PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding

Chunhui Liu
Institute of Computer Science and Technology, Peking University
Beijing, China 100080
liuchunhui@pku.edu.cn

Yueyu Hu
Institute of Computer Science and Technology, Peking University
Beijing, China 100080
hyy@pku.edu.cn

Yanghao Li
Institute of Computer Science and Technology, Peking University
Beijing, China 100080
lyttonhao@pku.edu.cn

Sijie Song
Institute of Computer Science and Technology, Peking University
Beijing, China 100080
ssj940929@pku.edu.cn

Jiaying Liu[*] [†]
Institute of Computer Science and Technology, Peking University
Beijing, China 100080
liujiaying@pku.edu.cn

ABSTRACT

Despite the fact that many 3D human activity benchmarks being proposed, most existing action datasets focus on the action recognition tasks for the segmented videos. There is a lack of standard large-scale benchmarks, especially for current popular data-hungry deep learning based methods. In this paper, we introduce a new large scale benchmark (PKU-MMD) for continuous skeleton-based human action understanding and cover a wide range of complex human activities with well annotated information. PKU-MMD contains 1076 long video sequences in 51 action categories, performed by 66 subjects in three camera views. It contains almost 20,000 action instances and 5.4 million frames in total. Our dataset also provides multi-modality data sources, including RGB, depth, Infrared Radiation and Skeleton. To the best of our knowledge, it is the largest skeleton-based detection database so far. We conduct extensive experiments and evaluate different methods on this dataset. We believe this large-scale dataset will benefit future researches on action detection for the community.

KEYWORDS

Video Analysis; Action Detection; Skeleton-Based Action Understanding; Video Benchmark

1 INTRODUCTION

The tremendous success of deep learning have made data-driven learning methods get ahead with surprisingly superior performance for many computer vision tasks.

This methods all need a large scale dataset for prior knowledge and feature learning. Activity understanding which contains several tasks like action recognition and action detection is still challenging. For RGB dataset, several famous large scale datasets have been collected to boost the research in this area [3, 22]. ActivityNet [3] is a superior RGB video dataset gathered from Internet media like YouTube with well annotated label and boundaries. Nevertheless, 3D action dataset is hard to obtain due to the lack of well annotated activities in 3D modal on the Internet.

Thanks to the prevalence of the affordable color-depth sensing cameras like Microsoft Kinect, and the capability to obtain depth data and the 3D skeleton of human body on the fly, 3D activity analysis has drawn great attentions. As an intrinsic high level representation, 3D skeleton is valuable and comprehensive for summarizing a series of human dynamics in the video, and thus benefits the more general action analysis. Besides succinctness and effectiveness, it has a significant advantage of great robustness to illumination, clustered background, and camera motion. However, as a kind of popular data modality, 3D action analysis suffers from the lack of large-scale benchmark datasets. To the best of our knowledge, existing 3D action benchmarks have limitations in two aspects.

• **Shortage in large action detection datasets:** Action detection plays an important role in video analytics and can be effectively studied through analysis and learning from massive samples. However, most existing skeleton datasets mainly target at the task of action recognition for segmented videos. There is a lack of large scale skeleton dataset for action detection. Additionally, previous detection benchmarks only contain a small number of actions in each video even in some large scale RGB datasets [3]. There is no doubt that more actions within one untrimmed video will promote the robustness of action detection algorithms based on the sequential action modeling and featuring.

• **Limited variations:** Existing models also suffer over-fitting problems due to limited action categories and sample variations. On the one hand, more action categories lead to ambiguity in some actions (*e.g.* drinking *vs.* eating), making the dataset more challenging and more consistent with real life. On the other hand, most datasets are collected under solo view with actors facing to the

[*] Corresponding author

[†] This dataset and work is funded by Microsoft Research Asia (project ID FY17-RES-THEME-013) and CCF-Tencent Open Research Fund.

VSCC'17, October 23, 2017, Mountain View, CA, USA.
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ISBN 978-1-4503-5506-3/17/10…$15.00
DOI: https://doi.org/10.1145/3132734.3132739

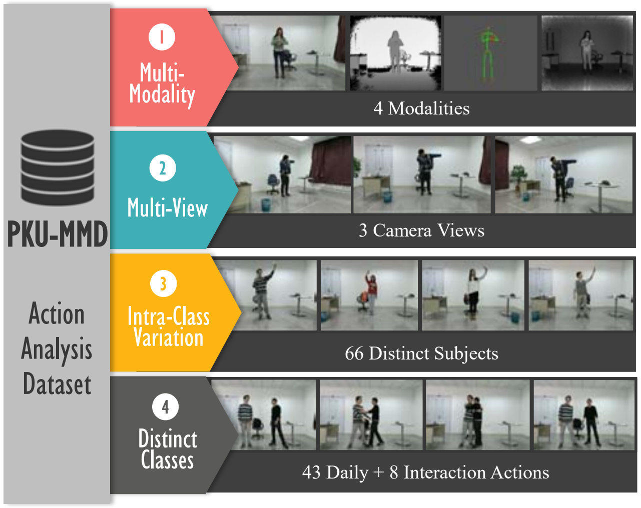


Figure 1: PKU Multi-Modalilty Dataset is a large-scale multi-modalities action detection dataset. This dataset contains 51 action categories, performed by 66 distinct subjects in 3 camera views.

camera, resulting in limited methods for multi-view analysis. The varieties of subjects and camera views will lead to more intra-class variation.

To overcome these limitations, we develop a new large scale continuous multi-modality 3D human activity dataset (PKU-MMD)[1] for facilitating further study on human activity understanding, especially action detection. As shown in Figure 1, our dataset contains 1076 videos composed by 51 action categories, and each of the video contains more than twenty action instances performed by 66 subjects in 3 camera views. The total number of our dataset is 3,000 minutes and 5,400,000 frames. Although this paper focus on skeleton-based action detection, we provide four raw modalities: RGB frame, depth map, skeleton data, and infrared for further study. More modalities can be further calculated such as optical flow and motion vector.

Besides, we propose a new 2D protocol to evaluate the precision-recall curve of each method in a much straightforward manner. Taking over-lapping ratio and detection confidence into account jointly, each algorithm can be evaluated with a single value, instead of a list of mean average precisions with corresponding overlap ratios. Several experiments are implemented to test both the capabilities of different approaches for action detection and the combination performance of different modalities.

We organize the paper as follows. We first review the development of action understanding and summarize existing benchmarks for 3D activity analysis in Sec. 2. In Sec. 3, we present the details of our dataset with collection and annotation details. Experimental details about proposed protocols and experimental results are shown in Sec. 4. Concluding remarks are given in Sec. 5.

2 RELATED WORK

In this section, we briefly summarize the development of activity analysis. As a part of pattern recognition, activity analysis shows a common way of development in machine learning, where large scale benchmarks share familiar significance with magnificent methods. Here, we briefly introduce a series of benchmarks and approaches. For a more extensive conclusion of activity analysis we refer to corresponding survey papers [1, 4, 5, 48].

2.1 Development of Activity Analysis

Early activity analysis mainly focuses on action recognition which consists of a classification task for segmented videos. Traditional methods mainly focus on hand-crafting features for 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) encoded by Fisher Vector [20, 36] achieved a good performance. Recently, deep learning has been exploited for action recognition [27, 41]. Deep approaches automatically learn robust feature representations directly from raw data and recognize actions synchronously with deep neural networks [32]. To model temporal dynamics, Recurrent Neural Network (RNN) have also been exploited for action recognition. In [8, 45], CNN layers are constructed to extract visual features while the followed recurrent layers are applied to handle temporal dynamics.

For action detection, existing methods mainly utilize either sliding-window scheme [26, 39], or action proposal approaches [40]. These methods usually have low computational efficiency or unsatisfactory localization accuracy due to the overlapping design and unsupervised localization approach. Most methods are designed for offline action detection [26, 33, 43]. However, in many new works, recognizing the actions on the fly before the completion of the action is well studied by a learning formulation based on a structural SVM [11], or a non-parametric moving pose framework [47] and a dynamic integral bag-of-words approach [23]. LSTM is also used for online action detection and forecast which provides frame-wise class information. It forecasts the occurrence of start and end of actions.

As the fundamental requirement of research, videos source also determines the branches of action analysis. Early action analysis dataset mainly focuses on home surveillance activities like drinking or waving hands. The analysis of those simple indoor activities are the start of action recognition process. The advantages of this kind of videos lie in that they are usually easy and cheap to capture. However, collecting a large scale benchmark with cameras can be troublesome. Fortunately, the rapid development of Internet technology and data mining algorithms enable a new approach of collecting dataset from Internet third-way media like YouTube [3, 29]. As a result, RGB-based datasets achieve a grant level with hundreds of action labels and video sources in TB level. Recently, there are also several works focus on collecting different datasets of action type like TV-series [7], Movies [14] and Olympic Games [13].

With the launch of Microsoft Kinect, the diversity of action source becomes possible. Different input sources have been discussed such as Depth data and Skeleton data. Depth data provides a 3D information which is beneficial for action understanding. Skeleton, as a kind of high level representation of human body, can provide valuable and condensed information for recognizing actions. As Kinect devices provide a real-time algorithm to generate

[1] http://www.icst.pku.edu.cn/struct/Projects/PKUMMD.html

Table 1: A comparison among different skeleton-based detection datasets.

Datasets	Classes	Videos	Labeled Instances	Actions per Video	Modalities	Temporal Localization	Year
MSR-Action3D [15]	20	567	567	1	D+Skeleton	No	2010
RGBD-HuDaAct [18]	13	1189	1189	1	RGB+D	No	2011
MSR-DailyActivity [37]	16	320	320	1	RGB+D+Skeleton	No	2012
Act4 [6]	14	6844	6844	1	RGB+D	No	2012
MHAD [19]	11	660	660	1	RGB+D+Skeleton	No	2013
Multiview 3D Event [42]	8	3815	3815	1	RGB+D+Skeleton	No	2013
Northwestern-UCLA [38]	10	1475	1475	1	RGB+D+Skeleton	No	2014
UWA3D Multiview II [21]	30	1075	1075	1	RGB+D+Skeleton	No	2015
NTU RGB+D [24]	60	56880	56880	1	RGB+D+IR+Skeleton	No	2016
G3D [2]	20	210	1467	7	RGB+D+Skeleton	Yes	2012
SBU Kinect interaction [46]	8	21	300	14.3	RGB+D+Skeleton	Yes	2012
CAD-120 [31]	20	120	~1200	~ 8.2	RGB+D+Skeleton	Yes	2013
compostable Activities [17]	16	693	2529	3.6	RGB+D+Skeleton	Yes	2014
Watch-n-Patch [44]	21	458	~2500	2~7	RGB+D+Skeleton	Yes	2015
OAD [16]	10	59	~700	~12	RGB+D+Skeleton	Yes	2016
PKU-MMD	**51**	**1076**	**21545**	**20.02**	**RGB+D+IR+Skeleton**	**Yes**	**2017**

skeleton data from the information of RBG, depth, and infrared, skeleton becomes an ideal source to support real-time algorithm and to be transferred and utilized on some mobile devices like robots or telephones.

Despite of the diversity of source, action understanding still faces several problems, among which the top priority is the accuracy problem. Another problem is the poor performance of cross-data recognition. That is, existing approaches or machine learning models achieve good performances with training and test sets in similar environments conditions. Open domain action recognition and detection is still challenging.

2.2 3D Activity Understanding Approaches

For skeleton-based action recognition, many generative models have been proposed with superior performance. Those methods are designed to capture local features from the sequences and then to classify them by traditional classifiers like Support Vector Machine (SVM). Those local features includes rotations and translations to represent geometric relationships of body parts in a Lie group [34, 35], or the covariance matrix to learn the co-occurrence of skeleton points [12]. Additionally, Fourier Temporal Pyramids (FTP) or Dynamic Time Warping (DTW) are also employed to temporally align the sequences and to model temporal dynamics. Furthermore, many methods [10, 25] divide the human body into several parts and learn the co-occurrence information, respectively. A Moving Pose descriptor [47] is proposed to mine key frames temporally via a k-NN approach in both pose and atomic motion features.

Most methods mentioned above focus on designing specific hand-crafted features and thus being limited in modeling temporal dynamics. Recently, deep learning methods are proposed to learn robust feature representations and to model the temporal dynamics without segmentation. In [9], a hierarchical RNN is utilized to model the temporal dynamics for skeleton based action recognition. Zhu *et al.* [49] proposed a deep LSTM network to model the inherent correlations among skeleton joints and the temporal dynamics in various actions. However, there are few approaches proposed for action detection on 3D skeleton data. Li *et al.* [16] introduced a Joint Classification Regression RNN to avoid sliding window design which demonstrates state-of-the-art performance for online action detection. In this work, we propose a large-scale detection benchmark to promote the study on continuous action understanding.

2.3 3D Activity Datasets

We have also surveyed other tens of well-designed action datasets which greatly improved the study of 3D action analysis. These datasets have promoted the construction of standardized protocols and evaluations of different approaches. Furthermore, they often provide some new directions in action recognition and detection previously unexplored. A comparison among several datasets and PKU-MMD is given in Table 1.

G3D [2] is designed for real-time action recognition 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.

CAD-60 [30] & *CAD-120* [31] are two special multi-modality datasets. Compared to *CAD-60*, *CAD-120* provides extra labels of temporal locations. However, the limited number of video instants is their downside.

Watch-n-Patch [44] and *Compostable Activities* [17] are the first datasets focusing on the continues sequences and the inner combination of activities in supervised or unsupervised methods. Those consist of moderate number of action instances. Also, the number of instance actions in one video is limited and thus cannot fulfill the basic requirement for deep network training.

OAD [16] dataset is a new dataset focusing on online action detection and forecast. 59 videos were captured by Kinect v2.0 devices which composed of daily activities. This dataset proposes a series of new protocols for 3D action detection and raises an online demand.

However, as the quick development of action analysis, 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 perspectives in Table 2.

Table 2: The desirable properties of PKU-MMD dataset.

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

3 THE DATASET

3.1 PKU-MMD Dataset

PKU-MMD is our new large-scale dataset focusing on long continuous sequences action detection and multi-modality action analysis. The dataset is captured via the Kinect v2 sensor, which can collect color images, depth images, infrared sequences and human skeleton joints synchronously. We collect 1000+ long action sequences, each of which lasts about 3∼4 minutes (recording ratio set to 30 FPS) and contains approximately 20 action instances. The total scale of our dataset is 5,312,580 frames of 3,000 minutes with 20,000+ temporally localized actions.

We choose 51 action classes in total, which are divided into two parts: 43 daily actions (drinking, waving hand, putting on the glassed, *etc.*) and 8 interaction actions (hugging, shaking hands, *etc.*). Table 3 illustrates more details on action categories.

We invite 66 distinct subjects for our data collection. Each subjects takes part in 4 daily action videos and 2 interactive action videos. The ages of the subjects are between 18 and 40. We also assign a consistent ID number over the entire dataset in a similar way in [24].

To improve the sequential continuity of long action sequences, the daily actions are designed in a weak connection mode. For example, we design an action sequence of *taking off shirt, taking off cat, drinking water*and *sitting down* to describes the scene that occur after going back home. Note that our videos only contain one part of the actions, either daily actions or interaction actions. We design 54 sequences and divide subjects into 9 groups, and each groups randomly choose 6 sequences to perform.

For the multi-modality research, we provide 5 categories of resources: depth maps, RGB images, skeleton joints, infrared sequences, and RGB videos. 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. Joint information consists of 3-dimensional locations of 25 major body joints for detected and tracked human bodies in the scene. We further provide the confidence of each joints point as appendix. RGB videos are recorded in the provided resolution of 1920 × 1080. Infrared sequences are also collected and stored frame by frame in 512 × 424.

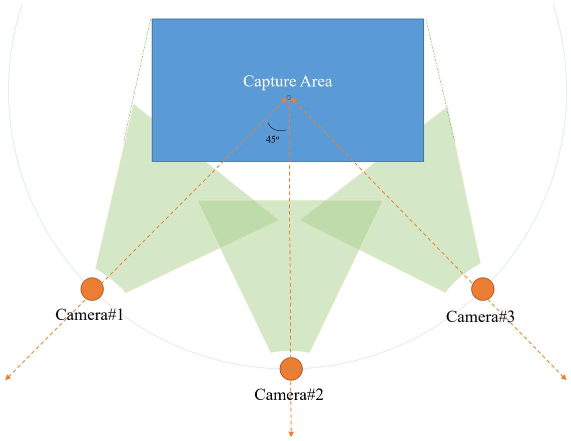


Figure 2: Camera setting for multi-view video recording of PKU-MMD dataset. Three camera views are included. Note that each subject will perform an action instance toward a random camera.

3.2 Developing the Dataset

Building a large scale dataset for computer vision task is traditionally a difficult task. To collect untrimmed videos for detection task, the main time-consuming work is labeling the temporal boundaries. The goal of PKU-MMD is to provide a large-scale continuous multi-modality 3D action dataset, the items of which contain a series of compact actions. Thus we combine traditional recording approaches with our proposed validation methods to enhance the robustness of our dataset and improve the efficiency.

We now fully describe the collecting and labeling process for obtaining PKU-MMD dataset. Inspired by [24], we firstly capture long sequences from Kinect v2 sensors with a well-designed standards. Then, we rely on volunteers to localize the occurrences of dynamic and verify the temporal boundaries. Finally, we design a cross-validation system to obtain labeling correction confidence evaluation.

•**Recording Multi-Modality Videos:** After designing several action sequences, we carefully choose a daily-life indoor environment to capture the video samples where some irrelevant variables are fully considered. Considering that the temperature changes will lead to the deviation of infrared sequences, we fully calculate the distance among the action occurrence, windows and Kinect devices. Windows are occluded for illumination consistency. We use three cameras in the fixed angle and height at the same time to capture three different horizontal views. We set up an action area with 180*cm* as length and 120*cm* as width. Each subject will perform each action instances in a long sequence toward a random camera, and it is accepted to perform two continuous actions toward different cameras. The horizontal angles of each camera is

Table 3: A detailed list about 51 action categories used in PKU-MMD dataset.

Interaction with items(13)	drop	pointing to something with finger	check time (from watch)
	pickup	put something inside pocket	playing with phone/tablet
	taking a selfie	use a fan (with hand or paper)	reading
	writing	make a phone call/answer phone	typing on a keyboard
	tear up paper		
Dressing related(7)	take off glasses	take off jacket	take off a hat/cap
	wear on glasses	take out something from pocket	put on a hat/cap
	wear jacket		
Home related(5)	drink water	brushing hair	wipe face
	eat meal/snack	brushing teeth	
Health related(4)	touch head (headache)	touch chest (stomachache/heart pain)	touch back (backache)
	touch neck (neckache)		
Interaction with person(8)	kicking other person	point finger at the other person	hugging other person
	pushing other person	giving something to other person	pat on back of other person
	handshaking	punching/slapping other person	
others(14)	bow	cross hands in front (say stop)	hand waving
	cheer up	hopping (one foot jumping)	clapping
	jump up	rub two hands together	standing up
	salute	kicking something	falling
	sitting down	throw	

$-45°$, $0°$, and $+45°$, as shown in Figure 2, with a height of $120cm$. An example of our multi-modality data can be found in Figure 4.

•**Localizing Temporal Intervals:** At this stages, captured video sources are labeled on frame level. We employ volunteers to review each video and give the proposal temporal boundaries of each action presented in the long video. In order to keep high annotation quality, we merely employ proficient volunteers who have experiences in labeling temporal actions. Furthermore, there will be a deviation for the temporal labels of a same action from different persons. Thus we divide actions into several groups and the actions in each group are labeled by only one person. At the end of this process, we have a set of verified untrimmed videos that are associated to several action intervals and label correspondingly.

•**Verifying and Enhancing Labels:** Unlike recognition task which merely need one label for an trimmed video clip, the probability of error on temporal boundaries will be much higher. Moreover, during the labeling process we observe that approximate 10-frames expansion of action interval is sometimes accepted in some instance. To further improve the robustness of our dataset, we propose a system of labeling correction confidence evaluation to verify and enhance the manual labels. Firstly, we design basic evaluation protocol of each video, like *If there is overlap of actions or Is the length of an action reasonable.* Thanks to multi-view capturing, we then use cross-view method to evaluate and verify the data label. The protocol guarantees the consistency of videos of each view.

4 EXPERIMENTS

4.1 Evaluation Protocols

To obtain a standard evaluation for the results on this benchmark, we define several criteria for the evaluation of the precision and recall scores in detection tasks. We propose two dataset partition settings with several precision protocols.

To evaluate the precision on the proposed action intervals with confidences, two tasks must be considered. One is to determine if the proposed interval is positive, and the other is to evaluate the performance of precision and recall. For the first task, there is a basic criterion to evaluate the overlapping ratio between the predicted action interval I and the ground truth interval I^* with a threshold θ. The detection interval is correct when

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

where $I \cap I^*$ denotes the intersection of the predicted and ground truth intervals and $I \cup I^*$ denotes their union. So, with θ, the $p(\theta)$ and $r(\theta)$ can be calculated.

•**F1-Score:** With the above criterion to determine a correction detection, the F1-score is defined as

$$\mathrm{F1}(\theta) = 2 \cdot \frac{p(\theta) \times r(\theta)}{p(\theta) + r(\theta)}. \tag{2}$$

F1-score is a basic evaluation criterion regardless of the information of the confidence of each interval.

•**Interpolated Average Precision (AP):** Interpolated average precision is a famous evaluation score using the information of confidence for ranked retrieval results. With confidence changing, precision and recall values can be plotted to give a precision-recall curve. The interpolated precision p_{interp} at a certain recall level r is defined as the highest precision found for any recall level $r' \geq r$:

$$p_{interp}(r, \theta) = \max_{r' \geq r} p(r', \theta). \tag{3}$$

Note that r is also determined by overlapping confidence θ. The interpolated average precision is calculated by the arithmetic mean of the interpolated precision at each recall level.

$$\mathrm{AP}(\theta) = \int_0^1 p_{interp}(r, \theta)\, \mathrm{d}r. \tag{4}$$

Method	Cross-view					
	θ	F1	AP	mAP$_a$	mAP$_v$	2D-AP
JCRRNN	0.1	0.671	**0.728**	**0.699**	**0.642**	**0.460**
	0.5	**0.526**	**0.544**	**0.533**	**0.473**	
SVM	0.1	0.399	0.236	0.240	0.194	0.073
	0.5	0.131	0.031	0.036	0.031	
BLSTM	0.1	**0.676**	0.525	0.545	0.508	0.187
	0.5	0.333	0.124	0.159	0.139	
STA-LSTM	0.1	0.613	0.468	0.476	0.439	0.180
	0.5	0.316	0.130	0.155	0.134	
Method	Cross-subject					
	θ	F1	AP	mAP$_a$	mAP$_v$	2D-AP
JCRRNN	0.1	0.500	**0.479**	0.452	0.431	**0.288**
	0.5	**0.366**	**0.339**	**0.325**	**0.297**	
SVM	0.1	0.332	0.179	0.181	0.143	0.051
	0.5	0.092	0.016	0.021	0.018	
BLSTM	0.1	**0.629**	0.464	**0.479**	**0.442**	0.164
	0.5	0.291	0.095	0.130	0.108	
STA-LSTM	0.1	0.586	0.427	0.444	0.405	0.156
	0.5	0.284	0.101	0.131	0.116	

Table 4: Comparison of results among several approaches on 3D action detection with various metrics.

•**Mean Average Precision (mAP)**: With several parts of retrieval set Q, each part $q_j \in Q$ proposes m_j action occurrences $\{d_1, \ldots d_{m_j}\}$ and r_{jk} is the recall result of ranked k retrieval results, then mAP is formulated by

$$\mathrm{mAP}(\theta) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} p_{interp}(r_{jk}, \theta). \qquad (5)$$

Note that with several parts of retrieval set Q, the AP score (4) is discretely formulated.

We design two splitting protocols: mean average precision of different actions (mAP$_a$) and mean average precision of different videos (mAP$_v$).

•**2D Interpolated Average Precision:** Though several protocols have been designed for information retrieval, none of them takes the overlap ratio into consideration. We can find that each AP score and mAP score is associated to θ. To further evaluate the performance of precisions of different overlap ratios, we now propose the 2D-AP score which takes both retrieval result and overlap ratio of detection into consideration:

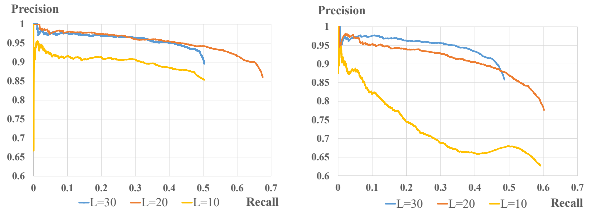
$$\text{2D-AP} = \iint_{r\in[0,1],\,\theta\in[0,1]} p_{interp}(r, \theta)\ \mathrm{d}r\mathrm{d}\theta. \qquad (6)$$

This section presents a series of evaluation of basic detection algorithms on our benchmark. Due to the fact that there is few implementation for 3D action detection, these evaluations also serve to illustrate the challenge activity detection is and call on new explorations.

4.2 Experiment Setup

In this part, we implement several detection approaches for the benchmarking scenarios for the comparison on PKU-MMD dataset.

4.2.1 Dataset Partition Setting. This section introduces the basic dataset splitting settings for various evaluation, including cross-view and cross-subject settings.



(a) Stride = 30

(b) Stride = 20

Figure 3: Different Precision-Recall curves (overlapping ratio θ is set to 0.2) under different settings with different window size and stride. L stands for the length of sliding windows.

•**Cross-View Evaluation:** For cross-view evaluation, the videos sequences from the middle and right Kinect devices are chosen for training set and the left is for testing set. Cross-view evaluation aims to test the robustness in terms of transformation (*e.g.*translation, rotation). For this evaluation, the training and testing sets have 717 and 359 video samples, respectively.

•**Cross-Subject Evaluation:** In cross-subject evaluation, we split the subjects into training and testing groups which consists of 57 and 9 subjects respectively. For this evaluation, the training and testing sets have 944 and 132 long video samples, respectively. Cross-subject evaluation aims to test the ability to handle intra-class variations among different actors.

4.2.2 Temporal Detection Method. Here we introduce several approaches for action detection.

•**Sliding Window + SVM:** Leveraging the insight from the RGB-based activity detection approaches, we design several slide-window detection approaches. For the classifier, one basic method is using SVM classifier which is agility ans easy to train.

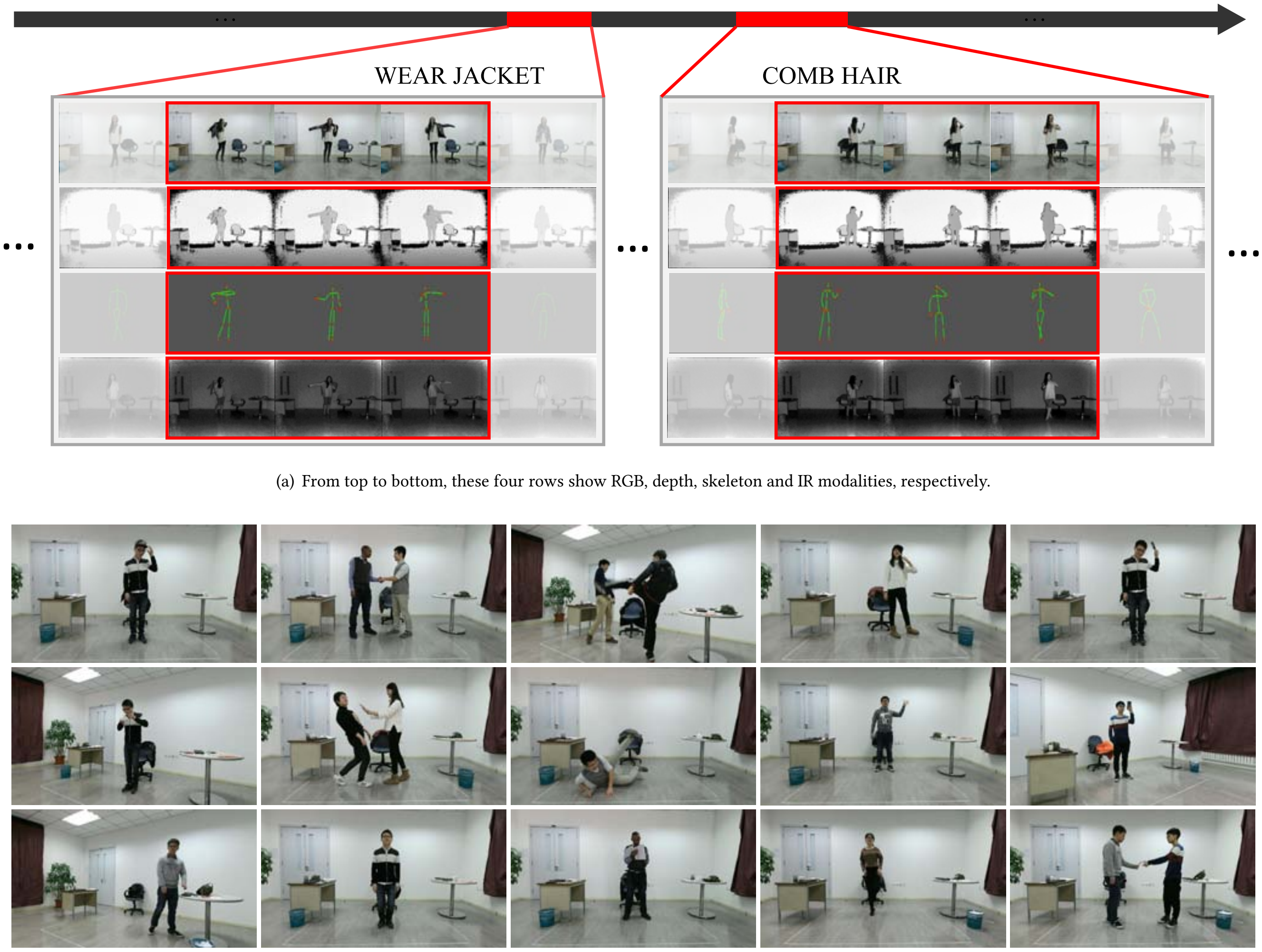
•**Sliding Window + BLSTM:** Due to the ability to model long-term and short-term dynamics, three stacked bidirectional LSTM (BLSTM) [49] network has been proved effective to model skeleton-based activities. The succinctness of skeleton data limit parameter explosion of LSTM and make it easy to converge.

•**Sliding Window + STA-LSTM:** Spatial-temporal attention network [28] is a state-of-the-art work proposed for action recognition with unidirectional LSTM. It proposes a regularized cross-entropy loss to drive the model learning process which conducts automatic mining of discriminative joints together with explicitly learning and allocating the content-dependent attentions to the output of each frame to boost recognition performance.

•**Joint Classification Regression RNN (JCRRNN):** Besides proposing the online action detection task, Li *et al.* [16] proposed a Joint Classification Regression RNN which implement frame level real-time action detection.

4.3 Action Detection Results

In the detection task, the goal is to find and recognize all activity instances in an untrimmed video. Detection algorithms should provide the start and end points with action labels. We exploit the



(a) From top to bottom, these four rows show RGB, depth, skeleton and IR modalities, respectively.

(b) We collect 51 actions performed by 66 subjects, including actions for single and pairs.

Figure 4: Sample frames from PKU-MMD. The top figure shows an example of continuous action detection in multi-modality, and about 20 action instances can be found within one sequences. The bottom figure depicts the diversity in categories, subjects and camera viewpoints.

location annotations of PKU-MMD to compare the performances of above methods.

As the skeleton is an effective representation, we implement several experiments to evaluate the ability to model dynamics and activity boundaries localizing. Table 4 shows the comparison of different combination of skeleton representation and temporal featuring methods. SVM performs worst because it only learn a linear transformation and weak to model high-level semantics. STA-LSTM performs worse than BLSTM possibly due to the large margin in amount of parameters. And STA-LSTM also learn a spatial-temporal attention of an entire activities thus may vulnerable to sliding-windows approaches. Joint classification regression RNN achieves remarkable results, because it utilizes frame-level predictions and thus is more compatible with stricter localization requirements.

We further analyze the different performances with several sliding-window approaches. We show Precision-Recall curves of *BLSTM* method in Figure 3. The performance is influenced by window size and stride. When stride is fixed, windows in smaller size contain less context information while noises can be involved by larger window size. However, smaller window size always leads to higher computation complexity. And obviously, too large window size will mix several dynamics information and confuse classifiers. So it is essential to balance window scale according to dataset, which illustrates the limitation of sliding window approaches. For stride, obviously, smaller stride achieves better results due to dense sampling while costing more time. In our following experiments, we try different settings of window size and stride as a trade-off between performance and speed. The results of different sliding-window approaches are shown in Figure 3.

5 CONCLUSION

In this paper, we propose a large-scale multi-modality 3D dataset (PKU-MMD) for human activity understanding, especially for action detection which demands localizing temporal boundaries and recognizing activity category. Performed by 66 actors, our dataset includes 1076 long video sequences, each of which contains 20 action instances of 51 action classes. Compared with current 3D datasets for temporal detection, our dataset is much larger (3000 minutes and 5.4 million frames in total) and contains much varieties (3 views, 66 subjects) in different aspects. The multi-modality attribution and larger scale of the collected data enable further experiments on deep networks like LSTM or CNN. Based on several detection retrieval protocols, we design a new 2D-AP evaluation for action detection task which takes both overlapping and detection confidence into consideration. We also design plenty experiments to evaluate several detection methods on PKU-MMD benchmarks. The results show that existing methods are not satisfied in terms of performance. Thus, large-scale 3D action detection is far from being solved and we hope this dataset can draw more studies in action detection methodologies to boost the action detection technology.

REFERENCES

[1] Jake K Aggarwal and Lu Xia. 2014. Human activity recognition from 3D data: A review. *PRL* (2014).

[2] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. 2012. G3D: A gaming action dataset and real time action recognition evaluation framework. In *CVPR*.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

[4] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao. 2016. RGB-D datasets using Microsoft Kinect or similar sensors: a survey. *Multimedia Tools and Applications* (2016).

[5] Lulu Chen, Hong Wei, and James Ferryman. 2013. A survey of human motion analysis using depth imagery. *PRL* 34 (2013).

[6] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. 2012. Human daily action analysis with multi-view and color-depth data. In *ECCV*.

[7] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. *Online Action Detection.*

[8] 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 *CVPR*.

[9] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*.

[10] Yusuke Goutsu, Wataru Takano, and Yoshihiko Nakamura. 2015. Motion Recognition Employing Multiple Kernel Learning of Fisher Vectors Using Local Skeleton Features. In *ICCV*.

[11] Minh Hoai and Fernando De la Torre. 2014. Max-margin early event detectors. *IJCV* (2014).

[12] Mohamed E Hussein, Marwan Torki, Mohammad Abdelaziz Gowayyed, and Motaz El-Saban. 2013. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations.. In *IJCAI*.

[13] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

[14] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *CVPR*.

[15] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. 2010. Action recognition based on a bag of 3D points. In *CVPR*.

[16] 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 *ECCV*.

[17] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. 2014. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *CVPR*.

[18] Bingbing Ni, Gang Wang, and Pierre Moulin. 2013. RGBD-hudaact: A color-depth video database for human daily activity recognition. Springer.

[19] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley MHAD: A comprehensive multimodal human action database. In *WACV*.

[20] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the fisher kernel for large-scale image classification. In *ECCV*.

[21] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. 2016. Histogram of oriented principal components for cross-view action recognition. *TPAMI* (2016).

[22] 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. 2014. ImageNet Large Scale Visual Recognition Challenge. *CoRR* (2014).

[23] Michael S Ryoo. 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*.

[24] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*.

[25] Amir Shahroudy, Tian-Tsong Ng, Qingxiong Yang, and Gang Wang. 2016. Multimodal multipart learning for action recognition in depth videos. *TPAMI* (2016).

[26] Amr Sharaf, Marwan Torki, Mohamed E Hussein, and Motaz El-Saban. 2015. Real-time multi-scale action detection from 3D skeleton data. In *WACV*.

[27] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*.

[28] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2016. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. *AAAI* (2016).

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

[30] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2011. Human Activity Detection from RGBD Images. *AAAI* (2011).

[31] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2012. Unstructured human activity detection from RGBD images. In *ICRA*.

[32] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-resnet and the impact of residual connections on learning. *arXiv* (2016).

[33] Yicong Tian, Rahul Sukthankar, and Mubarak Shah. 2013. Spatiotemporal deformable part models for action detection. In *CVPR*.

[34] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. 2014. Human action recognition by representing 3D skeletons as points in a lie group. In *CVPR*.

[35] Raviteja Vemulapalli and Rama Chellapa. 2016. Rolling rotations for recognizing human actions from 3D skeletal data. In *CVPR*.

[36] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *ICCV*.

[37] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*.

[38] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. 2014. Cross-view action modeling, learning and recognition. In *CVPR*.

[39] Limin Wang, Yu Qiao, and Xiaoou Tang. 2014. Action recognition and detection by combining motion and appearance features. *THUMOS* (2014).

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

[41] 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 *ECCV*.

[42] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. 2013. Modeling 4D human-object interactions for event and object recognition. In *ICCV*.

[43] Ping Wei, Nanning Zheng, Yibiao Zhao, and Song-Chun Zhu. 2013. Concurrent action detection with structural prediction. In *CVPR*.

[44] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. 2015. Watch-n-patch: Unsupervised understanding of actions and relations. In *CVPR*.

[45] 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 *ACM MM*.

[46] 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 *CVPR*.

[47] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. 2013. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *CVPR*.

[48] Jing Zhang, Wanqing Li, Philip O Ogunbona, Pichao Wang, and Chang Tang. 2016. RGB-D-based action recognition datasets: A survey. *PR* (2016).

[49] 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. *AAAI* (2016).