MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding

Quan Kong[1] Ziming Wu[2*] Ziwei Deng[1] Martin Klinkigt[1] Bin Tong[1,3†] Tomokazu Murakami[1]
[1] Hitachi, Ltd. R&D Group, Japan
[2]Hong Kong University of Science and Technology [3] Alibaba Group, China
{quan.kong.xz, ziwei.deng.qq, martin.klinkigt.ut, tomokazu.murakami.xr}@hitachi.com
zwual@connect.ust.hk, tongbin.tb@alibaba.com

Abstract

Unlike vision modalities, body-worn sensors or passive sensing can avoid the failure of action understanding in vision related challenges, e.g. occlusion and appearance variation. However, a standard large-scale dataset does not exist, in which different types of modalities across vision and sensors are integrated. To address the disadvantage of vision-based modalities and push towards multi/cross modal action understanding, this paper introduces a new large-scale dataset 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 to realize an adaptive knowledge transfer from sensor-based modalities to vision-based modalities. The proposed model significantly improves performance of action recognition compared to models trained with only RGB information. 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.

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 dominate the community of action understanding due to the advance of deep learning technologies [39, 27, 34]. Meanwhile, utilizing of body-worn inertial sensors, e.g. accelerator, gyroscope and orientation, to capture human motions is another typical way of realizing human action recognition [28, 22, 7]. It is well known that vision-based and sensor-based information in action recognition is complementary. To go beyond vision-only modalities which can not address vision related challenges, e.g. occlusions and appearance variation, it is considerable to leverage both vision-based and sensor-based modalities to improve performance of action understanding in multimodal [26, 10, 20] and cross-modal [38, 3, 19] manners.

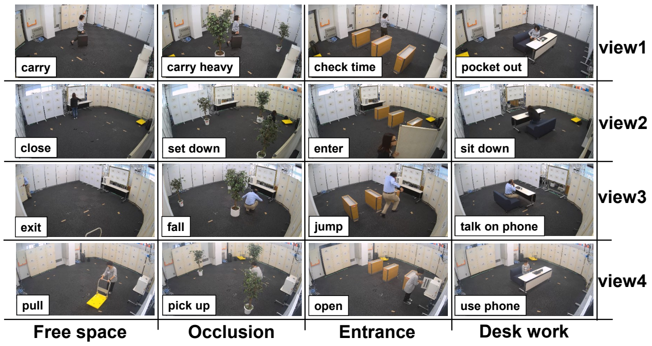

Figure 1. The illustration of our dataset. Each column shows actions under a scenario. Each row denotes the action under one of four camera views.

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 [25, 17] focusing on 3D human action recognition or detection. However, only three to four vision related

[*] Work is done during internship at Hitachi.
[†] Work is done at Hitachi. Current at Alibaba.

modalities are provided in the existing datasets. Second, there is the limited number of supported action understanding tasks 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 variance under the camera view. Forth, there is the limited number of instances for each modality with distinct subject, scenario, view and session in a factored data structure, especially for crossmodal related researches. This paper proposes a new multimodal dataset to overcome the above limitations, especially for expanding the crossmodal research on human action understanding.

Our dataset, named as multimodal action dataset (MMAct), consists of 36,000+ trimmed clips with seven types of modalities captured from 20 subjects, which include RGB videos, keypoints, acceleration, gyroscope, orientation, Wi-Fi and pressure signal. MMAct is designed under a semi-natural data collection protocol [4] that a random walk is performed between the end of current action and the start of next action. The action is only performed after a start sign given from the outside monitor. This protocol makes sure that the action will be occurred randomly in the action area to provide various 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, i.e. knowledge distillation [12], is a useful way to allow a model with only one modality input to achieve the performances close to the use of multiple modalities. For example, a student model with RGB input learns complementary information from other modalities, e.g. depth [13], 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 trained with only RGB information.

Different from the existing methods that focus on modality transfer across 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 multi modality 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 multi modality distillation model with attention mechanism. This model has a student network with input of RGB information, which learns useful side 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 where RGB modality may fail to recognize the actions.

2. Related Work

In this section, we illustrate 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, 40, 41].

2.1. Related Datasets

Some traditional and typical multimodal datasets for action understanding are discussed below, with a comparison between them and MMAct in Table 1.

MSR-Action3D [14] 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 [35] and Northwestern-UCLA [32] 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 [25] and it's extension [18] are the state-of-the-art large-scale benchmarks for 3D human activities analysis. NTU RGB+D 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. 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 [36] and Compostable Activities [16] 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 [17] is a large-scale benchmark for human action detection, which has a 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 [28] is a multi modality human activity dataset combining vision modalities with sensor signals, in-

Table 1. Comparison between different multimodal datasets for action understanding. 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 [14]	20	567	10	1	1	D+Keypoints	No	No	No	2010
CAD-60 [29]	12	60	4	5	-	RGB+D+Keypoints	No	No	No	2011
RGBD-HuDaAct [21]	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 [37]	10	200	10	1	4	RGB+D+Keypoints	No	No	No	2012
3D Action Pairs [23]	12	360	10	1	1	RGB+D+Keypoints	No	No	No	2013
Multiview 3D Event [35]	8	3815	8	1	3	RGB+D+Keypoints	No	No	No	2013
Northwestern-UCLA [32]	10	1475	10	1	1	RGB+D+Keypoints	No	No	No	2014
Office Activity [33]	20	1180	10	-	3	RGB+D+Keypoints	No	No	No	2014
NTU RGB+D [25]	60	56880	40	1	80	RGB+D+Keypoints+IR	No	No	No	2016
G3D [5]	20	1467	10	1	-	RGB+D+Keypoints	Yes	No	No	2012
CAD-120 [30]	20	1200	4	1	-	RGB+D+Keypoints	Yes	No	No	2013
Compostable Activities [16]	16	2529	14	1	1	RGB+D+Keypoints	Yes	No	No	2014
Watch-n-Patch [36]	21	2500	7	13	-	RGB+D+Keypoints	Yes	No	No	2015
OAD [15]	10	700	-	1	1	RGB+D+Keypoints	Yes	No	No	2016
PKU-MMD [17]	51	21545	66	1	3	RGB+D+IR+Keypoints	Yes	No	No	2017
CMU-MMAC [28]	5	186	39	1	5	RGB+D+Keypoints+Acc+Mic	No	No	No	2010
MHAD [22]	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+Presure**	**Yes**	**Yes**	**Yes**	**2019**

cluding 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 [22] 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 scene, 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 relatively new topic due to the development of deep learning technology and hardware such as depth cameras and wearable devices. There are some typical ideas of dealing with multi modality data. The two-stream architecture introduced by [27] has been widely developed in several works. [31] proposed a 3D ConvNets for extracting spatiotemporal features to model appearance and motion information simultaneously. [26] designed a deep auto-encoder architecture to decompose its multimodal 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. [10] 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. 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

The concept most related to our work is the transfer learning across different modalities. While conventional transfer learning works only focus on category-level knowledge transfer, crossmodal transfer works devote to modality shift, which transfers knowledge learned in one data modality to another. [13] proposed a modality hallucination architecture to mimic the depth mid-level features to enhance an RGB object detection model. [38] 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, [42] 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 due to 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[1] is a novel large-scale dataset focusing on action recognition/detection tasks and crossmodal action analysis. We collected 36,000+ temporally localized action instances in 1,900+ 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.

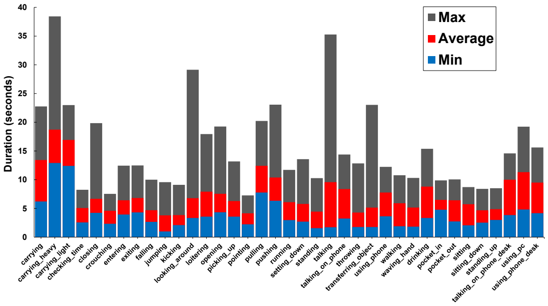

Figure 2. Average length of trimmed action clip per class. Overall there is high variation of the duration among each action.

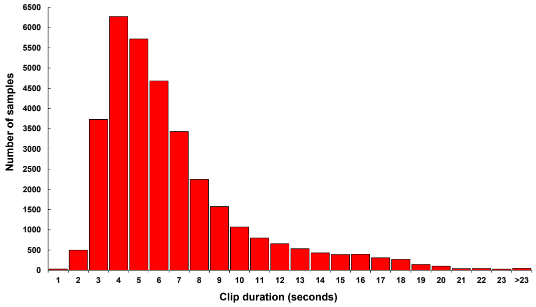

Figure 3. Distribution of the trimmed action clip length. Most samples are in a range from about 3 sec. to 8 sec.

3.1. Data Modalities

Seven types of modality are provided with the MMAct dataset: RGB videos, keypoints, acceleration, gyroscope, orientation, Wi-Fi and pressure signal.

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 smart glass (Google Glass) to record egocentric videos with a resolution of 1280 × 720 at 30 FPS to support action recognition research in this direction.

A smartphone (ASUS ZenPhone AR) installed with some initial sensors, such as accelerator and gyroscope, was used to obtain data of acceleration, gyroscope, orientation, Wi-Fi and pressure signal. The smartphone was carried and put inside the pocket of the subject's pants. 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 and 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 Figure 2, which illustrates that each action class has plenty of distinct samples with high variation in our dataset. All the classes we collected are illustrated in the horizontal axis of Figure 2. Figure 3 shows the distribution of number of samples for different clip duration, illustrating that we have a large number of sequences among different duration and most sequences last 3∼8 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, entrance 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 entrance scene, 3 gates like objects were set in parallel with a space large enough 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 desk 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

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

one was recorded from the egocentric view by wearing the smart 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 almost 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.

3.3. Data Collection

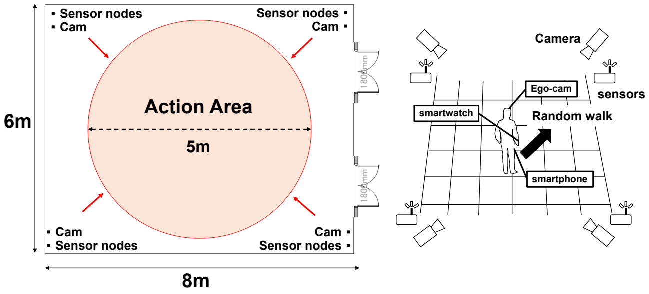

Figure 4. The environmental setup of the action area showing the size and location of the cameras and sensors.

Generally, collecting untrimmed data for action recognition is a difficult task. The recording environment and process must be appropriately designed and temporal boundaries must be controlled. 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 Figure 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 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 Figure 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.

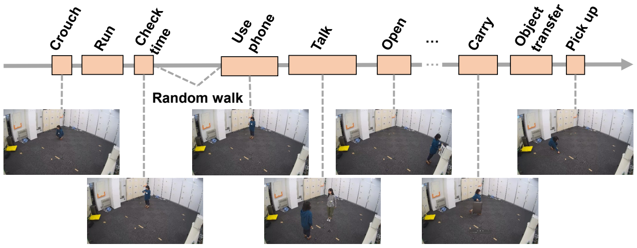

Figure 5. Sample of our collected action sequence.

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 (KD) [12] as our preliminary in advance. The idea of KD 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:

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

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 Figure 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.

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 its corresponding label, N_t 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, 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 $R_i = arcsin(\frac{z_i}{\sqrt{x^2{}_i + y^2{}_i + z^2{}_i}})$ [9], 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 [24] to train the teacher models along with the cross-entropy loss for classification. 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. Thus we want, $||T_m(g^a_{ij}) - T_m(g^p_{ij})||^2_2 + \alpha < ||T_m(g^a_{ij}) - T_m(g^n_{ij})||^2_2$, where α is used as a margin to enforce the anchor to be closer to the positive than negative samples. The triplet loss in our model that is being minimized is then $L_{\mathrm{t}} =$

$$\Sigma[||T_m(g^a_{ij}) - T_m(g^p_{ij})||^2_2 - ||T_m(g^a_{ij}) - T_m(g^n_{ij})||^2_2 + \alpha] \tag{2}$$

where $T_m(g_{ij})$ represents the semantic embedding from teacher model T_m. We use offline triplet mining to ensure the positive segment of a specific action from the other subject included in each batch.

Multi modality attention distillation. Let $D_s = \{(x_i, y_i)\}^s_{i \in N_t}$ denote the training set for the student modality s. Our student network is a TSN [34] 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 the attention layer consists of a four-layer feed-forward neural network with three convolutional layers and one FC layer activated by ReLU functions to nonlinearly project the concatenated semantic codes F_{ij} from each teacher into a common subspace used for representing attention weights as $(w^1_{ij}, ..., w^m_{ij}, m \in N_m)$, with softmax regression.

The ensemble layer is used to aggregate each weighted semantic codes $w^m_{ij} T_m(g^m_{ij})$ from ith action clip's multiple teachers and output an ensemble soft target $\widehat{z_i}$ as follows,

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

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 $L_D = H(\widehat{z_i}, \widehat{s_i})$, where H refers to the cross entropy, that $H(\widehat{z_i}, \widehat{s_i}) = -\sum_i \widehat{z_i} log(\widehat{s_i})$, $\widehat{s_i}$ represents class probability prediction of student network. The student network loss 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 minimizing the distillation loss and ensemble teacher classification loss simultaneously:

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

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

5. Evaluations

5.1. Evaluation Setting

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

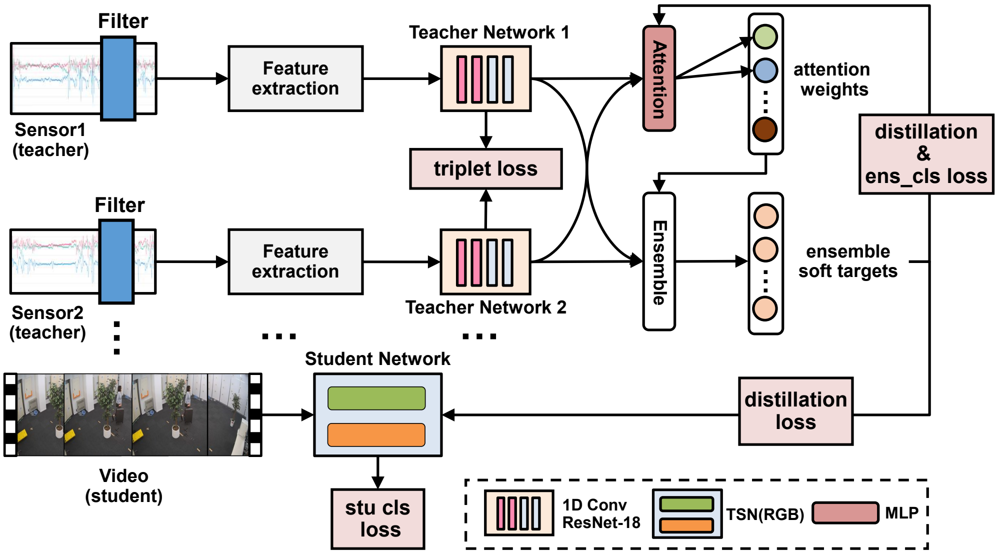

Figure 6. Architecture of our proposed multi modality attention distillation learning framework. We first train the teacher model separately on its corresponding modality, each teacher model is a 1D Convolutional Neural Network (CNN). Then we use the semantic embedding from the output of softmax layer as the teacher information of corresponding modality in trained teacher model. As for the softmax layer where the influence of domain gap 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 teacher 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.

Cross-Subject: samples from 80% of the subjects (subject id from 1 to 16) 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 (right upper in Figure 4) 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 top-80% sessions in ascending order of session id for each subject have been used for training the model and the remaining sessions for testing.

Out of these settings, Cross-Subject is typically applied on action classification works to confirm the realistic variation of methods for different subjects. 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 would be typical challenges. Cross-Session is a standard setting, 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**: our proposed multi modality attention distillation method. We used 1D conv ResNet-18[11] 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 about 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 some challenge settings for vision-only modality, such as the cross-scene setting.

Introducing accelerator sensor data to the training process improves the performance of the SMD model in most settings, with the cross-view setting the most significant improvement of almost 4.1%. Increasing the number of modalities for the MMD model even further, still improves the performance, but not as significantly as with the intro-

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[12]	Acc+RGB	RGB	63.89	66.31	61.56	71.23
MMD	Acc+Gyo+Ori+RGB	RGB	64.33	68.19	62.23	72.08
MMAD	**Acc+Gyo+Ori+RGB**	**RGB**	**66.45**	**70.33**	**64.12**	**74.58**

Table 3. Proposed method compared with the vision modality based methods under Cross-Session evaluation.

Method	Precision	Recall	F-measure
SVM+HOG[22]	45.31	47.81	46.52
TSN(RGB)[34]	68.32	70.11	69.20
TSN(Optical-Flow)[34]	71.89	73.27	72.57
TSN(Fusion)[34]	75.68	78.57	77.09
MMAD	**73.34**	**75.67**	**74.58**
MMAD(Fusion)	**77.58**	**80.12**	**78.82**

duction of the first additional modality. In the proposed model MMAD, a more significant improvement in performances while utilizing the same modalities in training and testing as the MMD model.

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 settings except cross-scene due to the intended setting of cross-scene emphasizing visual distortion caused by occlusion. The result sheds light on incorporating body-worn sensor modalities for improving human action recognition in the wild with vision-only modalities. The improvements obtained by additional support of multi modalities during training range from about 2% to 8% over various settings.

We further evaluated the proposed method of knowledge distillation compared to other state-of-the-art methods in Table 3 for the cross-session setting. SVM+HOG[22] 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(OF) as input. We also examine our approach with TSN(Fusion) as the student of MMAD(Fusion). In this case, RGB and OF networks are trained separately with MMAD and then, we fuse the results from the trained RGB and OF networks to produce the final prediction of MMAD(Fusion) method. These results further verify the effectiveness of the proposed method.

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.

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)[34]	11.12	28.41	31.57	61.53	48.79
MMAD	**64.51**	**78.67**	**52.63**	**81.31**	**65.30**

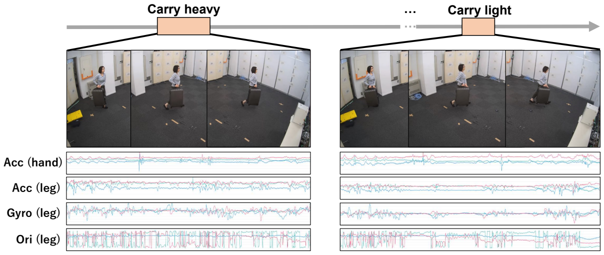

Figure 7. Sample clips with their paired sensors data related to action ”carry heavy luggage” and ”carry light luggage”.

In MMAct, carry related actions are designed to be composed of carrying the luggage with the same appearance but different weight from light to heavy. Figure 7 shows the example of ”carry” related action clips with their paired sensors data. Without any further modalities it is difficult to distinguish class ”carry light” 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. Compared to the current datasets for multimodal action understanding, MMAct has the largest number of modalities include both vision-based and sensor-based modalities. We also proposed a novel multi modality distillation model with attention mechanism, which makes student network with input of RGB learn useful information from a teacher network trained by multiple sensor-based modalities. Experimental results under 4 different setting show the availability of MMAct on the potential of cross modal action understanding across vision and sensors modalities.

References

[1] Jake K. Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[2] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Qunot, Joao Magalhaes, David Semedo, and Saverio Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016.

[4] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *Pervasive*, 2004.

[5] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *CVPR Workshops*, pages 7–12, 2012.

[6] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao. Rgb-d datasets using microsoft kinect or similar sensors: a survey. *Multimedia Tools and Applications*, 76:4313–4355, 2016.

[7] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *ICIP*, pages 168–172, 2015.

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

[9] Davrondzhon Gafurov, Kirsi Helkala, and Torkjel Søndrol. Biometric gait authentication using accelerometer sensor. *JCP*, 1:51–59, 2006.

[10] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *ECCV*, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[12] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[13] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *CVPR*, pages 826–834, 2016.

[14] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *CVPR Workshops*, pages 9–14, 2010.

[15] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. Online human action detection using joint classification-regression recurrent neural networks. In *ECCV*, 2016.

[16] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *CVPR*, pages 812–819, 2014.

[17] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. abs/1703.07475, 2017.

[18] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[19] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li Fei-Fei. Label efficient learning of transferable representations across domains and tasks. In *NIPS*, 2017.

[20] Pradeep Natarajan, Shuang Wu, Shiv Naga Prasad Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, pages 1298–1305, 2012.

[21] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *ICCV Workshops)*, pages 1147–1153, 2011.

[22] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. pages 53–60, 2013.

[23] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, pages 716–723, 2013.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[25] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016.

[26] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1045–1058, 2018.

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

[28] Ekaterina H. Spriggs, Fernando De la Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPR Workshops*, pages 17–24, 2009.

[29] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. *CoRR*, abs/1107.0169, 2011.

[30] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Unstructured human activity detection from rgbd images. In *ICRA*, pages 842–849, 2012.

[31] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[32] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *CVPR*, pages 2649–2656, 2014.

[33] Keze Wang, Xiaolong Wang, Liang Lin, Meng Wang, and Wangmeng Zuo. 3d human activity recognition with reconfigurable convolutional neural networks. In *ACM Multimedia*, 2014.

[34] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[35] Ping Wei, Yibiao Zhao, Narming Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1165–1179, 2017.

[36] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *CVPR*, pages 4362–4370, 2015.

[37] Lu Xia, Chia-Chih Chen, and Jake K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, pages 20–27, 2012.

[38] Dong Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *CVPR*, pages 4236–4244, 2017.

[39] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Joseph Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015.

[40] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. A survey on human motion analysis from depth data. In *Time-of-Flight and Depth Imaging*, 2013.

[41] Jing Zhang, Wanqing Li, Philip Ogunbona, Pichao Wang, and Chang Tang. Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, 2016.

[42] Mingmin Zhao and Dina Katabi. Through-wall human pose estimation using radio signals. In *CVPR*, pages 7356–7365, 2018.