A Large Scale RGB-D Dataset for Action Recognition

Jing Zhang[1], Wanqing Li[1], Pichao Wang[1], Philip Ogunbona[1], Song Liu[1], and Chang Tang[1,2]

[1] School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia,
jz960@uowmail.edu.au, wanqing@uow.edu.au, pw212@uowmail.edu.au, philipo@uow.edu.au, songl@uow.edu.au

[2] School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China,
tangchang@wust.edu.cn

Abstract. Human activity understanding from RGB-D data has attracted increasing attention since the first work reported in 2010. Over this period, many benchmark datasets have been created to facilitate the development and evaluation of new algorithms. However, the existing datasets are mostly captured in laboratory environment with small number of actions and small variations, which impede the development of higher level algorithms for real world applications. Thus, this paper proposes a large scale dataset along with a set of evaluation protocols. The large dataset is created by combining several existing publicly available datasets and can be expanded easily by adding more datasets. The large dataset is suitable for testing algorithms from different perspectives using the proposed evaluation protocols. Four state-of-the-art algorithms are evaluated on the large combined dataset and the results have verified the limitations of current algorithms and the effectiveness of the large dataset.

Keywords: Large scale RGB-D dataset, action recognition, evaluation protocol

1 Introduction

Action recognition has wide applications including human-computer interaction, video surveillance, health monitoring, and content-based video retrieval. The introduction of low-cost integrated depth sensors (such as Microsoft Kinect ™) that can capture both RGB (red, green and blue) video and depth (D) information has significantly advanced the research of human action recognition. More than 40 publicly available RGB-D action datasets have been collected during the past several years to simulate specific real world applications. According to Zhang et al. [18], the state-of-the-art results obtained on many of these existing RGB-D-based action datasets are nearly perfect. However, do the nearly perfect

results really reflect the robustness of these algorithms? As pointed in [18], most of current datasets only consist of a small number of action classes with restricted types of actions and small sample sizes. In addition, most RGB-D datasets are collected in laboratory environment and the execution style of actions generally follow strict instructions, suggesting that even with different subjects, the variations in performing style are subtle and maybe indiscernible. Hence, the gap between the laboratory collected datasets and real world scenarios is still large. This impedes the development of higher level action recognition algorithms for real world applications. We hypothesise that the algorithms designed for these datasets would likely overfit on the individual datasets. However, the real world applications require the algorithms to be able to generalize well across subjects, backgrounds, view angles, and other environmental factors. For example, a well trained video surveillance system in laboratory environments should be directly usable in the customer environments without the need of labelling new data and retraining the whole system afresh. A well trained content-based action video retrieval system should be able to retrieve unseen action videos with reasonable accuracy, where the unseen action videos can be captured with different sensor types, resolutions, environments, and performed by different subjects at different view angles.

To narrow the gap between current action datasets and realistic scenarios, we create a large scale dataset by combining several existing publicly available datasets and propose a set of evaluation protocols. The combined dataset will be useful to the research community in many respects. First, it aggregates different individual datasets wherein the action execution manners, backgrounds, acting positions, view angles, resolutions, and sensor types are different. In addition, as a large number of action classes are involved, it will be more challenging since the variation between classes will be smaller compared to that of the individual datasets. The large dataset can also be used for testing the scalability of algorithms by randomly selecting a subset of actions from the combined dataset which contains a large choice of distinctive actions. More importantly, as the combined dataset contains many small datasets, some actions exist in more than one small dataset and this allows easy accommodation of cross-dataset evaluations. In cross-dataset set-up, the actors, environment and manner of performing actions in training and test data are all different. Lastly, the large scale dataset will also enable the community to apply, develop and adapt various data-driven and data-hungry learning techniques, such as deep learning.

Based on these arguments, we provide the structure of the large dataset and propose a set of evaluation protocols. We also provide the settings of the large dataset to test the algorithms from both real world applications and algorithm development perspectives using specific protocols. In addition, four state-of-the-art algorithms are evaluated on the large combined dataset using different settings to verify the limitations of current algorithms and the effectiveness of the large dataset.

2 Structure of the Dataset

A key step in the creation of the large dataset is to merge semantically and visually similar action samples from different individual datasets into one action. Currently, we have combined 9 single-view RGB-D action datasets into one with 94 actions. The large dataset can be expanded easily by merging more individual datasets. All the 9 datasets are publicly available. Hence, the combined dataset can be easily produced according to the structure provided in Table 5. The combined dataset with unified data format and naming convention will also be released to the public with the permission of the authors of the original datasets. The researchers who would use the combined dataset should cite this paper and the papers of the original datasets.

2.1 Statistics

Table 1 gives the statistics of the individual datasets included in the large dataset to date and the statistics of current combined dataset.

The detailed structure of the combined dataset is shown in Table 5. The first and second column are the action labels and action names of the combined dataset. The third to fifth column give the sample numbers of different data modalities in each action. Column six is the number of datasets that have the actions. The rest of the columns show which datasets the action samples come from.

Table 1. Statistics of individual datasets and combined dataset. Dataset notation: D1: MSRAction3DExt [8, 14, 13]; D2:UTKinect [15]; D3:DailyActivity [12]; D4:ActionPair [10]; D5:CAD120 [7]; D6:CAD60 [11]; D7:G3D [2, 1]; D8:RGBD-HuDa [9]; D9:UTD-MHAD [4]. Notation for the header: #a: number of actions; #s: number of subjects; #e: number of total examples. Notation for data modalities: C: Colour; D: Depth; S: Skeleton; I: Inertial sensor data. Notation for protocols: CSub: Cross Subject; LOSeqO: Leave One Sequence Out; LOSubO: Leave One Subject Out.

Dataset	Sensor	Modalities	Resolution(C,D)	#a,#s,#e	Protocol
D1 [8, 14, 13]	Kinect v1	D,S	-,640*480	20,23,1369	CSub
D2 [15]	Kinect v1	C,D,S	640*480,320*240	10,10,200	LOSeqO
D3 [12]	Kinect v1	C,D,S	640*480,320*240	16,10,320	CSub
D4 [10]	Kinect v1	C,D,S	640*480,320*240	12,10,360	CSub
D5 [7]	Kinect v1	C,D,S	640*480,640*480	10,4,120	LOSubO
D6 [11]	Kinect v1	C,D,S	320*240,320*240	12,4,60	LOSubO
D7 [2, 1]	Kinect v1	C,D,S	640*480,640*480	20,10,659	CSub
D8 [9]	Kinect v1	C,D	640*480,640*480	12,30,1189	LOSubO
D9 [4]	Kinect v2	C,D,S,I	640*480,320*240	27,8,861	CSub
Combined	v1 or v2	C,D,S	multiple,multiple	94,107,4953	See Section 3

2.2 Properties

As shown in Tables 1 and 5, the combined dataset has the following properties;

- Large size: The dataset is large in terms of number of actions, number of subjects, and number of video sequences.
- Variations: As the large dataset consists of different small datasets, the variations are very large with respect to subjects, backgrounds, environments, execution rates, performing manners, resolutions, and sensors, etc.
- Imbalance: The issue of data imbalance is common in real world applications. For example, the abnormal actions (kick, punch, or fall down) are generally much rarer than normal actions in surveillance or health monitoring systems and generally require higher recognition rate. After the combination, the number of samples for each action is different, which allows the development of algorithms addressing the data imbalance issue.

3 Evaluation Protocols

According to the properties of the combined dataset, we propose four evaluation protocols to take advantage of the large combined dataset from different perspectives. The four protocols are *Random Cross Subject*, *Random Cross Subject Balanced*, *Random Cross Sample*, and *Random Cross Sample Balanced*, which can be used for different real world applications, such as human-computer interaction, video surveillance, health monitoring, and content-based video retrieval. Besides the real world applications, the combined dataset can also be used for algorithm development using specific protocols.

Below, we provide the detailed description of each protocol and corresponding evaluation metric, properties and applications, as well as the settings on how to use the combined dataset for simulating each application. The researchers can choose to test their algorithms freely from any aspect(s) on the combined dataset by following different settings.

3.1 Random Cross Subject

- Protocol: Half of the subjects are randomly selected as training data and the rest subjects are used as test data.
- Evaluation metric: Precision & Recall are used because the data are unbalanced when using this protocol on the combined dataset.
- Properties:
  1. Each action is performed by a different set of subjects because not all the subjects performed all the actions in the combined dataset.
  2. The variations among subjects can be evaluated, since the test subjects are unseen in the training stage.
  3. As the combined dataset contains different individual datasets, the training and test subjects have a certain possibility to be from different individual datasets. Thus, the variations among datasets can be evaluated.

4. The numbers of samples among actions are different due to the combination of different datasets. The data imbalance among classes allows the development of algorithms addressing the data imbalance issue.

– Applications: This protocol can be used on the real world scenarios that require the subjects in test data not to appear in the training stage, such as human computer interaction, health monitoring and video surveillance. The content-based video retrieval can also use this protocol.
– Settings:
  • Human Computer Interaction(HCI): In HCI,the actions are generally with relatively low complexity and short duration. There are usually no interactions with other objects when the actions are meant to control or interact with computers. In the combined dataset, the set of actions with above properties, such as actions a001-a020, a067-a075, and a082-a094, can be chosen to simulate this scenario.
  • Health Monitoring (HM) and Video Surveillance (VS): In HM and VS, the movements are more complex and with more variations in speed and style compared with that in HCI. Some more challenging factors, such as view angle variation, human-object interaction, and body part occlusion, are involved in HM and VS system. Hence, actions like a021-a066 and a076-a081 in the combined dataset can be selected to simulate HM or VS scenarios.
  • Video Retrieval (VR): In VR scenarios, all the challenges in action recognition can potentially occur. Videos can be recorded in different environments and involve different persons. The actions contained in videos can be performed in different manners and observed from different view angles. In addition, the action types are various and can be both HCI and video surveillance related actions. The combined dataset can satisfy the conditions because it contains multiple individual datasets with different properties. All the actions in the combined dataset can be used jointly to simulate the VR scenarios.

3.2 Random Cross Subject Balanced

– Protocol: Half of the subjects are randomly selected as training data and the rest subjects are used as test data. The subjects in both training set and test set are further down sampled to the smallest number of subjects among all the actions.
– Evaluation metric: Accuracy.
– Properties:
  1. Same as property 1 of *Random Cross Subject* protocol in Section 3.1.
  2. Same as property 2 of *Random Cross Subject* protocol in Section 3.1.
  3. Same as property 3 of *Random Cross Subject* protocol in Section 3.1.
  4. The numbers of samples among actions are approximately the same by the procedure of down sampling.
– Applications: Previous action recognition algorithms are generally designed for balanced data. Hence, this protocol can be used for algorithm development, such as evaluating the scalability of algorithms.

– Settings:
  • Algorithm Scalability (AS): The AS can be evaluated by randomly selecting some subsets with different number of actions from the combined dataset and the number of actions increases monotonically in these subsets. For example, 20 actions can be randomly selected to form the first subset, then add 20 more actions to form the second subset with 40 actions in total, and so forth.

3.3 Random Cross Sample

– Protocol: For each action, half of the samples are randomly selected as training data while the rest of the samples are used as test data.
– Evaluation metric: Precision & Recall.
– Properties:
  1. The test samples are unseen in the training stage.
  2. The training data are from more subjects than that in *Random Cross Subject* protocol.
  3. The training and test samples have a certain possibility to be from different individual datasets.
  4. The numbers of samples among actions are different.
– Applications: The real world scenarios that do not necessarily require the test samples to be constrained to specific subjects can apply this protocol; for example video retrieval.
– Settings:
  • Video Retrieval (VR): We can simulate VR system by using all the actions in the combined dataset. Compared with VR in *Random Cross Subject* protocol, more subjects can be involved to train the model using this protocol.

3.4 Random Cross Sample Balanced

– Protocol: For each action, half of the samples are randomly selected as training data while the rest of the samples are used as test data. The samples in both training set and test set are further down sampled to the smallest number of samples among all the actions.
– Evaluation metric: Accuracy.
– Properties:
  1. Same as property 1 of *Random Cross Sample* protocol in Section 3.3.
  2. The samples selected are from more subjects than that in *Random Cross Subject Balanced* protocol.
  3. Same as property 3 of *Random Cross Sample* protocol in Section 3.3.
  4. The numbers of samples among actions are approximately the same by using the procedure of down sampling.
– Applications: This protocol can also be used for evaluating algorithm scalability because of the balanced data.
– Settings:
  • Algorithm Scalability: The settings are similar as that of *Random Cross Subject Balanced* protocol in 3.2. The only difference is the samples used for training and test in each subset.

4 Experiments

To verify our hypothesis on the limitations of current algorithms, we conducted several experiments on the current version of the combined dataset. Four state-of-the-art algorithms are chosen to be evaluated on the combined dataset. Two of them are designed for depth data only: one is global feature-based algorithm, namely SNV [17], and the other is local feature-based algorithm, namely DSTIP+DCSF [16]. The third method uses both depth and skeleton data, namely local HON4D [10], which extracts features within the depth cuboids centred at each skeleton joint. The last method is designed on skeleton data only, namely dynamic skeleton(DS) [6].

4.1 Settings

Table 2 and Table 3 give the experimental settings of each test on the four methods. The difference between the settings of Table 2 and Table 3 is that the samples without skeleton modality are removed in Table 3, since the local HON4D and DS methods use the skeleton data to extract features. From Table 1, it can be seen that only dataset D8 (RGBD-HuDa) does not have skeleton data. Hence, in the experiments of Table 3, samples from RGBD-HuDa dataset are removed.

The settings are based on the proposed evaluation protocols and corresponding applications as described in Section 3. Table 2 and Table 3 give the information of actions involved, protocols and applications, total number of actions, and total number of video samples in each test. For example, in the experiments of SNV and DSITP methods (see Table 2), forty-two HCI related actions in the large dataset are selected, resulting in 2844 action samples in total to evaluate the algorithms on the application of HCI with the *Random Cross Subject* protocol. The action labels of the selected actions can be found in the second column of Table 2 and the corresponding actions are listed in Table 5. For the application of VS/HM, fifty-two daily life actions are selected also using the *Random Cross Subject* protocol. All the actions are selected to simulate the VR scenarios and the algorithms are evaluated using both the *Random Cross Sample* and *Random Cross Subject* protocol, respectively. We test the algorithm scalability of the four algorithms by setting the incremental quota between two subsets to be 20 actions. In the experiments on AS, eight subjects per action are randomly selected when using the *Random Cross Subject Balanced* protocol and 30 samples per action are randomly selected when using the *Random Cross Sample Balanced* protocol to approximately balance the numbers of samples across action classes. For the classes with less than 8 subjects or 30 samples, we just use data of all the subjects or all the samples. To make the results repeatable and comparable, we fix the random seed as default in Matlab for all the experiments.

For all the evaluated methods, we use the codes provided by the authors and select the parameters using cross validation. We use LIBLINEAR [5] SVM for classification on SNV, local HON4D, and DS methods, and use LIBSVM [3] with

histogram intersection kernel on DSTIP+DCSF method as used in the original papers. The parameters of the four methods are as follows.

SNV: we set a 9×3 neighborhood for each cloud point to form the polynormal, use 100 visual words in the sparse coding, and the spatio-temporal pyramid to be $4 \times 3 \times 3$ space-time grids in height, width, and time, respectively.

DSTIP+DCSF: we set the spatial scale of the filter to be $\sigma = 5$, the temporal scale of the filter to be $\tau = T/17$ (T denotes the duration of the action sequence), the number of interest points to be $N_p = 500$, the number of voxels for each cuboid to be $n_{xy} = 4$, $n_t = 2$, the cuboid spatial size to be adapted according to the depth value of the cuboid $\Delta_x^{(i)} = \Delta_y^{(i)} = \sigma \frac{L}{d^{(i)}}$ with $L = 6$ be the support region size and $d^{(i)}$ be the depth pixel value of the i-th cuboid, the codebook size to be $k = 1500$ in the bag-of-codewords.

Local HON4D: we extract descriptors around 15 skeleton joints by following the process similar to [12, 10]. The selected joints include head, neck, left knee, right knee, left elbow, right elbow, left wrist, right wrist, left shoulder, right shoulder, hip, left hip, right hip, left ankle, and right ankle. We use a patch size of $24 \times 24 \times 4$ for depth map with resolution of 320×240 and $48 \times 48 \times 4$ for depth map with resolution of 640×480 to reduce the effects of different depth map sizes in the combined dataset, then divide the patches into a $3 \times 3 \times 1$ grid.

Dynamic Skeleton(DS): we select 15 joints to extract dynamic skeleton feature. The selected joints are the same as that of local HON4D method.

4.2 Results

Table 2 and Table 3 also give the results of the four algorithms on corresponding protocols and applications. We could not evaluate the SNV methods on the applications of VR due to memory issue.

We firstly analyse the results obtained on SNV and DSTIP+DCSF methods shown in Table 2. Since the actions in the combined dataset are without much occlusion and changes in viewpoint, the advantage of global representations increases, which leads to the generally lower results obtained by DSTIP+DCSF than SNV. However, the results obtained from the large dataset are lower than that from individual datasets [17, 16]. The locations of actors and backgrounds have large variations among individual datasets in the combined dataset. Based on the characteristics of the algorithm and the experimental results, we conjecture that the SNV algorithm may be sensitive to the background of the data or the locations of the actors due to the global representation. By contrast, DSTIP+DCSF is an interest points-based representation. It extracts spatio-temporal interest points first and local cuboid is calculated around these points. Although local representations are less sensitive to the static background and location of actors, DSTIP+DCSF still performed poorer than SNV. This is because DSTIP+DCSF rely on more parameters to extract discriminative feature points and remove noise; one set of parameters is not suitable for all individual datasets in the combined dataset. The scalability of the two algorithms is both

Table 2. Results of SNV and DSTIP on combined dataset. Notation for the header: #A: number of actions; #S: number of total samples; Prec.: Precision; Rec.: Recall; Acc.: Accuracy. Notation for the protocols: RCSub: *Random Cross Subject*; RCSap: *Random Cross Sample*; RCSubB: *Random Cross Subject Balanced*; RCSapB: *Random Cross Sample Balanced*.

Test	Actions	Protocols (Applications)	#A,#S	SNV		DSTIP	
				Prec.	Rec.	Prec.	Rec.
1	a001-a020, a067-a075, a082-a094	RCSub (HCI)	42,2844	82.9%	81.6%	58.7%	54.6%
2	a021-a066, a076-a081	RCSub (VS/HM)	52,2109	63.9%	67.7%	55.3%	54.0%
3	All 94	RCSap (VR)	94,4953	-	-	70.8%	65.5%
4	All 94	RCSub (VR)	94,4953	-	-	59.5%	53.8%

Test	Actions	Protocols (Applications)	#A,#S	SNV	DSTIP
				Acc.	Acc.
5	Random 20	RCSubB (AS)	20,453	89.8%	80.9%
	Random 40		40,891	75.9%	71.4%
	Random 60		60,1337	68.6%	60.3%
	Random 80		80,1793	58.4%	55.2%
6	Random 20	RCSapB (AS)	20,513	95.7%	91.0%
	Random 40		40,989	89.1%	77.1%
	Random 60		60,1527	86.9%	74.0%
	Random 80		80,2053	81.3%	67.1%

poor (Test 5 and Test 6) because when the number of actions increases, the overlap between classes is higher. The lower results on Test 5 than that on Test 6 show the large variations among subjects in the combined dataset.

Thre results of local HON4D and DS methods are shown in Table 3. Compared to SNV and DSTIP+DCSF, local HON4D and DS methods used the skeleton data. The benefit of using skeleton data is the accurate locations of subjects. Hence, the effects of different locations of subjects among datasets can be reduced. However, the SNV method is superior compared to local HON4D on most of the tests. This is because the polynormals extracted by SNV are more robust to noise and the higher level features obtained by sparse coding further improve the results. Compared to DSTIP+DCSF, local HON4D extracts features on the cuboid of depth data around each skeleton joint rather than the interest points extracted from the depth map, which is more stable across subjects and datasets. Thus, the results of local HON4D are better than DSTIP+DCSF on almost all the settings. However, the effects of noise on depth data and different backgrounds among datasets still exist in HON4D method, resulting in poorer results than the skeleton-based feature (DS). Though the skeleton data are noisy on some individual datasets, the DS feature performs the best among all the methods on the combined dataset. This is because the skeleton data are

Table 3. Results of local HON4D and Dynamic Skeleton(DS) on combined dataset. Notation for the header: #A: number of actions; #S: number of total samples; Prec.: Precision; Rec.: Recall; Acc.: Accuracy. Notation for the protocols: RCSub: *Random Cross Subject*; RCSap: *Random Cross Sample*; RCSubB: *Random Cross Subject Balanced*; RCSapB: *Random Cross Sample Balanced*.

Test	Actions	Protocols (Applications)	#A,#S	local HON4D		DS	
				Prec.	Rec.	Prec.	Rec.
1	a001-a020, a067-a075, a082-a094	RCSub (HCI)	42,2821	69.3%	67.6%	83.1%	82.2%
2	a021-a066	RCSub (VS/HM)	46,1077	59.8%	59.4%	73.7%	73.3%
3	All 88	RCSap (VR)	88,3898	84.6%	84.1%	85.9%	85.6%
4	All 88	RCSub (VR)	88,3898	63.1%	59.3%	74.5%	73.7%

Test	Actions	Protocols (Applications)	#A,#S	local HON4D	DS
				Acc.	Acc.
5	Random 20	RCSubB (AS)	20,444	83.2%	88.6%
	Random 40		40,836	71.0%	83.9%
	Random 60		60,1308	66.7%	81.7%
	Random 80		80,1751	60.0%	74.2%
6	Random 20	RCSapB (AS)	20,504	95.6%	96.0%
	Random 40		40,966	89.0%	89.2%
	Random 60		60,1442	86.5%	89.6%
	Random 80		80,1974	81.5%	85.0%

Note that the samples without skeleton data are removed in local HON4D and DS methods, then the total number of actions is 88.

invariant to body sizes and backgrounds. However, the skeleton feature ignores the surrounding objects which may be discriminative among different actions. This may be the reason why DS method performs poorer on the VS/HM related actions with human-object interaction (test 2) compared to HCI scenarios (test 1). Similarly, results obtained on Test 5 and Test 6 in Table 3 show that the scalability of both local HON4D and DS is also poor, since the performance drops dramatically as more actions are involved. The results of both methods obtained on Test 5 are even lower than that on Test 6, which further shows the large variations among subjects in the combined dataset.

4.3 Comparison to Conventional Protocols

Experiments are also conducted to compare the proposed evaluation protocols with conventional protocols. We argue that the proposed protocols on the large dataset are different from the previously commonly used protocols on individual datasets. For example, one may argue that the *Random Cross Subject Balanced* protocol is the same as conventional *Cross Subject* protocol. However, we quickly point out that the combined dataset possesses the special property of including

multiple and different individual datasets which implies that the cross dataset features are involved in all the proposed protocols. We simulate the conventional *Cross Subject* protocol on the combined dataset by avoiding cross individual datasets when randomly selecting subjects for training and test, and then compare the results with that using the proposed *Random Cross Subject Balanced* protocol.

Table 4 shows the comparison results between the proposed *Random Cross Subject Balanced* and the conventional *Cross Subject* on the combined dataset. It can be seen that all the algorithms perform much poorer on the proposed *Random Cross Subject Balanced* protocol. This is probably due to the variations among datasets, since the cross dataset could be involved in the proposed *Random Cross Subject Balanced* protocol as mentioned and the algorithms may overfit to the samples of the datasets used for training the model.

Table 4. Comparisons between the proposed *Random Cross Subject Balanced* and the conventional *Cross Subject* on the combined dataset. Notation for the header: Acc.: Accuracy; RCSubB: *Random Cross Subject Balanced*; CSub: *Cross Subject*.

Actions	Applications	SNV		DSTIP	
		RCSubB (Acc.)	Conventional CSub (Acc.)	RCSubB (Acc.)	Conventional CSub (Acc.)
Random 20	AS	89.8%	96.1%	80.9%	89.0%
Random 40		75.9%	82.5%	71.4%	77.4%
Random 60		68.6%	79.4%	60.3%	69.7%
Random 80		58.4%	72.7%	55.2%	61.0%
Actions	Applications	Local HON4D		DS	
		RCSubB (Acc.)	Conventional CSub (Acc.)	RCSubB (Acc.)	Conventional CSub (Acc.)
Random 20	AS	83.2%	88.8%	88.6%	92.1%
Random 40		71.0%	79.6%	83.9%	84.3%
Random 60		66.7%	76.9%	81.7%	82.4%
Random 80		60.0%	73.5%	74.2%	80.5%

5 Conclusion and Future Work

A large scale RGB-D-based action dataset was introduced along with a set of evaluation protocols to overcome the limitations of both individual datasets and currently used evaluation protocols. Several experiments are conducted on the large datasets from different perspectives. Results show that current algorithms are not robust enough for real world applications. The experiments also verified the effectiveness of the combined dataset on evaluation of algorithms from different perspectives. The collection and design of this large dataset is an ongoing

effort and it will be enlarged to include more datasets, especially those with multiple views.

Table 5. The structure of the combined dataset. Dataset notation: D1: MSRAction3DExt [8, 14, 13], D2:UTKinect [15], D3:DailyActivity [12], D4:ActionPair [10], D5:CAD120 [7], D6:CAD60 [11], D7:G3D [2, 1], D8:RGBD-HuDa [9], D9:UTD-MHAD [4]

Action Label	Action Name	#Depth	#RGB	#Skeleton	#datasets	D1	D2	D3	D4	D5	D6	D7	D8	D9
a001	wave	131	131	131	3	✓						✓		✓
a002	horizontal arm wave	69	69	69	1	✓								
a003	hammer	69	69	69	1	✓								
a004	hand catch	100	100	100	2	✓								✓
a005	forward punch	128	128	128	2	✓						✓		
a006	throw	140	140	140	4	✓	✓	✓						✓
a007	draw x	99	99	99	2	✓								✓
a008	draw tick	69	69	69	1	✓								
a009	draw circle (clockwise)	101	101	101	1	✓								✓
a010	hand clap	151	151	151	4	✓	✓					✓		✓
a011	two hand high wave	89	89	89	2	✓	✓							
a012	side-boxing	69	69	69	1	✓								
a013	bend	70	70	70	1	✓								
a014	forward kick	129	129	129	2	✓						✓		
a015	side kick	59	59	59	1	✓								
a016	jogging	131	131	131	3	✓						✓		✓
a017	tennis swing forehand	131	131	131	3	✓						✓		✓
a018	tennis serve	131	131	131	3	✓						✓		✓
a019	golf swing	99	99	99	2	✓						✓		
a020	pickup & throw	103	103	103	2	✓								✓
a021	walk	101	101	101	4		✓	✓				✓		✓
a022	sit down	156	156	72	4		✓	✓					✓	✓
a023	stand up	154	154	72	4		✓	✓					✓	✓
a024	pick up from floor	62	62	62	3		✓		✓	✓				
a025	carry	19	19	19	1		✓							
a026	push	20	20	20	1		✓							
a027	pull	20	20	20	1		✓							
a028	drink	133	133	20	3			✓			✓		✓	
a029	eat	20	20	20	1			✓						
a030	read book	20	20	20	1			✓						
a031	talking on the phone	136	136	24	3			✓			✓		✓	
a032	write on a paper	20	20	20	1			✓						
a033	working on computer	24	24	24	2			✓			✓			
a034	use vacuum cleaner	20	20	20	1			✓						
a035	cheer up	20	20	20	1			✓						
a036	play game	20	20	20	1			✓						
a037	lay down on sofa	89	89	20	3			✓			✓		✓	
a038	play guitar	20	20	20	1			✓						
a039	pick up from table	30	30	30	1				✓					
a040	put down to table	30	30	30	1				✓					
a041	place a box	30	30	30	1				✓					
a042	push a chair	30	30	30	1				✓					
a043	pull a chair	30	30	30	1				✓					
a044	wear a hat	30	30	30	1				✓					
a045	take off a hat	30	30	30	1				✓					
a046	put on a backpack	30	30	30	1				✓					
a047	take off a backpack	30	30	30	1				✓					
a048	stick a poster	30	30	30	1				✓					
a049	remove a poster	30	30	30	1				✓					
a050	making cereal	16	16	16	1					✓				
a051	taking medicine	12	12	12	1					✓				
a052	stacking objects	12	12	12	1					✓				
a053	unstacking objects	12	12	12	1					✓				
a054	microwaving food	12	12	12	1					✓				
a055	cleaning objects	12	12	12	1					✓				
a056	taking food	12	12	12	1					✓				
a057	arranging objects	12	12	12	1					✓				
a058	having a meal	94	94	12	2					✓			✓	
a059	writing on whiteboard	4	4	4	1						✓			
a060	rinsing mouth	4	4	4	1						✓			
a061	brushing teeth	4	4	4	1						✓			
a062	wearing contact lens	8	8	8	1						✓			
a063	talking on couch	4	4	4	1						✓			
a064	cooking(chopping)	4	4	4	1						✓			
a065	cooking(stirring)	4	4	4	1						✓			
a066	opening pill container	12	12	12	1						✓			
a067	defend	30	30	30	1							✓		
a068	tennis swing backhand	30	30	30	1							✓		
a069	throw bowling ball	61	61	61	2							✓		✓
a070	aim and fire gun	90	90	90	1							✓		
a071	jump	30	30	30	1							✓		
a072	climb	30	30	30	1							✓		
a073	crouch	30	30	30	1							✓		
a074	steer a car	30	30	30	1							✓		
a075	flap	30	30	30	1							✓		
a076	mop the floor	84	84	0	1								✓	
a077	enter the room	83	83	0	1								✓	
a078	exit the room	82	82	0	1								✓	
a079	get up from bed	72	72	0	1								✓	
a080	take off the jacket	83	83	0	1								✓	
a081	put on the jacket	83	83	0	1								✓	
a082	swipe to the left	32	32	32	1									✓
a083	swipe to the right	32	32	32	1									✓
a084	cross arms in the chest	32	32	32	1									✓
a085	basketball shoot	32	32	32	1									✓
a086	draw circle(counter clockwise)	32	32	32	1									✓
a087	draw triangle	32	32	32	1									✓
a088	front boxing	32	32	32	1									✓
a089	baseball swing	32	32	32	1									✓
a090	arm curl	32	32	32	1									✓
a091	two hand push	32	32	32	1									✓
a092	knock on door	32	32	32	1									✓
a093	forward lunge	32	32	32	1									✓
a094	squat with two arms stretch out	31	31	31	1									✓

Bibliography

[1] Bloom, V., Argyriou, V., Makris, D.: Dynamic feature selection for online action recognition. In: Human Behavior Understanding, Lecture Notes in Computer Science, vol. 8212, pp. 64–76 (2013)

[2] Bloom, V., Makris, D., Argyriou, V.: G3D: A gaming action dataset and real time action recognition evaluation framework. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 7–12 (June 2012)

[3] Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27 (2011)

[4] Chen, C., Jafari, R., Kehtarnavaz, N.: UTD-MAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: Proc. IEEE International Conference on Image Processing (2015)

[5] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of machine learning research 9(Aug), 1871–1874 (2008)

[6] Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-D activity recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 5344–5352 (2015)

[7] Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. The International Journal of Robotics Research 32(8), 951–970 (2013)

[8] Li, W., Zhang, Z., Z.Liu: Action recognition based on a bag of 3D points. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 9–14 (June 2010)

[9] Ni, B., Wang, G., Moulin, P.: RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In: Proc. IEEE Conference on Computer Vision Workshops. pp. 1147–1153 (Nov 2011)

[10] Oreifej, O., Liu, Z.: HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 716–723 (2013)

[11] Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from RGBD images. In: Proc. AAAI workshop on Pattern, Activity and Intent Recognition (2011)

[12] Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1290–1297 (2012)

[13] Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., Ogunbona, P.O.: Action recognition from depth maps using deep convolutional neural networks. IEEE Transactions on Human-Machine Systems 46(4), 498–509 (Aug 2016)

[14] Wang, P., Li, W., Gao, Z., Tang, C., Zhang, J., Ogunbona, P.: Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1119–1122. ACM (2015)

[15] Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3D joints. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–27 (June 2012)

[16] Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2834–2841 (June 2013)

[17] Yang, X., Tian, Y.: Super normal vector for activity recognition using depth sequences. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 804–811 (June 2014)

[18] Zhang, J., Li, W., Ogunbona, P.O., Wang, P., Tang, C.: RGB-D-based action recognition datasets: A survey. Pattern Recognition 60, 86–105 (2016)