A Large Scale Dataset for Cross Modal Action Understanding

Quan Kong[1,a)] Ziming Wu[2,3] Ziwei Deng[1] Martin Klinkigt[1] Bin Tong[1]
Tomokazu Murakami[1]

Abstract: In recent years, many vision-based multimodal datasets have been proposed for human action understanding. Except RGB, most of them provide only one additional modality like depth. Unlike vision modalities, body-worn sensors or passive sensing can however avoid the failure of action understanding in cases of occlusion. Among the state-of-the-art bechmarks, a standard large-scale dataset does not exist, in which different types of modalities are integrated. To address the disadvantage of vision-based modalities, this paper introduces a new large-scale benchmark recorded from 20 distinct subjects with seven different types of modalities: RGB videos, keypoints, acceleration, gyroscope, orientation, Wi-Fi and pressure signal. The dataset consists of more than 36k video clips for 37 action classes covering a wide range of daily life activities such as desktop-related and check-in-based ones in four different distinct scenarios. On the basis of our dataset, we propose a novel multi modality distillation model with attention mechanism that appropriately utilizes both RGB-based and sensor-based modalities. The proposed model significantly improves performance of action recognition by up-to 8% compared to models without using sensor-based modalities. The experimental results confirm the effectiveness of our model on cross-subject, -view, -scene and -session evaluation criteria. We believe that this new large-scale multimodal dataset will contribute the community of multimodal-based action understanding.

Keywords: Dataset, Wearable device, Cross modal learning, Action recognition.

1. Introduction

Human action understanding is an important fundamental technology for supporting several real world applications such as surveillance system, health care services and factory efficiency services. In recent years, vision-based models dominates the community of action understanding due to the advance of deep learning technologies [30], [37], [42]. Meanwhile, utilizing of body-worn inertial sensors e.g., accelerator, gyroscope and orientation to capture human motions is a newly emerged way of realizing human action recognition [7], [25], [31]. It is well known that vision-based and sensor-based information in action recognition is complementary. Sensor information is difficult to be affected by occlusion, illumination changes in which vision-based models may encounter problems. Therefore, it is considerable to utilize both vision-based and sensor-based modalities to improve performance of action understanding in multimodal [13], [23], [29] and crossmodal [3], [22], [41] manners.

However, in the community of action understanding, a standard large-scale benchmark does not exist, in which both vision-based and sensor-based modalities are aggregated and a wide range of activities are provided. The current multimodal datasets for action understanding have following four limitations. First, there is the limited scale of vision-based and sensor-based modalities. There are some but limited number of large-scale multimodal action datasets [21], [28] focusing on 3D human action recognition or detection. However, only three to four vision related modality are provided in the existing datasets. Second, there is the limited number of supported action understanding task with enough instances per action. Most existing datasets only support action recognition but can hardly be utilized for action detection. Third, actions in the existing datasets are taken in a fixed location. Therefore, the distance between the actor and the camera does not change. In addition, the actions always appear in the center of the camera. These limit the naturality and perspective

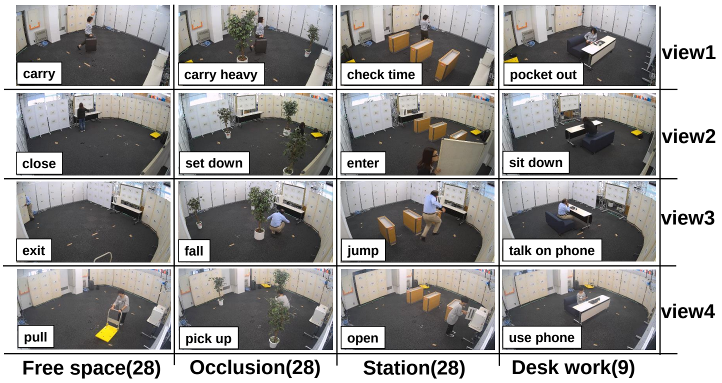

Fig. 1: The illustration of our dataset. Each column shows actions under a scenario. The number after the name of scenario denotes the amount of action category in the scenario. Each row denotes the action under one of four camera views.

[1] Hitachi,Ltd. Research & Development Group
[2] Hong Kong University of Science and Technology
[3] Work is done during internship at Hitachi.
[a)] quan.kong.xz@hitachi.com

feature under each camera view. Forth, the limited number of instances for each modality with distinct subject, scenario, view and session in a factored data structure, especially for crossmodal related research across large domain gaps. This paper proposes a new multimodal dataset to overcome the above limitations, especially for expanding the multimodal research on human action understanding across modalities, like from vision to Wifi-signal stength.

Our dataset, named as multimodal action dataset (MMAct), consists of 36,764 trimmed clips with seven types of modalities for 20 subjects, which include RGB videos, acceleration sensor, gyroscope sensor, orientation sensor, Wi-Fi signal and keypoints. The illustration of our dataset is shown on Fig. 1. MMAct is designed under a semi-natural data collection protocol that a random walk will be performed between the end of current action and the start of next action. The action is only performed after a start sign was given from the outside monitor. This protocol makes sure that the action will occur randomly in the action area to provide various perspective action video in different camera views.

For traditional multimodal models, the more modality a model uses, the higher cost is taken for the model to be deployed in a realistic environment. The technique of crossmodal transfer, a kind of knowledge distillation [16], is a useful way to allow a model with only one modal input to achieve the performances using multiple modalities. For example, a student model with RGB input learns complementary information from other modalities from depth or keypoint, which is served as teacher information. At test phase, only RGB information is used in the student network that is able to achieve better performance of action recognition than the model with RGB information.

Different from the existing methods that focus on modality transfer cross vision-based modalities, we intend to move a further step towards knowledge transfer from sensor-based modalities to vision-based modalities. We propose a novel multimodality distillation model with attention mechanism to realize an adaptive knowledge distillation via the learning of teacher and student models. The main contributions of our work are three-folds:

- To the best of our knowledge, MMAct is the largest multimodal dataset that includes both vision-based and sensor-based modalities. It helps research community to move towards crossmodal action analysis.
- Inspired by the knowledge distillation, we propose a novel multimodality distillation model with attention mechanism. This model has a student network with input of RGB information, which learn useful information from a teacher network with input of multiple sensor-based modalities.
- Our experimental results confirm the effectiveness of our model in our dataset. A significant improvement can be achieved in cases in which RGB modality may fail to recognize the actions.

2. Related Work

In this section, we illustrates some related datasets and works in action understanding. The most traditional and famous ones are listed with brief introductions. For a more complete conclusion, readers could refer to these survey papers [1], [6], [43], [44].

2.1 Related datasets

Some traditional and typical multimodal datasets for action understandin are dicussed below, with a comparison between them and MMact in Table.1.

MSR-Action3D dataset [18] is one of the earliest datasets which has contributed to several 3D action analysis researches. This dataset is composed of depth sequences of gaming actions and 3D body keypoints data made up by 20 different body joints. Multiview 3D event [38] and Northwestern-UCLA [35] datasets utilized a multi-view method to capture the 3D videos using more than one Kinect cameras. This method has been widely utilized in many 3D datasets. NTU RGB+D [28] is the state-of-the-art large-scale benchmark for human activities analysis, which contains videos of 60 action classes captured from 80 views with 40 subjects. It illustrated a series of standards of large-scale dataset and was applied by many works. Achieving promising results on this benchmark shows great importance in this field. Since only clipped sequences are available in these datasets, they cannot be applied to action detection and some other researches. G3D [5] is the earliest action detection dataset, of which most sequences contain multiple gaming actions in an indoor environment with a fixed camera. Watch-n-Patch [39] and Compostable Activities [20] are the first datasets focusing on the hidden correlation of actions in supervised or unsupervised methods. However, the number of instance actions in each video is not enough to fulfill the basic requirement for training a deep network. PKU-MMD [21] is a large-scale benchmark for human action detection, which has large number of instances for different modalities, including RGB, depth, infrared radiation and keypoints. Nevertheless, it was still limited to the vision modalities.

CMU-MMAC [31] is a multi modality human activity dataset combining vision modalities with sensor signals, including RGB, depth, keypoints, and sensor signals obtained by accelerometers and microphones. This dataset was collected in a kitchen and 25 subjects were recorded cooking and food preparation. MHAD [25] and UTD-MHAD [7] include sensor signals as well, providing more action classes and instances to support the evaluation of new algorithms. However, these datasets are no longer sufficient and satisfied enough for fast developing data-driven algorithms. Thus, we considered to build a large-scale dataset MMAct with various kinds of modalities and actions, combining with random walk and occlusion, providing both untrimmed and action-clipped data to support different level researches.

2.2 Multimodal action recognition

Action recognition has been developed for a long period, but action recognition based on multi modalities is a reletively new topic due to the development of deep learning technology and hardwares such as depth cameras and wearable devices. There are some typical ideas of dealing with multi modality data. [34] proposed a 3D ConvNets for extracting spatiotemporal features to model appearance and motion information simultaneously. [29] designed a deep autoencoder architecture to decompose its mul-

Table 1: Comparison between different multimodal datasets for action understanding. Ego: Egocentric view, D:Depth, Acc:Acceleration, Mic:Microphone, Gyo:Gyroscope, Ori:Orientation.

Datasets	Classes	Instances	Subjects	Scene	Views	Modalities	Temporal Localization	Random Walk	Occlusion	Year
MSR-Action3D [18]	20	567	10	1	1	D+Keypoints	No	No	No	2010
CAD-60 [32]	12	60	4	5	-	RGB+D+Keypoints	No	No	No	2011
RGBD-HuDaAct [24]	12	60	4	1	-	RGB+D+Keypoints	No	No	No	2011
Act4^2[8]	14	6844	24	1	4	RGB+D	No	No	No	2012
UTKinect-Action3D [40]	10	200	10	1	4	RGB+D+Keypoints	No	No	No	2012
3D Action Pairs [26]	12	360	10	1	1	RGB+D+Keypoints	No	No	No	2013
Multiview 3D Event [38]	8	3815	8	1	3	RGB+D+Keypoints	No	No	No	2013
Northwestern-UCLA [35]	10	1475	10	1	1	RGB+D+Keypoints	No	No	No	2014
Office Activity [36]	20	1180	10	-	3	RGB+D+Keypoints	No	No	No	2014
NTU-RGB+D [28]	60	56880	40	1	80	RGB+D+Keypoints	No	No	No	2016
G3D [5]	20	1467	10	1	-	RGB+D+Keypoints	Yes	No	No	2012
CAD-120 [33]	20	1200	4	1	-	RGB+D+Keypoints	Yes	No	No	2013
Compostable Activities [20]	16	2529	14	1	1	RGB+D+Keypoints	Yes	No	No	2014
Watch-n-Patch [39]	21	2500	7	13	-	RGB+D+Keypoints	Yes	No	No	2015
OAD [19]	10	700	-	1	1	RGB+D+Keypoints	Yes	No	No	2016
PKU-MMD [21]	51	21545	66	1	3	RGB+D+IR+Keypoints	Yes	No	No	2017
CMU-MMAC [31]	5	186	39	1	5	RGB+D+Keypoints+Acc+Mic	No	No	No	2010
MHAD [25]	11	660	12	1	12	RGB+D+Keypoints+Acc+Mic	No	No	No	2013
UTD-MHAD [7]	27	861	8	1	1	RGB+D+Keypoints+Acc+Gyo	No	No	No	2015
MMAct	**37**	**36764**	**20**	**4**	**4+Ego**	**RGB+Keypoints+Acc+ Gyo+Ori+Wi-Fi+Pressure**	**Yes**	**Yes**	**Yes**	**2019**

timodal input (RGB and depth) to modality-specific parts and a structured sparsity learning machine for a proper fusion of decomposed feature components, achieving state-of-the-art accuracy for action classification on 5 challenging datasets. The two-stream architecture introduced by [30] has been widely developed in several works. How one could insert cross-stream connections to fuse the two networks are discussed in [11][9]. A novel spatiotemporal architecture was presented in[10], which applied multiplicative interaction of appearance and motion features by injecting motion streams signal into the residual unit of the appearance stream. The network was designed in an end-to-end manner and fully convolutional for both streams. [13] is the most related work sharing the same task with our work. It proposed a new multimodal stream network to exploit and leverage multiple data modalities. Meanwhile, a newly designed hallucination network based on [17] was proposed to mimic the depth stream when relying only on RGB data at test time. However, the modalities used in this work are still RGB and depth, the same as most multimodal works, which shows limitation in modality diversity.

2.3 Crossmodal transfer

Most related to our work is the concept of transfer learning across different modalities. While conventional transfer learning works only focus on category-level knowledge transfer, cross-modal transfer works devote to modality shift, which transfers knowledge learned in one data modality to another.

[17] proposed a modality hallucination architecture to mimic the depth mid-level features to enhance an RGB object detection model. [22] [14] both contributes to supervision transfer, which transfer information from a large labeled source domain to a sparsely labeled or unlabeled target domain. They also contributes to transferring across different tasks: image object recognition to video action recognition. [41] designed a network to learn a non-linear feature mapping from the RGB channels to the thermal channel, in order to reconstruct the thermal channel when only RGB images are available in the pedestrian detection task. Unlike most works focusing on transfer between vision modalities, [45] suggests using vision data to provide crossmodal supervision for a radio data based human pose estimation task. And [3] learns sound representations by transferring discriminative visual knowledge from visual recognition models to the sound modality using unlabeled videos. These works provided promising evaluation results on some multi modality datasets, but nonetheless for most of them, only limited modalities were tested. The reason may be the lack of large-scale multimodal datasets, which can provide more than vision modalities and reach the demand of enough samples for network training.

3. MMAct Dataset

MMAct is a novel large-scale dataset focusing on action recognition/detection tasks and cross-modality action analysis [*1]. We collected 36,000+ temporally localized action instances in 1,968 continuous action sequences, each of which lasts about 3~4 minutes for desk work scene containing 9 action instances, 7~8 minutes for the other scenes with approximately 26~28 action instances. More details are introduced in the following parts.

3.1 Data Modalities

Seven types of modality are provided with the MMAct dataset: RGB videos, acceleration sensor, gyroscope sensor, orientation sensor, Wi-Fi signal, pressure sensor and keypoints of persons. RGB videos were captured by four commercial surveillance cameras (Hitachi DI-CB520) aligned at the four top corners of the space capturing the scene with a resolution of 1920 × 1080 at 30 FPS. Subjects are wearing a Google Glass to record egocentric videos to support action recognition research in this direc-

[*1] https://mmact19.github.io/2019/

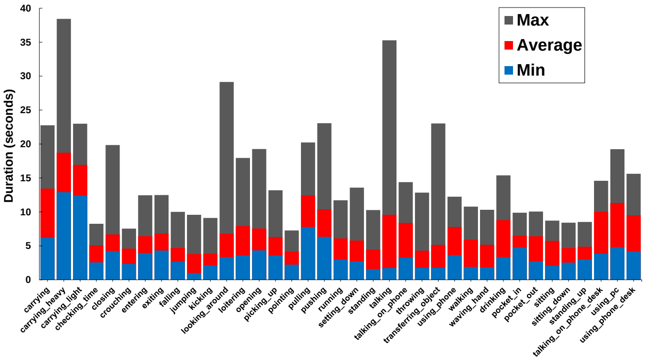

Fig. 2: Average trimmed action clip length per class. Overall the dataset is well balance with only a few outlayers like carry heavy, looking around and talking being longer, due to nature of the class.

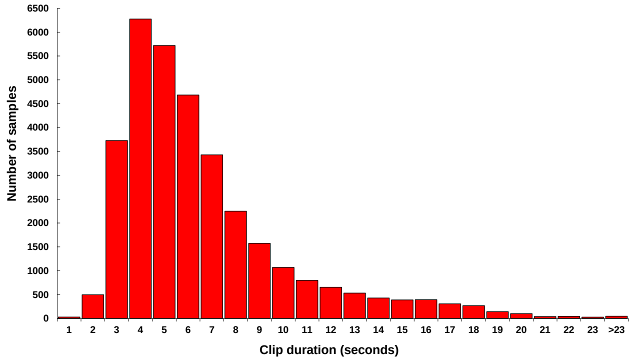

Fig. 3: Distribution of the trimmed action clip length. The average length is in a range from about 3 sec. to 8 sec.

tion. A smartphone (ASUS ZenPhone AR) installed with some initial sensors, such as accelerator and gyroscope, was used to obtain data of acceleration, gyroscope signal, orientation, Wi-Fi signal and pressure. The smartphone was carried and put inside the pocket of the subject's clothes. The acceleration and gyroscope signal both have 3-dimensional axis information, and the orientation modality is represented by 3 types: azimuth, pitch, roll. These 3 modalities are collected at a 100 Hz, 50Hz and 50 Hz sampling rate respectively, while for the Wi-Fi signal and the pressure is 1 Hz and 25 Hz respectively. Subjects are also wearing a smartwatch which further extends the provided acceleration data. Wi-Fi access points were installed at the four corners of the space in order to transmit as well as receive the Wi-Fi signals from the smartphone or each other.

3.2 Data Construction

Class: A total of 37 action classes were considered, which have been categorized into 3 major groups: 16 *complex actions:* carrying, talking, exiting, etc. 12 *simple actions:* kicking, talking on phone, jumping, etc. and 9 *desk actions:* sitting, using PC, pocket out, etc. The grouping of actions tries to follow the pattern introduced by [2]. We summarized the duration of each class and printed the minimum, average and maximum duration of each class in Fig. 2, which illustrates that each action class has plenty of distinct samples in our dataset. Fig. 3 shows the distribution of number of samples for different clip duration, illustrating that we have large number of sequences among different duration and most sequences last 4~6 seconds.

Subject: We invited 20 subjects balanced between 10 males and 10 females for our data collection. The ages of the subjects are between 21 and 49 and their heights are between 147 cm and 180 cm. Each subject has a consistent ID number over the entire dataset.

Scene: We designed 4 scenes in an indoor environment: free space, occlusion, station and desk work. In the scene of free space, there's nothing set up in the area. This is a standard scene following most related datasets. In the scene of occlusion, 3 potted plants were arranged in the space in order to mimic blind spots for the cameras. The subject could be occluded by the potted plants at some directions and positions. Occlusion is a weak point of vision based algorithms, thus we provide this scene aiming to prove that sensor signals are worth exploited to enhance the vision relied systems. In the station scene, 3 gates were set in parallel with a space to go through with a suitcase. It was designed to simulate a real world application scene. In the scene of desk work, a sofa and a deck was arranged in the center of the space for the purpose of recording desk actions.

View: We have videos from 5 views in total. Four of them were recorded from 4 top corners of the space, and one was recorded from the egocentric view by wearing the Google Glass. The cameras were located at the same height recording from a top view.

Session: We defined a session as one untrimmed video consisting of 9 actions for desk work scene and 26 to 28 actions for the other scenes. Each subject was asked to perform each session for 5 times with random changes in motion, direction and position. In this way, the collected data could be distinct and well balanced for each scene, view and subject.

Fig. 8 shows the variety of our views, scenes, also the subjects in age, gender, and height.

3.3 Data Collection

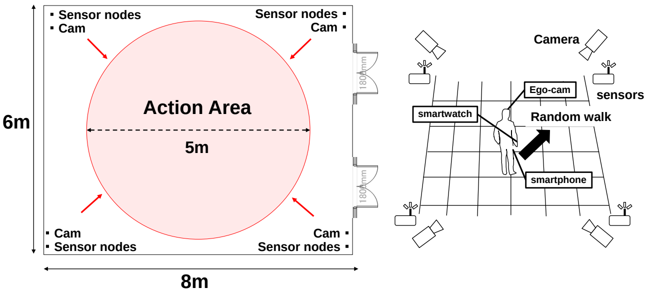

Fig. 4: The environmental setup of the action area showing the size and location of the cameras and sensors.

Generally, collecting untrimmed data for action recognition is a difficult task. The recording environment and process must be appropriate designed and temporal boundaries must be controlling. MMAct was deployed under a semi-naturalistic collection protocol [4] to make sure that the action will occur randomly in the action area to provide various perspective action videos in different camera views.

Recording environment: As Fig. 4 shows, we built our recording environment in a 6m×8m indoor space, with 4 cameras and 4 sensor nodes of the Wi-Fi access points equipped at 4 corners of the space. Subjects were asked to perform actions in a

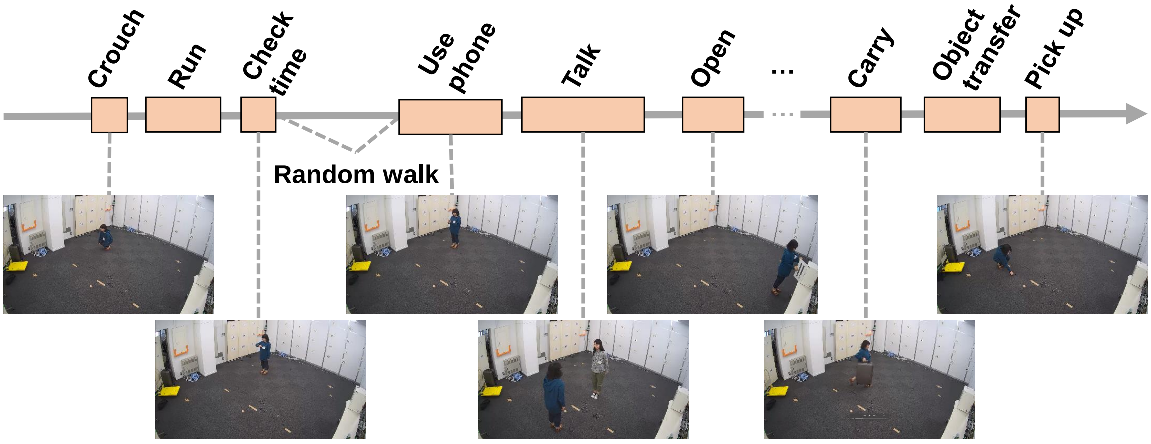

Fig. 5: Sample of our collected action sequence.

circular area of 5m radius, and were equipped with a smartwatch on the right hand, a smartphone in the right pocket of clothes and smart glasses.

Recording process: A series of actions was listed on a worksheet, as Fig. 5 shows as an example. Random walk was performed by subjects between the end of current action and the start of the next action. For the desk work scene, this random walk is with sitting still. Unlike other datasets recording subjects at certain positions and directions, subjects were captured at random positions and directions.

An outside monitor supervising through live videos would give an action command referring to the worksheet when the subject was random walking. Then the monitor gave a start and an end command while labelling the temporal annotation using a toolbox provided. Data collected between the start and end times were labeled with the name of the commanded action class. After hearing the start command, subjects should start within 3 seconds to perform the commanded action and stop after the end command announced. For some continuous actions such as talking and running, subjects were required to keep doing the action until the monitor gives the end command based on self-judgment. For some sudden actions such as throwing and kicking, the subject would randomly walk after the action ends and the monitor would record the end time label based on self-judgment. Thus, usually random walk of less than 3 seconds could be clipped into the action sequences, which is acceptable and reasonable for an action analysis dataset. Furthermore, subjects had freedom in how they performed each action. The monitor provided action classes for subjects to perform, but did not design the concrete motions involved, so that subjects can perform regarding their habits. We invited 20 professional actors to perform these actions in order to make our dataset more naturalistic, realistic and diverse.

4. Proposed Method on Cross Modal

In this section, we introduce a new crossmodal learning method, which is a multi modality attention distillation method to model the vision based human actions with the adaptive weighted side information from inertial sensors using our MMAct dataset.

4.1 Preliminary

As for our method is a distillation based method, we introduce the Knowledge Distillation [16] as our preliminary in advance. The pure Knowledge Distillation is a useful way to significantly improve the accuracy of a small model by transferring the generalization ability of an ensemble of networks, which leaded to a significant performance enhancement on the image classification task. The idea is to allow the student network to capture not only the information provided by the ground truth labels, but also the finer structure learned by the teacher network.

Neural networks generally output class probabilities by using a softmax output layer, which converts the classification score output z_i computed for each class into a probability $p_i = softmax(\frac{z_i}{T})$, where T is a temperature parameter to control the distribution of the probability. A higher value for T means a softer probability distribution over classes. The categorization predictions p_t of a teacher model or an ensemble of models are used as ”soft target” to guide the training of a student model. The student network is then trained by optimizing the following loss function based on cross entropy:

$$L_{KD} = H(y_{gt}, p_s) + \lambda H(p_t, p_s) \qquad (1)$$

where p_s is the probability prediction of the student model and H refers to the cross entropy. The hyper-parameter λ controls the balance between different losses. Note that the first term corresponds to the traditional cross entropy between the output of a network and ground truth labels, whereas the second term enforces the student network to learn from the ”soft target” to inherit hidden information discovered by the teacher network.

4.2 Proposed

The overview of our proposed model is shown in Fig. 6. In our framework, teachers are a set of trained specialist models for each teacher modality. We use acceleration, gyroscope and orientation signal as our teacher modalities, and RGB stream of video as our single student modality.

4.2.1 Training of teacher network

Let $D_t = \{(x_i, y_i)\}^m_{i\in N_t}$ denote the training set for the teacher modality $m \in N_m$, N_m represents the number of teacher modalities, x_i is ith action sample, and y_i is it’s corresponding label, N_t

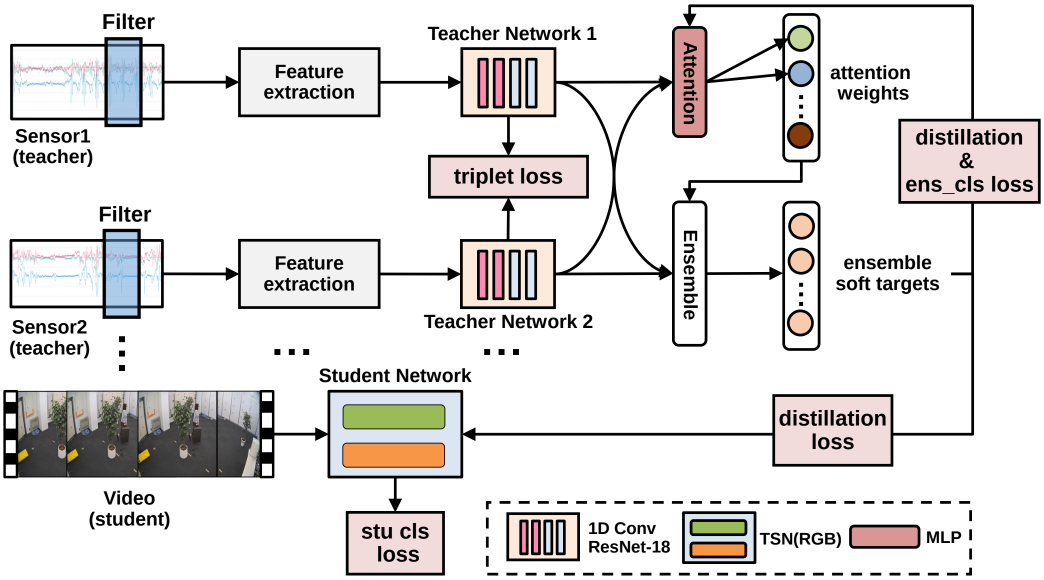

Fig. 6: Architecture of our proposed multi modality attention distillation learning framework. We first train the teacher model separately on its corresponding modality, each teacher model is a 1D Convolutional Neural Network (CNN). Then we use the semantic embedding from the output of softmax layer as the side information of corresponding modality in trained teacher model. As for the softmax layer where the influence of domain difference is the least due to teacher and student share the same semantic space. Afterwards, each semantic embedding is weighted by the attention layer which generates adaptive weights according to the feature representation of input modalities. The semantic embedding with their attention weights are incorporated as an ensemble soft targets for distillation. Finally, we transfer knowledge from multiple teachers into the student network by training it with classification loss and weighted ensemble soft targets distillation loss.

represents the number of samples. We use a sliding window to generate a set of segments $\{(g_{ij}, y_i)\}_{i \in N_t, j \in G_i}$ for sample x_i, where g_{ij} is jth segment for x_i, and all the segments in this set share with the same label y_i as x_i, G_i represents the number of segments for action sample x_i. Each teacher model is an adaption of CNN with 1D conv trained on a segment g_{ij} of the corresponding modality. Note that acceleration, gyroscope and orientation signals in three orthogonal directions (x, y, and z) might be sensitive to sensor placement, e.g., in pants. To cope with the problem, we use the previously proposed combined signal as feature extraction for sensor data, given by $R_i = arcsin(\frac{z_i}{\sqrt{x^2{}_i + y^2{}_i + z^2{}_i}})$ [12], where R_i is the ith combined signal. The combined signal R_i will be the input to the follows 1D conv network. We sampling 64-sample window for 100 Hz acceleration data and 32-sample for 50 Hz gyroscope and orientation data with 70% overlaps for each action clip. As for body-worn sensor is sensitive enough to capture the difference about the same action performed by different subject. Therefore, we use a standard triplet loss [27] instead of a cross-entropy loss to train the teacher models which is being minimized is then $L_\mathrm{t} =$

$$\Sigma[||T_m(g^a_{ij}) - T_m(g^p_{ij})||^2 - ||T_m(g^a_{ij}) - T_m(g^n_{ij})||^2 + \alpha] \tag{2}$$

where $T_m(g_{ij})$ represents the semantic embedding from teacher model T_m, α is a margin that is enforced between positive and negative pairs. Here we want to ensure that a segment g^a_{ij}(anchor) of a specific action of subject is closer to the other g^p_{ij}(positive) of the same action of herself or the other subject, than it is to any g^n_{ij}(negative) of any other actions. We use offline triplet mining to ensure the positive segment of a specific action from the other subject included in each batch.

4.2.2 Multi modality attention distillation

Let $D_s = \{(x_i, y_i)\}^s_{i \in N_t}$ denotes the training set for the student modality s. Our student network is a TSN [37] based network with only RGB branch trained on the sample x_i which is ith action's RGB stream. During the training of student network, the parameter of teacher models are fixed. Let w^m_{ij} be an attention weight of the jth segment for the ith action clip when mth modality. We use $M(F_{ij})$ as a mapping function which is a Multi Layer Perception (MLP) model M to map the feature F_{ij} of segment g_{ij} to $(w^1_{ij}, ..., w^m_{ij}, m \in N_m)$. F_{ij} is an ensemble feature by concatenating the last pooling layer's output from each teacher model. The ith action clip multiple teacher supervision signal is a weighted sum of semantic codes from each teacher modality as an ensemble soft targets:

$$\widehat{z} = \frac{1}{G_i} \sum_{j}^{G_i} \sum_{m}^{N_m} w^m_{ij} T_m(g^m_{ij}) \tag{3}$$

We use cross entropy loss to train the student network with student network classification loss $L_{CS} = H(\widehat{y_i}, \widehat{s_i})$ and distillation loss $L_D = H(\widehat{z_i}, s_i)$, where s_i represents the class probability prediction of the student model, H refers to the cross entropy, the student network loss L_s is organized as:

$$L_s = \sum_{x_i} [\lambda L_{CS} + (1 - \lambda) L_D] \tag{4}$$

where λ is the balance parameter. The attention model M aims to generate adaptive weights for providing more accurate teacher information, that it is optimized by minimize the distillation loss and ensemble teacher classification loss simultaneously as loss L_M:

$$L_M = \sum_{x_i} [\beta L_{CT} + (1 - \beta) L_D] \tag{5}$$

where β is the balance parameter, $L_{CT} = H(\widehat{y_i}, \widehat{z_i})$ is our multiple teacher classification loss.

Table 2: F-measure for action recognition results of all compared methods by using our MMAct dataset.

Method	Train Modality	Test Modality	Cross Subject	Cross View	Cross Scene	Cross Session
Student(Baseline)	RGB	RGB	64.44	62.21	57.91	69.20
Mutli-Teachers	Acc+Gyo+Ori	Acc+Gyo+Ori	62.67	68.13	67.31	70.53
SMD[16]	Acc+RGB	RGB	63.89	70.11	61.56	71.23
MMD	Acc+Gyo+Ori+RGB	RGB	64.33	68.19	62.23	72.08
MMAD (proposed)	**Acc+Gyo+Ori+RGB**	**RGB**	**66.45**	**70.33**	**64.12**	**74.58**

5. Evaluations

5.1 Evaluation Setting

Due to the distinct splitting of the dataset, several settings have been evaluated.

Cross-Subject: in this setting, samples from 80% of the subjects have been used for training the model and the remaining 20% for testing. **Cross-View:** samples from 3 views of all the subjects have been used for training the model and the 4th view for testing. **Cross-Scene:** samples from the scenes except for occlusion of all the subjects have been used for training the model and the occlusion scene from all the subjects for testing. **Cross-Session:** samples from 4 sessions of all the subjects have been used for training the model and the remaining session for testing.

Out of these settings, the cross-subject typically applied in general action classification works do not consider cross-modal settings. For cross-view, self-occlusion (the subject is standing in a way that the action cannot be seen from the camera) is a typical challenge to overcome. In cross-scene, normal occlusion as well as self-occlusion would be typical challenges. The last setting of cross-session is the easiest one, as no domain transfer takes place, e.g. same subjection, view, scenes are available during training and testing.

5.2 Evaluation Method

We evaluated the performance of our method based on the average F-measure ($\frac{2\cdot precision\cdot recall}{precision+recall}$). To investigate its effectiveness, we tested the performance of the other four different methods as shown in Table.2.

- **Student(Baseline)**:our student network trained with only RGB modality.
- **Mutli-Teacher**: our teacher networks trained with 3 types of inertial sensor modality separately with an ensemble testing.
- **SMD**: Single Modality Distillation by using standard knowledge distillation method. Acceleration is used as teacher modality.
- **MMD**: our proposed Multi Modality Distillation method without attention mechanism.
- **MMAD (proposed)**: our proposed multi modality attention distillation method.

We used 1D conv ResNet-18[15] as our teacher network, and TSN with ResNet-18 as our student network.

5.3 Evaluation Results

Evaluation results are presented in Tables 2, 3 and 4. We can see in Table 2 that the student model with only RGB input can already achieve a performance of 57% to 70% across the different settings. The multi-teachers trained and tested with the sensor modalities (accelerator, gyroscope and orientation) can significantly outperform the student model in general, achieving nearly 10% improvement in the cross-scene setting. It proves that models with sensor data generalize better over different settings and are robust against occlusions.

Table 3: Proposed method compared with the baseline vision modality based action recognition methods.

Method	Precision	Recall	F-measure
SVM+HOG[25]	45.31	47.81	46.52
TSN(RGB)[37]	68.32	70.11	69.20
TSN(Optical-Flow)[37]	71.89	73.27	72.57
TSN(Fusion)[37]	75.68	78.57	77.09
MMAD	**73.34**	**75.67**	**74.58**

Table 4: Top 5 improved action classes by the MMAD model compared to TSN with RGB input.

Method	Carry light	Open	Pocket out	Talk on phone	Throw
TSN(RGB)[37]	11.12	28.41	31.57	61.53	48.79
MMAD	**64.51**	**78.67**	**52.63**	**81.31**	**65.30**

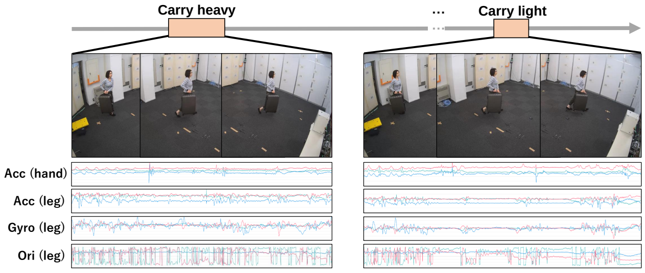

Fig. 7: Sample clips with their paired sensors data related to action "carry heavy luggage" and "carry light luggage".

Introducing accelerator sensor data to the training process improves the performance of the SMD model in most settings, with the cross-view seeing the most significant improvement of almost 8%. Increasing the number of modalities for the MMD model even further, still improves the performance, but not as significantly as with the introduction of the first additional modality. In the proposed model MMAD with attention, we see a more significant improvement in performances while utilizing the same modalities in training and testing as the MMD model.

Interestingly, the proposed MMAD model trained with RGB and sensor modalities can outperform the multi-teacher models with sensor modalities in both training and testing, under all the

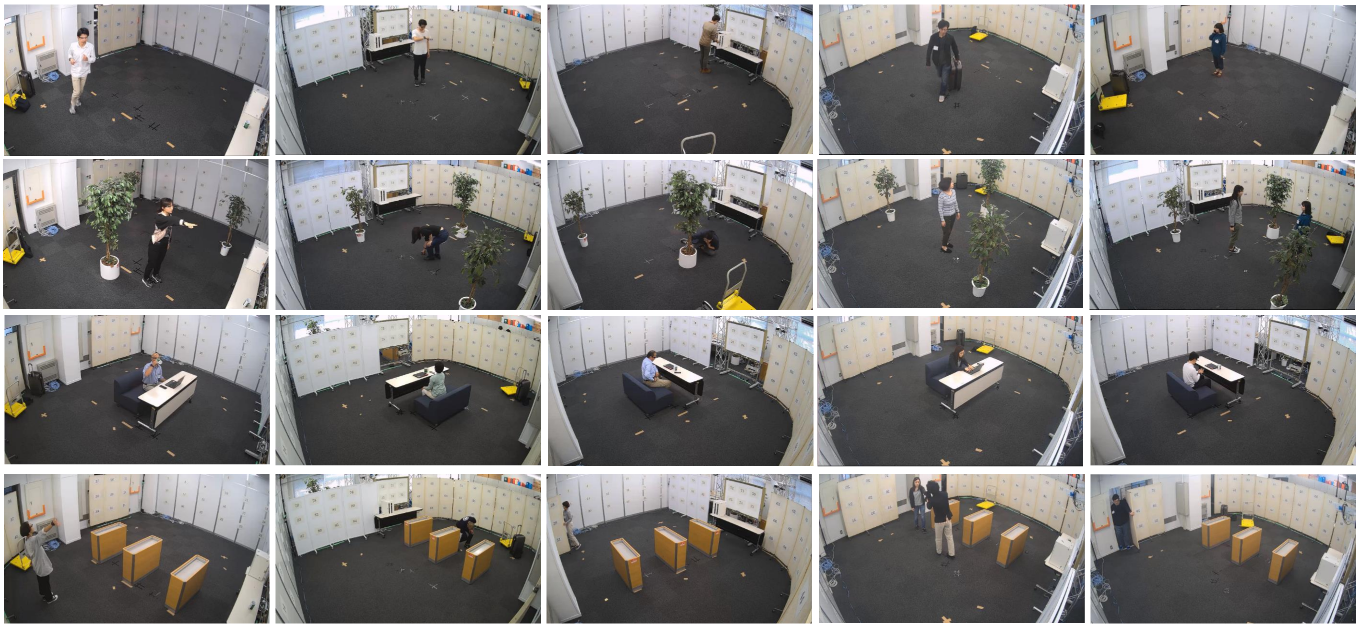

Fig. 8: More sample frames of the MMAct dataset to show the variety of our dataset. Each row shares the same scene setting. Each column shares the same camera view.

settings except cross-scene. This confirms the significance of introducing additional modalities during the training process. For the cross-scene setting, still only using sensors data achieves the best performance. This is not surprising, and confirms the robustness of sensor data against occlusions. The improvements obtained by additional support of multi modalities during training range from 2% to 8% over various settings.

We further evaluated the proposed method of knowledge distillation compared to other state-of-the-art systems in Table 3 for the cross-session setting. SVM+HOG[25] is a state-of-the-art handcraft approach trained only with RGB modality in our case. The MMAD model reaches top performance and is only second to a TSN using RGB and Optical-Flow (Fusion) as input.

In Table 4 we compare the performance of a TSN with RGB input to the MMAD model split by the most significantly improved action classes. With more than 50% of the improvement on the class ”carry light luggage” is significant. As for in our dataset, ”carry” related actions are fine-grained classes, that consist of carrying the luggage owns the same appearance but with different weight from light to heavy under the same moving route. Fig. 7 shows the example of ”carry” related action clips with their paired sensors data. Without any further modalities but only with visual information, it is difficult to distinguish ”carry light luggage” with other carry actions, like ”carry heavy luggage”. The visual input of a person moving a luggage does not give enough mutual information during training. Similar arguments hold for open, pocket out, talk on phone, etc.

5.4 Conclusion

This paper introduces a new large-scale mutlimodal dataset MMAct for action understanding. MMAct includes 36,764 action video samples collected from 37 action classes performed by 20 distinct subjects. Compared to the current datasets for multimodal action understanding, MMAct has the largest multimodal dataset that include both vision-based and sensor-based modalities. we also proposed a novel multimodality distillation model with attention mechanism, which make student network learn useful information from a teacher network with input of multiple sensor-based modalities. Experimental results under 4 different setting show our proposed method significantly improves performance of action recognition by up-to 8% compared to models without using sensor-based modalities.

References

[1] Aggarwal, J. K. and Xia, L.: Human activity recognition from 3D data: A review, *Pattern Recognition Letters*, Vol. 48, pp. 70–80 (2014).

[2] Awad, G., Butt, A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., Joy, D., Delgado, A., Smeaton, A. F., Graham, Y., Kraaij, W., Qunot, G., Magalhaes, J., Semedo, D. and Blasi, S.: TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search, *Proceedings of TRECVID 2018*, NIST, USA (2018).

[3] Aytar, Y., Vondrick, C. and Torralba, A.: SoundNet: Learning Sound Representations from Unlabeled Video, *NIPS* (2016).

[4] Bao, L. and Intille, S. S.: Activity Recognition from User-Annotated Acceleration Data, *Pervasive* (2004).

[5] Bloom, V., Makris, D. and Argyriou, V.: G3D: A gaming action dataset and real time action recognition evaluation framework, *CVPR Workshops*, pp. 7–12 (2012).

[6] Cai, Z., Han, J., Liu, L. and Shao, L.: RGB-D datasets using microsoft kinect or similar sensors: a survey, *Multimedia Tools and Applications*, Vol. 76, pp. 4313–4355 (2016).

[7] Chen, C., Jafari, R. and Kehtarnavaz, N.: UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, *ICIP*, pp. 168–172 (2015).

[8] Cheng, Z., Qin, L., Ye, Y., Huang, Q. and Tian, Q.: Human Daily Action Analysis with Multi-view and Color-Depth Data, *ECCV Workshops* (2012).

[9] Feichtenhofer, C., Pinz, A. and Wildes, R. P.: Spatiotemporal Residual Networks for Video Action Recognition, *NIPS* (2016).

[10] Feichtenhofer, C., Pinz, A. and Wildes, R. P.: Spatiotemporal Multiplier Networks for Video Action Recognition, *CVPR*, pp. 7445–7454 (2017).

[11] Feichtenhofer, C., Pinz, A. and Zisserman, A.: Convolutional Two-Stream Network Fusion for Video Action Recognition, *CVPR*, pp. 1933–1941 (2016).

[12] Gafurov, D., Helkala, K. and Søndrol, T.: Biometric Gait Authentication Using Accelerometer Sensor, *JCP*, Vol. 1, pp. 51–59 (2006).

[13] Garcia, N. C., Morerio, P. and Murino, V.: Modality Distillation with Multiple Stream Networks for Action Recognition, *ECCV* (2018).

[14] Gupta, S., Hoffman, J. and Malik, J.: Cross Modal Distillation for Supervision Transfer, *CVPR*, pp. 2827–2836 (2016).

[15] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *CVPR*, pp. 770–778 (2016).

[16] Hinton, G. E., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *CoRR*, Vol. abs/1503.02531 (2015).

[17] Hoffman, J., Gupta, S. and Darrell, T.: Learning with Side Information through Modality Hallucination, *CVPR*, pp. 826–834 (2016).

[18] Li, W., Zhang, Z. and Liu, Z.: Action recognition based on a bag of 3D points, *CVPR Workshops*, pp. 9–14 (2010).

[19] Li, Y., Lan, C., Xing, J., Zeng, W., Yuan, C. and Liu, J.: Online Human Action Detection using Joint Classification-Regression Recurrent Neural Networks, *ECCV* (2016).

[20] Lillo, I., Soto, A. and Niebles, J. C.: Discriminative Hierarchical Modeling of Spatio-temporally Composable Human Activities, *CVPR*, pp. 812–819 (2014).

[21] Liu, C., Hu, Y., Li, Y., Song, S. and Liu, J.: PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding, Vol. abs/1703.07475 (2017).

[22] Luo, Z., Zou, Y., Hoffman, J. and Fei-Fei, L.: Label Efficient Learning of Transferable Representations across Domains and Tasks, *NIPS* (2017).

[23] Natarajan, P., Wu, S., Vitaladevuni, S. N. P., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R. and Natarajan, P.: Multimodal feature fusion for robust event detection in web videos, *CVPR*, pp. 1298–1305 (2012).

[24] Ni, B., Wang, G. and Moulin, P.: RGBD-HuDaAct: A color-depth video database for human daily activity recognition, *ICCV Workshops)*, pp. 1147–1153 (2011).

[25] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R. and Bajcsy, R.: Berkeley MHAD: A comprehensive Multimodal Human Action Database, pp. 53–60 (2013).

[26] Oreifej, O. and Liu, Z.: HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences, *CVPR*, pp. 716–723 (2013).

[27] Schroff, F., Kalenichenko, D. and Philbin, J.: FaceNet: A unified embedding for face recognition and clustering, *CVPR*, pp. 815–823 (2015).

[28] Shahroudy, A., Liu, J., Ng, T.-T. and Wang, G.: NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis, *CVPR*, pp. 1010–1019 (2016).

[29] Shahroudy, A., Ng, T.-T., Gong, Y. and Wang, G.: Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, pp. 1045–1058 (2018).

[30] Simonyan, K. and Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos, *NIPS* (2014).

[31] Spriggs, E. H., la Torre, F. D. and Hebert, M.: Temporal segmentation and activity classification from first-person sensing, *CVPR Workshops*, pp. 17–24 (2009).

[32] Sung, J., Ponce, C., Selman, B. and Saxena, A.: Human Activity Detection from RGBD Images, *CoRR*, Vol. abs/1107.0169 (2011).

[33] Sung, J., Ponce, C., Selman, B. and Saxena, A.: Unstructured human activity detection from RGBD images, *ICRA*, pp. 842–849 (2012).

[34] Tran, D., Bourdev, L. D., Fergus, R., Torresani, L. and Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks, *ICCV*, pp. 4489–4497 (2015).

[35] Wang, J., Nie, X., Xia, Y., Wu, Y. and Zhu, S.-C.: Cross-View Action Modeling, Learning, and Recognition, *CVPR*, pp. 2649–2656 (2014).

[36] Wang, K., Wang, X., Lin, L., Wang, M. and Zuo, W.: 3D Human Activity Recognition with Reconfigurable Convolutional Neural Networks, *ACM Multimedia* (2014).

[37] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Gool, L. V.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, *ECCV* (2016).

[38] Wei, P., Zhao, Y., Zheng, N. and Zhu, S.-C.: Modeling 4D Human-Object Interactions for Joint Event Segmentation, Recognition, and Object Localization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, pp. 1165–1179 (2017).

[39] Wu, C., Zhang, J., Savarese, S. and Saxena, A.: Watch-n-patch: Unsupervised understanding of actions and relations, *CVPR*, pp. 4362–4370 (2015).

[40] Xia, L., Chen, C.-C. and Aggarwal, J. K.: View invariant human action recognition using histograms of 3D joints, *CVPR Workshops*, pp. 20–27 (2012).

[41] Xu, D., Ouyang, W., Ricci, E., Wang, X. and Sebe, N.: Learning Cross-Modal Deep Representations for Robust Pedestrian Detection, *CVPR*, pp. 4236–4244 (2017).

[42] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C. J., Larochelle, H. and Courville, A. C.: Describing Videos by Exploiting Temporal Structure, *ICCV*, pp. 4507–4515 (2015).

[43] Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R. and Gall, J.: A Survey on Human Motion Analysis from Depth Data, *Time-of-Flight and Depth Imaging* (2013).

[44] Zhang, J., Li, W., Ogunbona, P., Wang, P. and Tang, C.: RGB-D-based Action Recognition Datasets: A Survey, *Pattern Recognition*, Vol. 60, pp. 86–105 (2016).

[45] Zhao, M. and Katabi, D.: Through-Wall Human Pose Estimation Using Radio Signals, *CVPR*, pp. 7356–7365 (2018).