2186–2200.

[19] Yueyu Hu, Chunhui Liu, Yanghao Li, Sijie Song, and Jiaying Liu. 2017. Temporal perceptive network for skeleton-based action recognition. In *Proceedings of the British Machine Vision Conference*. 1–12.

[20] Min Huang, Song-Zhi Su, Hong-Bo Zhang, Guo-Rong Cai, Dongying Gong, Donglin Cao, and Shao-Zi Li. 2018. Multifeature selection for 3D human action recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 2 (2018), 45.

[21] Arpit Jain, Abhinav Gupta, Mikel Rodriguez, and Larry S. Davis. 2013. Representing videos using mid-level discriminative patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2571–2578.

[22] Manan Jain, Jan Van Gemert, Hervé Jégou, Patrick Bouthemy, and Cees G. M. Snoek. 2014. Action localization with tubelets from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 740–747.

[23] Feng Jiang, Shengping Zhang, Shen Wu, Yang Gao, and Debin Zhao. 2015. Multi-layered gesture recognition with kinect. *J. Mach. Learn. Res.* 16, 1 (2015), 227–254.

[24] Yu-Gang Jiang, Qi Dai, Wei Liu, Xiangyang Xue, and Chong-Wah Ngo. 2015. Human action recognition in unconstrained videos by explicit motion modeling. *IEEE Trans. Image Process.* 24, 11 (2015), 3781–3795.

[25] Zhuolin Jiang, Viktor Rozgic, and Sancar Adali. 2017. Learning spatiotemporal features for infrared action recognition with 3D convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 115–123.

[26] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1725–1732.

[27] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 2556–2563.

[28] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[29] Bo Li, Huahui Chen, Yucheng Chen, Yuchao Dai, and Mingyi He. 2017. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In *Proceedings of the IEEE International Conference on Multimedia Expo Workshops*. 613–616.

[30] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2017. Skeleton-based action recognition with convolutional neural networks. In *Proceedings of the IEEE International Conference on Multimedia Expo Workshops*. 597–600.

[31] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 786–792.

[32] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. 2010. Action recognition based on a bag of 3D points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9–14.

[33] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. 2016. Online human action detection using joint classification-regression recurrent neural networks. In *Proceedings of the European Conference on Computer Vision*. 203–220.

[34] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. 2014. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 812–819.

[35] An-An Liu, Wei-Zhi Nie, Yu-Ting Su, Li Ma, Tong Hao, and Zhao-Xuan Yang. 2015. Coupled hidden conditional random fields for RGB-D human action recognition. *Signal Processing* 112 (2015), 74–82.

[36] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. 2017. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*. 1–8.

[37] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. 2018. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision*. 174–192.

[38] Behrooz Mahasseni and Sinisa Todorovic. 2016. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3054–3062.

[39] Jie Miao, Xiangmin Xu, Shuoyang Qiu, Chunmei Qing, and Dacheng Tao. 2015. Temporal variance analysis for action recognition. *IEEE Trans. Image Process.* 24, 12 (2015), 5904–5915.

[40] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. 2007. *Documentation Mocap Database HDM05*. Technical Report CG-2007-2.

[41] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley MHAD: A comprehensive multimodal human action database. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 53–60.

[42] Omar Oreifej and Zicheng Liu. 2013. Hon4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[43] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision*. 143–156.

[44] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. 2016. Histogram of oriented principal components for cross-view action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 12 (2016), 2430–2443.

[45] Hossein Rahmani, Arif Mahmood, Du Q. Huynh, and Ajmal Mian. 2014. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In *Proceedings of the European Conference on Computer Vision*. 742–757.

[46] Michalis Raptis, Iasonas Kokkinos, and Stefano Soatto. 2012. Discovering discriminative action parts from mid-level video representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1242–1249.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252.

[48] Michael S. Ryoo. 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 1036–1043.

[49] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1010–1019.

[50] Amr Sharaf, Marwan Torki, Mohamed E. Hussein, and Motaz El-Saban. 2015. Real-time multi-scale action detection from 3D skeleton data. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 998–1005.

[51] Zhiyuan Shi and Tae-Kyun Kim. 2017. Learning and refining of privileged information-based RNNs for action recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3461–3470.

[52] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the Advances in Neural Information Processing Systems*. 568–576.

[53] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4263–4270.

[54] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *Arxiv Preprint Arxiv:1212.0402* (2012).

[55] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2011. Human activity detection from RGBD images. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 47–55.

[56] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2012. Unstructured human activity detection from RGBD images. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 842–849.

[57] Yicong Tian, Rahul Sukthankar, and Mubarak Shah. 2013. Spatiotemporal deformable part models for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2642–2649.

[58] Raviteja Vemulapalli and Rama Chellapa. 2016. Rolling rotations for recognizing human actions from 3D skeletal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4471–4479.

[59] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*. 3551–3558.

[60] Hongsong Wang and Liang Wang. 2017. Learning robust representations using recurrent neural networks for skeleton based action classification and detection. In *Proceedings of the IEEE International Conference on Multimedia Expo Workshops*. 591–596.

[61] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1290–1297.

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

[63] Keze Wang, Xiaolong Wang, Liang Lin, Meng Wang, and Wangmeng Zuo. 2014. 3D human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the ACM International Conference on Multimedia*. 97–106.

[64] Limin Wang. 2014. Action recognition and detection by combining motion and appearance features. *THUMOS* (2014).

[65] LiMin Wang, Yu Qiao, and Xiaoou Tang. 2013. Motionlets: Mid-level 3d parts for human motion recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2674–2681.

[66] Limin Wang, Yu Qiao, and Xiaoou Tang. 2014. Video action detection with relational dynamic-poselets. In *Proceedings of the European Conference on Computer Vision*. 565–580.

[67] Limin Wang, Zhe Wang, Yuanjun Xiong, and Yu Qiao. 2015. CUHK&SIAT submission for THUMOS15 action recognition challenge. *THUMOS* (2015).

[68] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4325–4334.

[69] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*. 20–36.

[70] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. 2013. Modeling 4D human-object interactions for event and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 3272–3279.

[71] Ping Wei, Nanning Zheng, Yibiao Zhao, and Song-Chun Zhu. 2013. Concurrent action detection with structural prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3136–3143.

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

[73] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the ACM International Conference on Multimedia*. 461–470.

[74] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. 2012. View invariant human action recognition using histograms of 3D joints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 20–27.

[75] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7444–7452.

[76] Xiaodong Yang and YingLi Tian. 2014. Super normal vector for activity recognition using depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 804–811.

[77] Jun Ye, Hao Hu, Guo-Jun Qi, and Kien A. Hua. 2017. A temporal order modeling approach to human action recognition from multimodal sensor data. *ACM Trans. Multimedia Comput. Commun. Appl.* 13, 2 (2017), 14.

[78] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. 2012. Two-person interaction detection using body-pose features and multiple instance learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 28–35.

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

[80] Jing Zhang, Wanqing Li, Philip O. Ogunbona, Pichao Wang, and Chang Tang. 2016. RGB-D-based action recognition datasets: A survey. *Pattern Recogn.* 60, 1 (2016), 86–105.

[81] Lei Zhang, Shengping Zhang, Feng Jiang, Yuankai Qi, Jun Zhang, Yuliang Guo, and Huiyu Zhou. 2017. BoMW: Bag of manifold words for one-shot learning gesture recognition from kinect. *IEEE Trans. Circ. Syst. Vid. Technol.* 28, 10 (2017), 2562–2573.

[82] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*. 2117–2126.

[83] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Dahua Lin, and Xiaoou Tang. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2914–2923.

[84] Jun Zhu, Baoyuan Wang, Xiaokang Yang, Wenjun Zhang, and Zhuowen Tu. 2013. Action recognition with actions. In *Proceedings of the IEEE International Conference on Computer Vision*. 3559–3566.

[85] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3697–3703.

[86] Mohammadreza Zolfaghari, Gabriel L. Oliveira, Nima Sedaghat, and Thomas Brox. 2017. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2923–2932.

Received May 2019; revised August 2019; accepted September 2019