RGB-D-based Action Recognition Datasets: A Survey

Jing Zhang[a,*], Wanqing Li[a], Philip O. Ogunbona[a], Pichao Wang[a], Chang Tang[a,b]

[a]*School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia*
[b]*School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China*

Abstract

Human action recognition from RGB-D (Red, Green, Blue and Depth) 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. This raises the question of which dataset to select and how to use it in providing a fair and objective comparative evaluation against state-of-the-art methods. To address this issue, this paper provides a comprehensive review of the most commonly used action recognition related RGB-D video datasets, including 27 single-view datasets, 10 multi-view datasets, and 7 multi-person datasets. The detailed information and analysis of these datasets is a useful resource in guiding insightful selection of datasets for future research. In addition, the issues with current algorithm evaluation vis-á-vis limitations of the available datasets and evaluation protocols are also highlighted; resulting in a number of recommendations for collection of new datasets and use of evaluation protocols.

Keywords: Action recognition, RGB-D dataset, Evaluation protocol

1. Introduction

Human action recognition is an active research topic in Computer Vision. Prior to the release of Microsoft Kinect [TM], research has mainly focused on learning and recognizing actions from conventional two-dimensional (2D) video [1, 2, 3, 4]. There are many publicly available 2D video datasets dedicated to action recognition. Review papers categorizing and summarizing their characteristics are available to help researchers in evaluating their algorithms [5, 6, 7]. The introduction of low-cost integrated depth sensors (such as Microsoft Kinect [TM]) that can capture both RGB (red, green and blue) video and depth (D) information has significantly advanced the research of human action recognition. Since the first work reported in 2010 [8], many benchmark datasets have been created to facilitate the development and evaluation of new action recognition algorithms. However, available RGB-D-based datasets have insofar only been briefly summarized or enumerated without comprehensive coverage and in-depth analysis in the survey papers, such as [9, 10], that mainly focus on the development of RGB-D-based action recognition algorithms. The lack of comprehensive reviews on RGB-D datasets motivated the focus of this paper.

---

*Corresponding author.
Email addresses: `jz960@uowmail.edu.au` (Jing Zhang), `wanqing@uow.edu.au` (Wanqing Li), `philipo@uow.edu.au` (Philip O. Ogunbona), `pw212@uowmail.edu.au` (Pichao Wang), `tangchang@tju.edu.cn` (Chang Tang)

Datasets are important for the rapid development and objective evaluation and comparison of algorithms. To this end, they should be carefully created or selected to ensure effective evaluation of the validity and efficacy of any algorithm under investigation. The evaluation of each task-specific algorithm depends not only on the underlying methods but also on the factors captured by each dataset. However, it is currently difficult to select the most appropriate dataset from among the many Kinect sensor captured RGB-D datasets available and establish the most appropriate evaluation protocol. There is also the possibility of creating a new but redundant dataset because of the lack of comprehensive survey on what is available. This paper fills this gap by providing comprehensive summaries and analysis of existing RGB-D action datasets and the evaluation protocols that have been used in association with these datasets.

The paper focuses on action and activity datasets. "Gesture datasets" are excluded from this survey since, unlike actions and activities that usually involve motion of the entire human body, gesture involves only hand movement and gesture recognition is often considered as a research topic independent of action and activity recognition. For details of the available gesture datasets, readers are referred to the survey paper by Ruffieux et al. [7].

This rest of the survey is organized as follows. Section 2 summarises characteristics of publicly available and commonly used RGB-D datasets; the summaries (44 in total) are categorised under *single-view activity/action datasets*, *multi-view action/activity datasets* and *interaction/multi-person activity datasets*. Section 3 provides a comparative analysis of the reviewed datasets with regard to the applications, complexity, state-of-the-art results, and commonly employed evaluation protocols. In addition, some recommendations are provided to aid the future usage of datasets and evaluation protocols. Discussions on the limitations of current RGB-D action datasets and commonly used evaluation methods are presented in Section 4. At the same time, we provide some recommendations on requirements for future creation of datasets and selection of evaluation protocols. In Section 5, a brief conclusion is drawn.

2. RGB-D Action/Activity Datasets

This section summarizes most of the publicly available RGB-D action datasets, including the creation date, creation institution, number of actions, number of subjects involved, action repetition times, action classes, total number of video samples, capture settings, background and environment.

The datasets are categorized into three classes namely: *single-view action/activity*, *multi-view action/activity*, and *human-human interaction/multi-person activity*. In the single-view action/activity datasets, each action is captured from a single specific view point, while in the multi-view action/activity datasets, two or more view points of each action are captured. Note that in both single-view and multi-view datasets, each action/activity is performed by one actor at a time. The human-human interaction/multi-person activity datasets consist of interactions between two people or activities performed by multiple persons.

2

2.1. Single-view action/activity datasets

Table 1 is a list summarizing the basic specifications of single view action/activity datasets in descending order of citation frequency.

2.1.1. MSR-Action3D

MSR-Action3D [8](http://research.microsoft.com/en-_us/um/people/zliu/ActionRecoRsrc/) is the first public benchmark RGB-D action dataset collected by Microsoft Research Redmond and University of Wollongong in 2010. The dataset contains 20 actions: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis serve*, *golf swing*, *pickup and throw*. Ten subjects performed these actions three times. All the videos were recorded from a fixed point of view and the subjects were facing the camera while performing the actions. The background of the dataset was removed by some post-processing. Specifically, if an action needs to be performed with one arm or one leg, the actors were required to perform it using right arm or leg. The data are provided as segmented samples and the sample file names provide the information of action types, subject ID and number of repetitions.

2.1.2. RGBD-HuDaAct

RGBD-HuDaAct [11](http://adsc.illinois.edu/sites/default/files/files/ADSC-_RGBD-_dataset-_-download-_instructions.pdf) was collected by Advanced Digital Sciences Center Singapore in 2011. Compared to MSR-Action3D dataset, this dataset consists of fewer actions (12 actions) and performed by more subjects (30 subjects). The action types are also different from MSR-Action3D dataset. This dataset focuses on human daily activities, such as *make a phone call*, *mop the floor*, *enter the room*, *exit the room*, *go to bed*, *get up*, *eat meal*, *drink water*, *sit down*, *stand up*, t*ake off the jacket*, and *put on the jacket*. Each actor performed 2-4 repetitions of each action. The background is also fixed as the camera was fixed when recording. However, there was no restriction on which leg or hand was used in the actions and the dataset contains human-object interaction.

2.1.3. CAD-60

CAD-60 dataset [12](http://pr.cs.cornell.edu/humanactivities/data.php) was captured by Cornell University in 2011, motivated by the fact that true daily activities rarely occur in structured environments. Hence, the actions were performed within uncontrolled background. Twelve distinctive activities were performed within 5 environments: bathroom (*rinsing mouth*, *brushing teeth*, *wearing contact lens*), bedroom (*talking on the phone*, *drinking water*, *opening pill container*), kitchen (cooking (*chopping*), cooking (*stirring*), *drinking water*, *opening pill container*), living room (*talking on the phone*, *drinking water*, *talking on couch*, *relaxing on couch*), office (*talking on the phone*, *writing on whiteboard*, *drinking water*, *working on computer*). Four subjects performed all the activities and one of the subjects is left-handed. To determine whether test algorithms can distinguish the desired activities from other randomly performed

activities, additional random activity was collected, which contains a series of random movements that is different from any of other 12 activities in the dataset. In the original paper, this random activity was only used at testing stage.

2.1.4. MSRC-12

MSRC-12 dataset [13](http://research.microsoft.com/en-_us/um/cambridge/projects/msrc12/) was collected by Microsoft Research Cambridge and University of Cambridge in 2012. Although it is sometimes referred as gesture dataset, the movements involved whole body, so we categorize it as action/activity dataset. Two main goals motivated the collection of this dataset: first, to test whether semiotic modality of instructions for collecting data will affect the performance of the recognition system and, second, to determine whether the type of gesture makes a difference in the effect of modality. So, there are two types of gestures: Iconic gestures (*Crouch* or *hide*, *Shoot a pistol*, *Throw an object*, *Change weapon*, *Kick*, and *Put on night vision goggles*) and Metaphoric gestures (*Start Music/Raise Volume (of music)*, *Navigate to next menu*, *Wind up the music*, *Take a bow to end music session*, *Protest the music*, and *Move up the tempo of the song*). The authors provided three familiar and easy to prepare instruction modalities and their combinations to the participants. The modalities are (1) descriptive text breaking down the performance kinematics, (2) an ordered series of static images of a person performing the gesture with arrows annotating as appropriate, and (3) video (dynamic images) of a person performing the gesture. There are 30 participants in total and for each gesture, the data were collected as: Text (10 people), Images (10 people), Video (10 people), Video with text (10 people), Images with text (10 people). The dataset was captured using Kinect ^{TM}sensor and only the skeleton data are made available.

2.1.5. MSRDailyActivity3D

MSRDailyActivity3D Dataset [14](http://research.microsoft.com/en-_us/um/people/zliu/ActionRecoRsrc/) was collected by Microsoft and the Northwestern University in 2012 and focused on daily activities. The motivation was to cover human daily activities in the living room. There are 16 activity types: *drink*, *eat*, *read book*, *call cellphone*, *write on a paper*, *use laptop*, *use vacuum cleaner*, *cheer up*, *sit still*, *toss paper*, *play game*, *lay down on sofa*, *walk, play guitar*, *stand up, sit down*. The actions were performed by 10 actors while sitting on the sofa or standing close to the sofa. The camera was fixed in front of the sofa. In addition to depth data, skeleton data are also recorded, but the joint positions extracted by the tracker are very noisy due to the actors being either sitting on or standing close to the sofa.

2.1.6. UTKinect

UTKinect dataset [15](http://cvrc.ece.utexas.edu/KinectDatasets/HOJ3D.html) was collected by the University of Texas at Austin in 2012. Ten types of human actions were performed twice by 10 subjects. The actions include *walk, sit down, stand up, pick up, carry, throw, push, pull, wave, clap hands*. The subjects performed the actions from a variety of views. An added difficulty of recognition was afforded by the actions

being performed with actor-dependent variability. Furthermore, human-object occlusions and body parts being out of the field of view added to the difficulty of the dataset in recognition tasks. Ground truth in terms of action labels and segmentation of sequences are provided.

2.1.7. G3D

Gaming 3D dataset (G3D) [16, 17](http://dipersec.king.ac.uk/G3D/) captured by Kingston University in 2012 focuses on real-time action recognition in gaming scenario. It contains 10 subjects performing 20 gaming actions: *punch right*, *punch left*, *kick right*, *kick left*, *defend*, *golf swing*, *tennis serve*, *throw bowling ball*, *aim and fire gun*, *walk*, *run*, *jump*, *climb*, *crouch*, *steer a car*, *wave*, *flap*, and *clap*. Each subject performed these actions thrice. Two kinds of labels were provided as ground truth: the onset and offset of each action and, the peak frame of each action. In [17], the authors defined an *action point* as a single time instance that an action is clear and all instances of that action can be uniquely identified. The peak frame provided in this dataset represents the action point indicated by the authors. This action point can be used for evaluating on-line action recognition algorithms.

2.1.8. DHA

Depth-included Human Action video dataset (DHA) [18](http://mclab.citi.sinica.edu.tw/dataset/dha/dha.html) was created by CITI in Academia Sinica. It contains 23 different actions: *bend*, *jack*, *jump*, *run*, *side*, *skip*, *walk*, *one-hand-wave*, *two-hand-wave*, *front-clap*, *side-clap*, *arm-swing*, *arm-curl*, *leg-kick*, *leg-curl*, *rod-swing*, *golf-swing*, *front-box*, *side-box*, *tai-chi*, *pitch*, *kick*. The first 10 categories follow the same definitions as the Weizmann action dataset [19] and the 11th to 16th actions are extended categories. The 17th to 23rd are the categories of selected sport actions. The 23 actions were performed by 21 different individuals. All the actions were performed in one of three different scenes. Similarly to MSRAction3D dataset, the background information has been removed in the depth data.

2.1.9. Falling Event Detection

The Falling Event Detection dataset [20](http://media-_lab.engr.ccny.cuny.edu/~zcy/) was collected in 2012 by City University of New York with the aim of creating a dataset for evaluating a newly proposed method for falling event detection and recognition. There are five activities related to falling event including *standing, fall from standing, fall from sitting, sit on a chair*, and *sit on floor*, captured using a RGB-D camera. The activities were performed by five different subjects under two different lighting environments (sufficient and insufficient illumination) resulting in 150 video sequences (100 videos under sufficient and 50 videos under insufficient illumination). The authors set aside a training set comprising 50 videos which covers all 5 subjects and 5 types of activities performed under sufficient lighting. The remaining 100 video sequences (50 for each condition) were set aside for testing.

2.1.10. MSRActionPair

MSRActionPair dataset [21](http://www.cs.ucf.edu/~oreifej/HON4D.html) was collected by University of Central Florida and Microsoft in 2013, and has two foci. First, the authors argue that many actions share similar motion cues; hence, relying only on motion information is insufficient for recognition. Second, considering motion and shape information independently is inefficient because they are correlated in an action sequence. As a result, they collected a dataset with pairs of actions; for example, *pick up* and *put down*. The action pairs share similar motion and shape cues but the relation between motion and shape is different. The background of the dataset was fixed, without occlusion and change of lighting. To perform well on this dataset, the algorithm needs to be able to capture the prominent cues of motion and shape jointly. In this dataset, ten subjects performed six pairs of actions twice: *pick up a box/put down a box, lift a box/place a box, push a chair/pull a chair, wear a hat/take off a hat, put on a backpack/take off a backpack*, and *stick a poster/remove a poster*.

2.1.11. CAD-120

CAD-120 dataset [22](http://pr.cs.cornell.edu/humanactivities/data.php), collected by the Cornell University, focuses on high level activities and object interactions. This dataset contains 10 high level activities performed by 4 subjects, and each activity was performed thrice with different objects. The high level activities include: *making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, having a meal*. The high level activities consist of a sequence of sub-activities. Different subjects performed the sub-activities over different length of time and, in different order and manner of execution. In addition, the subjects may perform the same activity with different objects. The backgrounds are also varied among actions. Based on above features, CAD-120 dataset not only can be used for action recognition, but also can be used to evaluate some object detection and tracking algorithms. The dataset also provides some ground-truth, such as the bounding boxes of the objects involved in the activities, sub-activity labels and object affordance labels.

2.1.12. WorkoutSU-10 dataset

WorkoutSU-10 dataset [23](http://vpa.sabanciuniv.edu/databases/WorkoutSU-_10/) was collected by Sabanc University in 2013 and contains exercise actions selected by professional trainers for therapeutic purposes. There are 10 actions in total, namely *SL Balance with Hip Flexion, SL Balance-Trunk Rotation (A2), Lateral Stepping, Thoracic Rotation Bar on shoulder, Hip Adductor Stretch, Hip Adductor Stretch, DB Curl-to-Press, Freestanding Squats, Transverse Horizontal DB Punch, Transverse Horizontal DB Punch*. The performance instruction was the combination of an animated character performing the exercise and a subscripted text explaining the instructions. The RGB, depth, and skeleton data were all captured. Twelve subjects performed all the actions 10 times. There are 1200 action samples in total. The participants performed the action in front of a green screen, suggesting that the background of this dataset is clean.

2.1.13. Concurrent Action

The concurrent action dataset [24](http://www.stat.ucla.edu/~ping.wei/research/project/ConcurrentAction/ConcurrentAction.html) was collected by Xi'an Jiaotong University and University of California, Los Angeles in 2013. This dataset focuses on action detection. Twelve actions were performed by several subjects in a sequential fashion. The actions are: *drink, make a call, turn on monitor, type on keyboard, fetch water, pour water, press button, pick up trash, throw trash, bend down, sit,* and *stand.* Sixty-one long video sequences were captured. Each sequence contains several actions which are concurrent in the time and interact with others. The dataset is different from previously created dataset in that it contains multiple concurrent actions in each sequence and the actions semantically and temporally interact with each other. Only skeleton data format are available for this dataset.

2.1.14. IAS-lab Action

IAS-lab Action dataset [25, 26](http://robotics.dei.unipd.it/actions/index.php/overview) was collected by IAS Lab at the University of Padua in 2013. The authors claimed that in order to test as many different algorithms as possible, a dataset needs to contain sufficient variety of actions and number of people performing the actions. To this end, they captured 15 different actions performed by 12 different people thrice. The actions are: *check watch*, *cross arms*, *get up*, *kick*, *pick up*, *point*, *punch*, *scratch head*, *sit down*, *standing*, *throw from bottom up*, *throw over head*, *turn around*, *walk*, and *wave*. The subjects were asked to perform well defined actions rather than in free style, because the authors argued that variability could bias the evaluation of the performance of an algorithm. Notice that all actions were captured in the same indoor setting and with clean background.

2.1.15. UCFKinect

In order to explore the trade-off between accuracy and observational latency when recognizing actions, UCFKinect dataset [27](http://www.cs.ucf.edu/~smasood/datasets/UCFKinect.zip) was created. It was collected by University of Central Florida Orlando in 2013. This dataset can be used for measuring how fast a recognition system can overcome the ambiguity in initial poses when performing an action. The dataset is composed of 16 actions, including *balance, climb up, clumb ladder, duck, hop, vault, leap, run, kick, punch, twist left, twist right, step forward, step back, step left, step right.* Sixteen subjects (13 males and 3 females, all ranging between ages 20 to 35) were involved with each subject performing all 16 actions 5 times for a total of 1280 action samples. The dataset is only presented as skeleton data comprising 3-dimensional coordinates of 15 joints along with the correponding orientation and binary confidence values. Subjects were asked to stand in a relaxed posture with loosely downward hanging arms beside the body before performing different actions. They were then told what action to perform and if requested, given a demonstration of the action. The end of a countdown signalled the beginning of recording and performance of the action. The recording was manually stopped upon completion of the action. The authors claimed that gathering

the data in this fashion simulates a gaming scenario where the user performs a variety of actions, such as punches and kicks, and returns to a resting pose between actions.

2.1.16. Osaka University Kinect Action

The Osaka University Kinect Action Dataset [28](`http://www.am.sanken.osaka-_u.ac.jp/~mansur/dataset.html`) was collected by Osaka University in 2013 within laboratory environment. Ten actions were performed by 8 subjects and once. Action types are *jumping jack type 1, jumping jack type 2, jumping on both legs, jumping on right leg, jumping on left leg, running, walking, side jumps, skipping on left leg,* and *skipping on right leg.* RGB, depth, and skeleton data were all captured. The background and illumination conditions remained unchanged during the capture sessions.

2.1.17. Human Morning Routine Dataset

Human Morning Routine dataset [29](`http://www.uni-_tuebingen.de/fakultaeten/mathematisch-_-naturwissenschaftliche-_fakultaet/fachbereiche/informatik/lehrstuehle/human-_computer-_interaction/home/code-_datasets/morning-_routine-_dataset.html`) was collected by Technische Universität München and the Eberhard Karls Universität Tübingen in 2013. It is aimed at testing algorithms for recognizing and monitoring morning routine of a human in a kitchen. A robot was supposed to be able to react to these activities/actions. They include *preparing a drink*, *drinking a glass of water*, *preparing breakfast*, *having breakfast*, *cleaning the table*, *packing a bottle of water into the backpack*, and *leaving the room with the backpack*. A participant reenacted and logged his morning routine (including location he stood while performing those activities) in an experimental kitchen equipped with two Kinect ᵀᴹdevices (one for motion-tracking and the other for detection of objects). The actions were annotated to provide ground truth.

2.1.18. RGBD-SAR Dataset

RGBD-SAR Dataset [30](`http://www.uestcrobot.net/en/?q=download`), created by the University of Electronic Science and Technology of China and Microsoft, aimed at algorithms monitoring behaviours of seniors. Nine categories of elderly daily activities are collected: *put on the jacket*, *take off the jacket*, *enter the room*, *exit the room*, *sit down*, *stand up*, *drink water*, *eat meal*, and *walk*. Thirty elderly people were invited to perform these activities and each of them performed each activity thrice.

2.1.19. Mivia Dataset

Mivia dataset [31](`http://mivia.unisa.it/datasets/video-_analysis-_datasets/mivia-_action-_dataset/`) was acquired by Mivia Lab at the University of Salemo in 2013. It consists of 7 high-level actions performed by 14 subjects. Each subject performed 5 repetitions of each action. The actions include: *opening a jar*, *drinking*, *sleeping*, *random movements*, *stopping*, *interacting with a table* and *sitting*.

2.1.20. UPCV

The UPCV action dataset [32](http://www.upcv.upatras.gr/personal/kastaniotis/datasets.html) was collected by the University of Patras in 2014. The dataset consists of 10 actions performed by 20 subjects twice. The actions, representing activities usually performed by ppedestrians, include: *walk*, *seat*, *grab*, *phone*, *watch clock*, *scratch head*, *cross arms*, *punch*, *kick*, and *wave*. The published UPCV dataset only contains skeleton data. The subjects perform the actions in front of a fixed camera in a natural manner and against a stationary background. The ground truth provided is the annotation of data, which can isolate the action data from the overall motion.

2.1.21. TJU dataset

The TJU dataset [33](http://media.tju.edu.cn/tju_dataset.html) was captured by Tianjin University in 2014. and contains 22 actions performed by 20 subjects in two different environments; a total of 1760 sequences. Action types include: *boxing, side boxing, one hand wave, two hands wave, hand clap, side bend, forward bend, draw X, draw tick, draw circle, tennis serve, tennis swing, walking, side walking, jogging, running, jacks, jump, jump in place, forward kick, side kick,* and *sit down.* The background was fixed during capture and was subtracted from depth data before publishing the dataset.

2.1.22. MAD

Due to the fact that there were very few publicly available sequential action dataset which can be used in the development and evaluation of detection algorithms, the Multi-modal action detection (MAD) Dataset [34](http://humansensing.cs.cmu.edu/mad/download.html) was created by Carnegie Mellon University in 2014. It contains 35 sequential actions performed by 20 subjects. Each subject performed the sequential actions twice. There are 40 sequences in total (2 sequences for each subject). The actions include: *Running, crouching, jumping, walking, jump and side-kick, left arm swipe to the left, left arm swipe to the right, left arm wave, left arm punch, left arm dribble, left arm pointing to the ceiling, left arm throw, swing from left (baseball swing), left arm receive, left arm back receive, left leg kick to the front, left leg kick to the left, right arm swipe to the left, right arm swipe to the right, right arm wave, right arm punch, right arm dribble, right arm pointing to the ceiling, right arm throw, swing from right (baseball swing), right arm receive, right arm back receive, right leg kick to the front, right leg kick to the right, cross arms in the chest, basketball shooting, both arms pointing to the screen, both arms pointing to both sides, both arms pointing to right side, both arms pointing to left side.* The authors provided ground truth labels which indicated the start and end of the actions and are suitable for both detection and classification.

2.1.23. Composable activities

Composable activities dataset [35](http://web.ing.puc.cl/~ialillo/ActionsCVPR2014/) was created by Pontificia Universidad Catolica de Chile and Universidad del Norte in 2014. It was aimed at the problem of recognizing complex activities, such as *waving while walking, talking on the phone while running away to*

attend an urgent matter, etc. Different combinations of 26 atomic actions formed 16 activity classes which were performed by 14 subjects and annotations were provided. Each activity is composed of 3 to 11 atomic actions. For example, the activity *walk while hand waving* consists of 3 atomic actions: *walk*, *hand wave*, and *idle*; while the activity *composed-activity-4* is composed of 11 atomic actions: *idle*, *walk*, *call a friend with hands*, *hand wave*, *talking on cellphone*, *pick from the floor*, *dial cellphone*, *put an object*, *pick cellphone from pocket*, and *put cellphone in pocket*.

2.1.24. 3D Online Action

3D online action dataset [36](`https://sites.google.com/site/skicyyu/rgbd_recognition`) was collected by Microsoft and Nanyang Technological University in 2014 with the aim of developing and testing algorithms for continuous online human action recognition from RGB-D data. There are seven action categories: *drinking*, *eating*, *using laptop*, *reading cellphone*, *making phone call*, *reading book* and *using remote*. Thirty-six subjects performed the actions in this dataset. The dataset is intended for the evaluation of three categories of tasks: same-environment action recognition, cross-environment action recognition, and continuous action recognition. In order to achieve this purpose, the dataset was separated into four sections: first two sections contain single action in each sample and were captured in same environment; the third section also contains single action in each sample, but was captured in a different environment; the fourth section contains multiple, albeit orderless actions in each sample. The bounding box of the object involved in each frame is manually labelled.

2.1.25. RGB-D activity dataset

The RGB-D activity dataset [37](`http://watchnpatch.cs.cornell.edu/`) was collected by Cornell University and Stanford University in 2015. The dataset was recorded by the Kinect v2 camera. Each video in the dataset contains 2-7 actions involving interaction with different objects. Compared to previous Kinect v1 system, the Kinect v2 has higher resolution of RGB-D data (RGB: 1920*1080, depth: 512*424) and improved body tracking of human skeletons (25 body joints). In this dataset, 21 actions (10 in the office, 11 in the kitchen) interacted with 23 types of objects were performed by 7 subjects. The action categories are: *turn-on-monitor, turn-off-monitor, walking, play-computer, reading, fetch-book, put-back-book, take-item, put-down-item, leave-office, fetch-from-fridge, put-back-to-fridge, prepare-food, microwaving, fetch-from-oven, pouring, drinking, leave-kitchen, move-kettle, fill-kettle, and plug-in-kettle*. The background of the captured scene are relatively complex and in each environment the activities were performed relative to different views. In total, there are 458 videos with a total length of about 230 minutes.

2.1.26. SYSU 3D Human-Object Interaction Dataset

The SYSU 3D Human-Object Interaction dataset [38](`http://sist.sysu.edu.cn/~zhwshi/students/jianfang/HomePage.htm`) was created by Sun Yat-sen University in 2015. This dataset focuses on actions involving human-object interaction. Forty subjects perform 12 distinct activities, such as *drinking, pouring,*

calling phone, playing phone, wearing backpacks, packing backpacks, sitting chair, moving chair, taking out wallet, taking from wallet, mopping, and *sweeping.* For each activity, each subject manipulates one of the six different objects: phone, chair, bag, wallet, mop and besom. Hence, the dataset contains 480 video clips in total. The RGB frames, depth sequence and skeleton data of each video clips are captured by a Kinect camera. The authors claimed that their dataset presents some new challenges compared to previous datasets. For example, the motions and the appearance of manipulated objects are highly similar between some activities, and the number of participants is larger than that of any existing dataset.

2.1.27. UTD-MHAD

UTD-MHAD [39](`http://www.utdallas.edu/~cxc123730/UTD-_MHAD.html`) was collected by University of Texas at Dallas in 2015. Eight subjects performed 27 actions four times. The 27 actions are: *right arm swipe to the left, right arm swipe to the right, right hand wave, two hand front clap, right arm throw, cross arms in the chest, basketball shoot, right hand draw x, right hand draw circle (clockwise), right hand draw circle (counter clockwise), draw triangle, bowling (right hand), front boxing, baseball swing from right, tennis right hand forehand swing, arm curl (two arms), tennis serve, two hand push, right hand knock on door, right hand catch an object, right hand pick up and throw, jogging in place, walking in place, sit to stand, stand to sit, forward lunge (left foot forward),* and *squat (two arms stretch out).* All the actions were performed in a fixed background. An inertial sensor was worn on the subject's right wrist for action 1 to 21, and on the right thigh for action 22 to 27. Hence, four types of data modalities were captured, namely RGB videos, depth videos, skeleton joint positions, and the inertial sensor signals.

Dataset	Year(Cited[1])	Modality	#a,#s,#e	Protocol
MSR-Action3D [8]	2010 (333)	D,S	20,10,567	1. 1/3 training 2. 2/3 training 3. Half training, half testing CS
MSRDaily-Activity3D [14]	2012 (311)	C,D,S	16,10,320	Half training, half test CS
UTKinect [15]	2012,(193)	C,D,S	10,10,200	LOSeqO
CAD-60 [12]	2011(159)	C,D,S	12,4,60	1. LOSubO 2. Halved the testing subject's data and included one half in the training dataset
RGBD-HuDaAct [11]	2011(148)	C,D	12,30,1189	LOSubO
MSRAction-Pair [21]	2013(136)	C,D,S	12,10,180	First half training CS
MSRC-12 gesture [13]	2012(100)	S	12,30,594	LOSubO
CAD-120 [22]	2013(81)	C,D,S	10,4,120	LOSubO (4-fold CV)
UCFKinect [27]	2013(62)	S	16,16,1280	4-fold CV
G3D [16, 17]	2012(28)	C,D,S	20,10,234	CS (4 subjects training, 1 subject validation, 5 subjects test)
Falling Event [20]	2012(21)	C,D,S	5,5,150	50 samples covering 5 subjects and 5 activities with sufficient lighting for training, rest for testing
UPCV [32]	2014(18)	S	10,20,400	LOSubO
DHA [18]	2012(17)	C,HM,D	23,21,483	CS (10 training,11 test)

---

[1]Citations as of 31 August 2015

WorkoutSU-10 [23]	2013(16)	C,D,S	10,12,1200	CS
IAS-lab [25, 26]	2013(15)	C,D,S,P	15,12,540	LOSubO
Osaka [28]	2013(8)	C,D,S	10,8,80	LOSubO CV
Mivia [31]	2013(6)	C,D	7,14,490	1. Leave two repetitions of one person out. 2. LOSubO
Concurrent Action [24]	2013(5)	S	12,-,61	Not given
TJU [33]	2014(4)	C,D,S	22,20,1760	First 12 subjects training, rest test
3D Online [36]	2014(4)	C,D,S	7,36,386	1. Same-Environment (2-fold CV) 2. Cross-Environment (S1, S2 training, S3 test) 3. Continuous (S1, S2, S3 training, S4 test)
MAD [34]	2014(3)	C,D,S	20,35,40	5-fold CV (8 groups training, 2 groups test)
Composable [35]	2014(3)	C,D,S	16,14,693	LOSubO
RGBD-SAR [30]	2013(1)	C,D	12,6,810	Not given
SYSU [38]	2015(0)	C,D,S	12,40,480	1. Half samples training, rest test 2. CS
RGB-D activity [37]	2015(0)	C,D,S	21,7,458	Not given
UTD-MHAD [39]	2015(0)	C,D,S,I	27,8,861	CS (odd subjects training, even subjects test)
Morning-Routine [29]	2013(0)	C,D,S	7,1,14	Not given

Table 1: Summary of basic specifications of Single-view action/activity datasets. Notation for the header: #a: number of actions, #s: number of subjects, #e: number of total examples. Notation for data format: C: Colour, D: Depth, S: Skeleton, HM: Human Mask, P: Point clouds, I: Inertial sensor data. Notation for protocol: CS: Cross Subject, LOSeqO: Leave One Sequence Out, LOSubO: Leave One Subject Out, CV: Cross Validation

2.2. Multi-view action/activity datasets

A multi-view dataset can be generated in at least two ways. First, several cameras can be mounted at different positions and angles. Second, the same action can be repeated from different viewpoints. The reviewed multiview datasets are generated using these two approaches. However, most of them are captured by multiple cameras. Similarly to the review of single-view datasets, the descriptions of multiview datasets are given in chronological order. Table 2 shows a summary of basic specifications of multi-view datasets.

2.2.1. ATC4^2

ATC4^2 dataset [40](http://vipl.ict.ac.cn/rgbd-_action-_dataset)was collected by Institute of Computing Technology of Chinese Academy of Science in 2012 for the purpose of providing an evaluative framework that supports view variations of actions. The dataset focuses on facilitating practical applications, such as smart house or e-healthcare, and contains 14 daily activities: *Collapse, Drink, MakePhonecall, MopFloor, PickUp, PutOn, ReadBook, SitDown, SitUp, Stumble, TakeOff, ThrowAway, TwistOpen, WipeClean*. Note that *Collapse* and *Stumble* are two activities specific to homecare applications. The authors distinguished between *Collapse* (people falling as a result of inner factors, such as hurt or giddiness) and *Stumble* (body dropping caused by outside effects such as tripping on an obstacle). The dataset was captured by 4 Kinect

sensors from different heights and view angles. Twenty-four subjects performed the 14 activities for several times. The labels of start/stop points of single actions are provided.

2.2.2. Falling Detection

The Falling Detection dataset [41](http://vlm1.uta.edu/~zhangzhong/fall_detection/) was collected by the University of Texas in 2012. It focused on falling actions captured in a laboratory-based simulated apartment set up. Six subjects in two sceneries performed a series of actions continuously, including both *real fall actions* and fall-like actions, such as *picking up a coin from floor, sitting down on the floor, tying shoelaces, sleeping down on the bed, opening the lower drawer which is close to the floor, jumping on to the floor, and sleeping down on the floor.* Only depth data sequences are published along with annotation of the start and end frame for every fall process, but not other actions. There are 12 real falls in video from the first scene, and 14 real falls in the second scene. For the fall like actions, there are 23 examples of picking up something from the floor, 12 cases of sitting on the floor, 10 examples of tying shoelaces, 9 examples of lying down on the bed, 5 examples of opening/closing a drawer at floor level, 1 example of jumping on the bed, and 1 example of lying on the floor.

2.2.3. Berkeley MHAD

Berkeley Multimodal Human Action Database (MHAD) [42](http://tele-_immersion.citris-_uc.org/berkeley_mhad#dl), collected by University of California at Berkeley and Johns Hopkins University in 2013, was captured in five different modalities to expand the fields of application. The modalities are derived from: optical mocap system, four multi-view stereo vision cameras, two Microsoft Kinect ^{TM}cameras, six wireless accelerometers and four microphones. Twelve subjects performed 11 actions, five times each. Three categories of actions are included: (1) actions with movement in full body parts, e.g., *jumping in place, jumping jacks, throwing*, etc., (2) actions with high dynamics in upper extremities, e.g., *waving hands, clapping hands*, etc. and (3) actions with high dynamics in lower extremities, e.g., *sit down, stand up*. The actions were executed with style and speed variations. This dataset can be used in the development and evaluation of multimodal algorithms, such as action recognition, pose estimation, motion segmentation and dynamic 3D scene reconstruction.

2.2.4. DMLSmartActions

DMLSmartActions dataset [43](http://dml.ece.ubc.ca/data/smartaction/) was collected by the University of British Columbia in 2013 and aimed at demonstrating the real situation in a home environment. Two high-definition (HD) RGB cameras and one Kinect sensor were utilized for collecting the data. Although the three cameras were static during acquisition, their location and orientation were not fixed so as to provide variability. The Kinect ^{TM}sensor was always located between the two HD RGB cameras in different scenes. Sixteen subjects performed 12 different actions in a natural manner. The actions include: *clean-table, drink, drop-and-pickup, fell-down, pick-something, put-something, read, sit-down, standup, use-cellphone, walk*, and

write. Subjects were asked to perform a series of the listed actions in a natural style, suggesting that there was no instruction on how or when to perform these actions. The data was manually labelled into samples.

2.2.5. ReadingAct

ReadingAct dataset [44] was collected by Reading University in 2013, using 2 Kinect sensors; one was in front of the subject and the other was placed orthogonally to capture a side view. Twenty actors performed the actions four times in free form style to ensure variability. The dataset includes a background scene and 19 actions: *coming in, going out, walking past, walking around, switching light, talking on phone, phone call (mobile), picking up from floor, putting on jacket, hoovering floor, sitting down, standing up, lying down, getting up, reading a book, typing on computer, having meal, drinking (sitting)* and *drinking (standing)*.

2.2.6. Multiview 3D Event

Multiview 3D Event dataset [45](`http://www.stat.ucla.edu/~ping.wei/research/project/4DHOI/4DHOI.html`) was created by University of California at Los Angles in 2013 using three simultaneous Kinect ^{TM}sensors from different viewpoints around the subjects. This dataset includes 8 categories of events performed by 8 subjects 20 times independently with different object instances and in various styles. The eight event categories are: *drink with mug, call with cellphone, read book, use mouse, type on keyboard, fetch water from dispenser, pour water from kettle,* and *press button.* These events involve 11 object classes: mug, cellphone, book, mouse, keyboard, dispenser, kettle, button, monitor, chair, and desk. To label the data, the videos were manually cut into sequences wherein each sequence contains one action.

2.2.7. Northwestern-UCLA Multiview Action 3D

Northwestern-UCLA Multiview Action 3D [46](`http://users.eecs.northwestern.edu/~jwa368/my_data.html`) was collected by Northwestern University and University of California at Los Angles in 2014. The capture settings were similar to Multiview 3D Event dataset but adds multiple locations. The actions were performed by 10 actors and captured by three simultaneous Kinect cameras. There are 10 action categories: *pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry.*

2.2.8. UWA3D Multiview

UWA3D Multiview Activity Dataset [47](`http://staffhome.ecm.uwa.edu.au/~00053650/databases.html`) was collected by the University of Western Australia in 2014. In this dataset, all actions were captured continuously without break or pause. Thirty activities were performed by 10 individuals: *one hand waving, one hand Punching, sitting down, standing up, holding chest, holding head, holding back, walking, turning around, drinking, bending, running, kicking, jumping, moping floor, sneezing, sitting down(chair), squatting, two hand waving, two hand punching, vibrating, falling down, irregular walking, lying down, phone answering, jumping jack, picking up, putting down, dancing,* and *coughing.* Each subject performed the 30 activities

twice or thrice continuously in random order. To achieve multiview, five subjects performed 15 activities from four different side views.

2.2.9. Muti-View TJU dataset

The Muti-View TJU dataset [48](`http://media.tju.edu.cn/tju_dataset.html`) was captured by Tianjin University in 2014 and represents similar action types as in TJU dataset. However, this dataset was captured with two Kinect cameras from two viewpoints (front view and side view) and the angle between the two views is around 65 degrees. The 22 actions were performed by 20 subjects four times in both light and dark environments. There are 7040 samples in total. Each action was recorded in modes RGB, depth, skeleton data, and human mask.

2.2.10. NJUST RGB-D Action

NJUST RGB-D Action dataset [49](`http://imag.njust.edu.cn/imag/NJUST_RGB-_D_Action_Dataset.html`) was collected by Nanjing University of Science and Technology in 2014. The dataset was collected in lab environments with subjects located at about three meters from the camera. There are 19 action categories: *Bending, Bending-side, Boxing, Checking-Time, Drinking, DroppingBag, Kicking, LyingDown, OpeningCloset, PickingUp, PullingOut, SittingDown, Squatting, StandingUp, TakingPhoto, Telephoning, Tossing, Walking*, and *Waving*. Each action was performed by ten subjects in two scenes. This dataset also provides some view variation samples of six actions. To achieve view variation, the subjects were asked to perform the six actions with 30 degree view angle to the camera. The six actions are: *Bending-30D, Boxing-30D, Drinkin-30D, SittingDown-30D, Squatting-30D and StandingUp-30D*. Altogether, there are 500 action samples. For each sample, RGB frames, depth frames, skeleton data, and body segmentation are provided.

Dataset	Year(Cited[2])	Modality	#a,#s,#e	Protocol
Berkeley MHAD [42]	2013(50)	C, D, M, A, Au	12,12,720	CS (First 7 training, last 5 test)
ATC4[2] [40]	2012(27)	C,D	14,24,6844	8 training, 16 test CS
Falling Detection [41]	2012(17)	D	8,6,12	CS
Multiview 3D Event [45]	2013(13)	C,D,S	8,8,3815	Not given
Multi-View TJU [48]	2013(6)	C,D,S	20,22,7040	6 subjects training, 6 validation, 8 test
Northwestern-UCLA [46]	2014(5)	C,D,S	10,10,-	1. LOSubO 2. 2 Camera training,1 Camera test 3. test on different environment
UWA3D Multiview [47]	2014(4)	D,S	30,10,600+	1. CS (Half training, half test) 2. 0° training
NJUST [43]	2014(2)	C,D,S,HM	19,10,500	LOSubO CV
DMLSmart Actions [43]	2013(2)	HDC,C,D	12,16,932	LOSubO
ReadingAct [44]	2013(1)	C,D	19,20,2340	CS (15 training, 5 test, 4-fold CV)

[2]Citations are as of 31 August 2015

Table 2: Summary of basic specifications of Multi-view action/activity datasets. Notation for the header: #a: number of actions, #s: number of subjects, #e: number of total examples. Notation for data format: C: Colour, D: Depth, S: Skeleton, M: Mocap, SV: Stereo Video, Au: Acceleration, A: Audio, HM: Human Masks, HDC: High Definition Colour. Notation for protocol: CS: Cross Subject, LOSubO: Leave One Subject Out, CV: Cross Validation

2.3. Interaction/Multi-person activity datasets

The human-human interaction datasets normally contain interaction between two persons. The number of persons involved in multi-person activity is not fixed. A summary of basic specifications of interaction/multi-person activity datasets is provided in Table 3.

2.3.1. SBU Kinect Interaction Dataset

SBU [50](`http://www3.cs.stonybrook.edu/~kyun/research/kinect_interaction/index.html`) was collected by Stony Brook University in 2012. It contains eight types of interactions, including: *approaching*, *departing*, *pushing*, *kicking*, *punching*, *exchanging objects*, *hugging*, and *shaking hands*. All videos were recorded with the same indoor background. Seven participants were involved in performing the activities which have interactions between two actors. The dataset is segmented into 21 sets and each set contains one or two sequences of each action category. Two kinds of ground truth information are provided: action labels of each segmented video and identification of "active" actor and "inactive" actor.

2.3.2. K3HI

Similarly to SBU dataset, K3HI [51](`http://www.lmars.whu.edu.cn:8086/prof_web/zhuxinyan/DataSetPublish/dataset.html`) is also a two-person interaction dataset. It was collected by Wuhan University in 2013. Fifteen volunteers performed 8 categories of activities, including *approaching*, *departing*, *kicking*, *punching*, *pointing*, *pushing*, *exchanging an object*, and *shaking hands*. In order to ensure the integrity and continuity of the spatial information of the skeleton data of the two persons, the RGB and depth data were ignored during data capture.

2.3.3. The LIRIS human activities dataset

LIRIS Human Activities Dataset [52](`http://liris.cnrs.fr/voir/activities-_dataset/`), collected by the French National Center for Scientific Research in 2014, was captured in complex scenarios. The Kinect TMsensor was mounted on a remotely controlled robot to capture activities involving human-human interactions, human-object interactions and human-human-object interactions. All the activities were examples from daily life, such as *discussing*, *telephone calls*, *giving an item*, etc. Full localization information with bounding boxes is provided as ground truth for each frame of each activity.

2.3.4. G3Di

G3Di [53](http://dipersec.king.ac.uk/G3D/) is a human interaction dataset for multiplayer gaming scenarios and was collected by the same group that colected G3D dataset at Kingston University in 2014. The dataset was captured using a gamesourcing approach where the users were recorded whilst playing computer games. This dataset contains 12 subjects split into 6 pairs. Each pair interacted through a gaming interface showcasing six sports involving several actions: boxing (*right punch, left punch, defend*), volleyball (*serve, overhand hit, underhand hit,* and *jump hit*), football (*kick, block* and *save*), table tennis (*serve, forehand hit* and *backhand hit*), sprint (*run*) and hurdles (*run* and *jump*). Most sequences contain multiple action classes in a controlled indoor environment with a fixed camera. Similar to G3D, action point and action segment are provided as ground truth.

2.3.5. Office Activity dataset

Office Activity dataset [54](http://vision.sysu.edu.cn/projects/3d-_activity/) was collected by Sun Yat-Sen University in 2014 aimed at complex activities that may typify an office environment. Three RGB-D cameras were set up in two scenes and at different viewpoints within the scene to capture activities in multiple views. The dataset consists of two parts: OA1 and OA2. In OA1, each activity was performed by a single subject. Five subjects performed 10 classes of activities, namely *answering-phones, arranging-files, eating, moving-objects, going-to-work, finding-objects, mopping, sleeping, taking-water, wandering.* The activities in OA2 are interactive activities performed by two subjects, and include *asking-and-away, called-away, carrying, chatting, delivering, eating-and-chating, having-guest, seeking-help, shaking-hands, showing.* In total, there are 1180 RGB-D activity sequences in Office Activity dataset.

2.3.6. M^2I dataset

The M^2I dataset [55](http://media.tju.edu.cn/tju_dataset.html) was captured by Tianjin University in 2015. This dataset contains both human-object interactive actions and human-human interactive actions captured from two different views. The human-object interactive actions include: *throwing basketball, bouncing basketball, twirling hula-hoop, tennis swing, tennis serve, calling cellphone, drinking water, taking photos, sweeping the floor, cleaning the desk, playing guitar, playing football, passing basketball,* and *carrying box*, where the last three actions were performed by two people. The human-human interactive actions include: *walking, crossing, waiting, chatting, hugging, handshaking, high-fives, bowing,* and *boxing.* Each human-object interaction was performed by 22 persons twice and they represent both daily life and sport actions. Each human-human interactive action was performed by 20 groups (two persons in a group) with 2 repetitions. This dataset contains 1760 action samples in total. The RGB, depth, human mask, and skeleton data are all available.

ShakeFive Dataset [56](`http://www.projects.science.uu.nl/shakefive/`), collected by Universiteit Utrecht in 2014, is a dyadic interactions dataset, which contains only two actions, namely *hand shake* and *high five*. This dataset is aimed at algorithms designed to recognize fine-grained interactions and consists of 100 RGB videos along with Kinect [TM]skeleton measurements for each subject. Fifty-seven videos contain *hand shake* interactions and 43 contain *high five* interactions. Metafiles provided store the ground truth, which contain frame numbers, twenty skeleton joint positions per person, and one of 5 possible labels describing the interaction in the frame: *standing*, *approaching*, *hand shake*, *high five* and *leaving*.

Dataset	Year(Cited[3])	Modality	#a,#s,#e	Protocol
SBU [50]	2012(33)	C,D,S	8,7,300	5-fold CV
K3HI [51]	2013(5)	S	8,15,320	4-fold CV
LIRIS [52]	2014(2)	C,D,G	10,21,828	1. D1 (305 action samples training, 156 samples test)
				2. D2 (242 action samples training, 125 samples test)
G3Di [53]	2014(0)	C,D,S	6,12,72	LOSubO
Office Activity [54]	2014(0)	C,D,S	20(10OA1+10OA2), 5,1180	5-fold cross validation
M^2I TJU [55]	2015(0)	C,D,S,HM	22,20,1760	-
ShakeFive [56]	2014(0)	C,S	2,37,100	1. 75% training (4-fold CV)
				2. 25% training (4-fold CV)

Table 3: Summary of the key specifications of the human-human interaction and multi-person action/activity datasets. Notation for the header: #a: number of actions, #s: number of subjects, #e: number of total examples. Notation for data format: C: Colour, D: Depth, S: Skeleton, G: Grayscale, HM: Human Mask. Notation for protocol: LOSubO: Leave One Subject Out, CV: Cross Validation

3. Analysis

The analysis presented in this section is framed by consideration for (i) the category of application scenarios, (ii) characteristics of dataset acquisition and presentation format, (iii) dependence of algorithm evaluation on dataset acquisition modes, (iv) complexity of the environmental factors inherent in dataset, (v) evaluation protocols commonly used for algorithm development and testing, and (vi) state-of-the-art results obtained to date with the datasets. Naturally, the discussions invite some recommendations and they are provided appropriately.

3.1. Application scenarios

The creation of a given dataset is usually motivated and targeted at some real-world applications. Lun et al. [10] summarized the major applications from the algorithm development perspective in [10]. In this

---

[3]Citations are as of 31 August 2015

paper, two broad categories of applications are identified and they are characterized by the types of actions in the dataset or the description provided by the dataset creators. The first category is human-computer interaction (HCI), example applications include video game interface and device control. The second category is daily activity (DA), including scene surveillance, elderly monitoring, service robotics, E-healthcare and smart rooms. Ostensibly, the various datasets model the applications well, but the various environmental factors and the size of examples need to be considered in determining how well a dataset mimics reality. Table 4 (columns one and two) presents a summary of the datasets reviewed and the target applications.

3.2. Characteristics of dataset acquisition

The characteristics of the dataset acquisition modes and the presentation format has bearing on how algorithms can use them for evaluation without repurposing. A set of *de facto* standard acquisition modes and presentation formats potentially provide a basis for objective comparative evaluation of algorithms. Based on the datasets reviewed, four modes of acquisition and presentation along with two modes that are variations of the third and fourth modes can be identified. They are listed below with some explanations:

- Mode 1: Captured as action samples and stored in segments where each segment contains only one action or activity.
- Mode 2: Captured as activity samples, but each activity contains a continuous sequence of labelled sub-activities.
- Mode 3: Captured as sequences of actions where the order of the actions in each sequence is fixed. The data is stored in sequential fashion and action segment points are provided.
- Mode 4: Captured as sequences of actions where the order of actions in each sequence is random. The data are stored in sequential fashion and action segment points are provided.
- Mode 3*: Captured as in Mode 3, but stored and presented as in Mode 1 after some processing.
- Mode 4*: Captured as in Mode 4, but stored and presented as in Mode 1 after some processing.

Table 4 (columns one and four) presents a summary of the datasets reviewed and the acquisition mode.

3.3. Algorithm evaluation and dataset acquisition modes

The development and implementation of a given application may require several algorithms and these will need to be evaluated objectively. Based on the acquisition and presentation modes, and available ground truth labels, the datasets can be used for testing five identifiable types of algorithms. These include action recognition, action detection, falling detection and online action recognition. Detailed explanations are provided as follows.

Action Recognition: In this paper, action recognition and action categorization are synonymous and we assume that a unique label can represent the entire video sequence. This casts the human action recognition problem as a classification problem.Datasets captured and presented in Mode 1, as well as Mode 3* and Mode 4*, can be directly used for action recognition. The datasets presented in other

modes can also be used for action recognition after some processing, e.g. segmenting sequence into action samples using the ground truth action segment points.

Action Detection: This focuses on identifying the occurrence of specific actions in an observed sequence.Thus, to test action detection algorithms the dataset should be captured continuously and provide accurate ground truth segmentation points of each action. Only the datasets captured in Modes 2, 3 and 4 can be used for action detection.

Falling Detection: This is an important but specific type of action detection which only focuses on falling event. Its importance has risen because of the potential application in health monitoring. A dataset meant for the evaluation of *falling detection* algorithm should be captured in similar modes as *action detection* but should also contain falling events and possibly other actions that are easily confused with falling actions.

Online Action Recognition: For the evaluation of online action recognition algorithms, the dataset must mimic the realistic scenario where unlabeled video sequence are continuously presented. Additionally, the actions should also be performed in random order. Datasets captured in Mode 4 are the only ones suitable for this category of algorithms.

3.4. Complexity of the environmental factors inherent in datasets

The comparative performance of a given algorithm depends on the environmental factors that are represented in the dataset being used for evaluation. Incidentally, the degree of complexity of the factors should also be considered. For example, a dataset with fixed but cluttered background may not be as challenging as one where the cluttered background varies from sample to sample. To judge the degree of challenege posed by a dataset consideration should be given to the complexity of the actions performed and the attending environmental factors. Ramanathan et al. [57] identified some of these factors as execution rate, anthropomorphic variations, viewpoint variation, occlusion, cluttered background, and camera motion. In order to evaluate an algorithm targeted at real-world applications, a good dataset should represent some of these factors and exercise the robustness of the algorithm. Ideally, the dataset should model the real-world application.

Most of the reviewed RGB-D datasets include execution rate and anthropomorphic variations to some extent, since these factors can be achieved by employing different individuals and several repetition. However, viewpoint variation is only found in multi-view dataset. Only small subset of the datasets include occlusion and cluttered background. The lack of occlusion and acquisition in relatively simple background limits the usefulness of any dataset in the design of realistic algorithms. Camera motion is not frequently found in RGB-D-based action datasets. Although the location and orientation of camera were not fixed in DMLSmartActions dataset, the camera was static during data capture and cannot be regarded as camera motion. Only LIRIS dataset incorporates camera motion because the camera was mounted on a mobile robot. Apart from these common challenges that are also typical of 2D video datasets, another issue related to RGB-D-based action dataset is the useful range (for depth data) of the Kinect $^{\text{TM}}$camera. This limitation

Single view	Applications	Algorithm Evaluation	Data acquisition/presentation	Ground truth
MSRAction3D	HCI	AR	Mode 1	AN
RGBD-HuDaAct	DA	AR	Mode 1	AN
CAD-60	DA	AR/AD	Mode 1	AN
MSRC-12	HCI	AR	Mode 3	AN/ASP/TD
MSRDaily	DA	AR	Mode 1	AN
UTKinect	HCI	AR	Mode 3	AN/ASP
G3D	HCI	AR	Mode 2	AN/SAN/SASP
DHA	HCI	AR	Mode 1	AN
Falling Event	DA	FD	Mode 1	AN
MSRActionPair	DA	AR	Mode 1	AN
CAD-120	DA	AR/AD/ObT	Mode 2	AN/SAN/SASP/OL
WorkoutSU-10	DA	AR	Mode 1	AN/TD
Concurrent Action	DA	AR/OAR	Mode 4	AN/ASP
IAS-lab	DA	AR	Mode 1	AN
UCFKinect	HCI	AR	Mode 1	AN
Osaka	HCI	AR	Mode 1	AN
Morning-Routine	DA	AR/AD/ObT	Mode 3	AN/ASP/ASL
RGBD-SAR	DA	AR	Mode 1	AN
Mivia	DA	AR	Mode 1	AN
UPCV	DA	AR	Mode 1	AN
TJU	HCI	AR	Mode 1	AN
MAD	HCI	AR/AD	Mode 3	AN/ASP
Composable	DA	AR/AD	Mode 2	AN/SAN/SASP/ArLg
3D Online	DA	AR/OAR/ObT	Mode 1/ Mode 4	AN/OL/ASP
RGB-D activity	DA	AR/AD	Mode 4	AN per frame
UTD-MHAD	HCI	AR	Mode 1	AN
SYSU	DA	AR	Mode 1	-

Multi-view	Applications	Algorithm Evaluation	Data acquisition/presentation	Ground truth
ATC4^2	DA	AR/FD	Mode 1	AN
Falling Detection	DA	FD	Mode 4	FSP
Berkeley MHAD	HCI	AR	Mode 1	AN
DMLSmartActions	DA	AR/OAR	Mode 4	AN/ASP
ReadingAct	DA	AR	Mode 1	-
Multiview 3D Event	DA	AR/AD/ObT	Mode 3*	AN/OL
Northwestern-UCLA	DA	AR	Mode 1	AN
UWA3D Multiview	DA/HCI	AR	Mode 4*	AN
Multi-view TJU	HCI	AR	Mode 1	AN
NJUST	HCI	AR	Mode 1	AN

Multi-person	Applications	Algorithm Evaluation	Data acquisition/presentation	Ground truth
SBU	DA	AR	Mode 1	AN
K3HI	DA	AR	Mode 1	AN
LIRIS	DA	AR	Mode 1	AN/ASL
G3Di	HCI	AR/AD	Mode 3	AN/ASP/AP
Office Activity	DA	AR	Mode 1	AN
3M TJU	DA/HCI	AR	Mode 1	AN
ShakeFive	DA	AR	Mode 1	AN

Table 4: Real world applications and algorithm evaluations. Notation for real world application: DA: Daily Activity; HCI: Human Computer Interaction. Notation for algorithm evaluations: AR: Action Recognition; ObT: Object Tracking; AD: Action Detection; OAR: Online Action Recognition; FD: Falling Detection. Notation for ground truth: AN: Action Name; ASP: Action Segment Point; TD: Text Description; SAN: Sub Action Label; SASP: Sub Action Segment Point; FSP: Falling Segment Point; ASL: Actor Spatial Location; ArLg: right or left Arm, right or left Leg; OL: Object Location; AP: Action Point.

has restricted the capture environment to indoors and hence also limits the usefulness of these datasets in testing algorithms meant to operate outdoor.

It is instructive to describe and assign level of complexity to a selection of these factors: background clutter and occlusion, kinematic complexity of the actions/activities, variability amongst the actions/activities within a dataset, execution speed and personal style, *composable actions*, and interactivity between human and objects. We define a composable action as one composed of two or more actions, which are recognisable actions in their own right. For example, *pick up& throw* and *high throw* are two individual actions contained in MSR Action 3D dataset, but *pick up& throw* contains *high throw*, which makes them confusable actions. Human-object interactivity is another important characteristic of a dataset because some algorithms may benefit from the objects that the actors interact with [58, 59, 22].

Table 5 summarizes the assignment of the level of complexity of environmental factors found in the datasets reviewed. The order of datasets are in chronological order. The first four factors could take on one of three levels of complexity (low, medium, and high) while the last two are binary valued (yes/no). The criteria for categorization are summarized as follows.

Background clutter and occlusion

- Low: the background is fixed and clean. There is no occlusion of the subjects.
- Medium: the background is fixed but is cluttered. Some occlusion of subjects may be present.
- High: the background is not fixed among action samples and/or is cluttered. Occlusions are present and the actions may be affected by the background and occlusion.

Kinematic complexity

- Low: the movements are relatively simple and with short duration.
- Medium: the movements are of medium complexity and the duration is longer than movements in the low level category.
- High: the movements are complex and with long duration.

Variability amongst actions

- Low: the variation of complexity levels amongst actions within a dataset is low.
- Medium: the variation of complexity levels amongst actions within a dataset is medium.
- High: the variation of complexity levels amongst actions within a dataset is high.

Execution rate

- Low: the variation in style of execution among different subjects or repetitions is low
- Medium: the variation in style of execution among different subjects or repetitions is medium.
- High: the variation in style of execution among different subjects or repetitions is high.

Composable actions: whether a dataset contain composable actions (Yes/No).

Human-object interaction: whether a dataset contain human-object interaction (Yes/No).

Single view	Background& occlusion	Kinematic complexity	Variability amongst actions	Execution rate	Composable actions	Object
MSRAction3D	Low	Low	Low	Low	Yes	No
RGBD-HuDaAct	Medium	Medium	Medium	High	Yes	Yes
CAD-60	Medium	Medium	Medium	Low	Yes	Yes
MSRC-12	No background	Low	Low	Medium	No	No
MSRDaily	Medium	High	Low	Low	No	Yes
UTKinect	Medium	Medium	Low	Medium	Yes	Yes
G3D	Low	Low	Low	Low	No	No
DHA	Low	Low	Low	Low	No	No
Falling Event	Low	Low	Low	Low	Yes	No
MSRActionPair	Low	Low	Low	Low	No	Yes
CAD-120	High	High	High	Medium	No	Yes
WorkoutSU-12	Low	Low	Low	Low	No	No
Concurrent Action	No background	Medium	High	High	No	No
IAS-lab	Low	Low	Low	Low	Yes	Yes
UCFKinect	No background	Low	Low	Low	No	No
Osaka	Low	Low	Low	Low	No	No
Morning-Routine	Medium	High	Medium	Only one subject	No	Yes
RGBD-SAR	High	High	Medium	High	No	Yes
Mivia	Low	Low	Low	Low	No	Yes
UPCV	No background	Low	Low	Low	No	No
TJU	Low	Low	Low	Low	No	No
MAD	Low	Low	Low	Low	No	No
Composable	Low	High	Medium	High	Yes	Yes
3D Online	Medium	Medium	Low	Medium	No	Yes
RGB-D activity	High	High	High	High	Yes	Yes
UTD-MHAD	Low	Low	Low	Low	Yes	No
SYSU	Not released	-	-	-	-	-
Multi view	**Background& occlusion**	**Kinematic complexity**	**Variability amongst actions**	**Execution rate**	**Compositional actions**	**Object**
ATC4^2	Low	Low	Low	Low	No	Yes
Falling Detection	High	Low	Low	Medium	Yes	Yes
Berkeley MHAD	Low	Low	Low	Low	No	No
DMLSmart	High	Low	Low	Medium	Yes	Yes
ReadingAct	Not released	-	-	-	-	-
Multiview 3D Event	High	Medium	Low	Low	No	Yes
Northwestern-UCLA	Low	Low	Low	Low	No	Yes
UWA3D Multiview	Low	Low	Low	Low	Yes	No
Multi-view TJU	Low	Low	Low	Low	No	No
NJUST	Low	Low	Low	Low	No	No
Multi person	**Background& occlusion**	**Kinematic complexity**	**Variability amongst actions**	**Execution rate**	**Compositional actions**	**Object**
SBU	Low	Low	Low	Low	No	No
LIRIS	High	High	High	High	Yes	Yes
K3HI	No background	Low	Low	Low	No	No
G3Di	Low	High	Low	Medium	No	No
Office Activity	High	Medium	Medium	High	No	Yes
M^2I TJU	Medium	Medium	Low	Low	No	Yes
ShakeFive	Low	Low	Low	Low	No	No

Table 5: Complexity level of the reviewed datasets from different aspects

3.5. Evaluation protocols

Careful design of the evaluation protocols is necessary to validate the results reported for each algorithm. Also important is the matching of the algorithm insofar as its purpose can be articulated, with the dataset represnting the enviromental factors that underpin the purpose. Several algorithms have been evaluated using the datasets reviewed in this paper. Using the algorithms that reported state-of-the-art results as a basis, a number of evaluation setup are found to be in common usage. They are listed and described below:

Leave-one-sequence-out cross validation setup: Randomly select one sequence from the entire dataset as test data and use the remaining sequences as training data. Perform a certain number of these tests and average the outcomes as the final result.

Leave-one-subject-out cross validation setup: Train with all but one subject and test with the unseen data. Repeat this for all subjects and report the average of the outcomes as the final result.

Cross-subject test: A number of the subjects are used for training and the remainder for testing.

- Select half of the subjects to be used for training and the remainder for testing. Some may use two-fold cross validation: repeat the evaluation using the previous test set as the training set and vice versa. The final result is the average of the two tests.
- Consider all the possible combinations of half subjects for training and the remaining for test.

Cross-view: Select one view as training set and the other views as test set. This only applies to multi-view datasets.

Cross-environment: Select the actions performed in one environment as training and test on actions performed in other environments. This is only applicable to datasets with specific actions captured in different environments.

3.6. State-of-the-art results

In this section, we tabulate the state-of-the-art methods[4] that used the reviewed datasets in order to highlight current status of research. For most of the datasets, we provide more than one algorithm because, not having used the same evaluation protocol, the qualifier "state-of-the-art" is not unequivocal. In addition, even when the same datasets and evaluation protocols have been used, the data modalities also need to be taken into consideration. This important observation has previously been ignored by researchers. There are instances where algorithms have been tested on skeleton data and claim of superior performance made over algorithms tested on depth data. In Tables 6, 7 and 8, we provide the state-of-the-art methods along with the reported results, the modalities of the algorithm used, and the protocol used for training and evaluation of the algorithms. The listing is in descending order of citation frequency of the original paper that published the datasets.

Dataset	State-of-the-art Methods	Acc.(%)	Data	Protocol

---

[4]Authors will maintain a website to keep updating the state-of-the-art results

Dataset	Method	Result	Modality	Protocol
MSR-Action3D [8]	1. TriViews +PFA [60] 2. Decision-Level Fusion (SUM) [61] 3. ConvNets [62, 63]	1. 98.2 2. 98.2 3. 100	1. D, S 2. D, S 3. D	1. CS (Half training, half test) 2. CS (2,3,5,7,9 subject training, 1,4,6,8,10 subject test) 3. CS (1,3,5,7,9 subject training, 2,4,6,8,10 subject test)
MSRDaily-Activity3D [14]	1. τ-test [64] 2. DL-GSGC +TPM [65] 3. 3D joint+CS-MLtp [66] 4. Depth-VSFR [67]	1. 95.63 2. 95 3. 92.5 4. 89.7	1. D,S 2. S 3. C,S 4. D	1. Not given 2. CS (Half training, half test) 3. CS (Half training, half test) 4. Not given
UTKinect [15]	1. Fused feature [68] 2. TriViews +PFA [60] 3. Grassman manifold [69]	1. 100 2. 98 3. 95.25	1. C, D, S 2. D, S 3. D	1. CS (Half training, half test) 2. CS (Half training, half test) 3. LOSubO
CAD-60 [12]	1. Decision-Level Fusion (Majority Voting) [61] 2. Pose Kinectic Energy [70] 3. SpatioTemporal Interest Pt [71]	1. 96.4(Prec.) 84.6(Rec.) 2. 93.8(Prec.) 94.5(Rec.) 3. 93.2(Prec.) 84.6(Rec.)	1. D, S 2. S 3. D	1. LOSubO(1,3,4 training, 2 test) 2. LOSubO 3. LOSubO
RGBD-HuDaAct [11]	1. BoW-Pyramid [72] 2. PA-Pooling [73]	1. 91.7 2. 85.9	1. C,D 2. C	1. LOSubO 2. LOSubO
MSRAction-Pair [21]	1. BHIM [74] 2. 3D Pose [75] 3. SNV [76]	1. 100 2. 99.4 3. 98.89	1. C, D 2. S 3. D	1. CS (First 5 test, rest training) 2. CS (Odd subjects training, even subjects test) 3. CS (First 5 test, rest training)
MSRC-12 gesture [13]	1. RDF-selected features [23] 2. Cov3DJ [77] 3. ESM(6 iconic gestures) [78]	1. 94.03 2. 93.6 & 91.7 3. 96.76	1. S 2. S 3. S	1. LOSubO(5-fold CV) 2. LOSubO(30-fold CV) &CS (half subjects training) 3. LOSubO
CAD-120 [22]	1. QQSTR-gt-tracks [79] 2. Skeleton feature+HMMs [80] 3. ATCRF [81]	1. 95.2(Activity Acc.) 95.2(Activity Prec.)95(Activity Rec.) 2. 94.4(Activity Acc.) 91.6(Sub-activity Acc.) 3. 93.5(Activity Acc.) 95(Activity Prec.) 93.3(Activity Rec.) 89.3(Sub-activity Acc.)	1. S 2. S 3. S	1. LOSubO (4-fold CV) 2. LOSubO (4-fold CV) 3. LOSubO (4-fold CV)
UCFKinect [27]	1. MvMF-HMM [82] 2. Hierarchical model [83] 3. Moving Pose [84]	1. 98.9 2. 98.7 3. 98.5	1. S 2. S 3. S	1. 4-fold CV 2. 2-fold CV 3. 4-fold CV
G3D [16, 17]	1. LRBM [85] 2. Clustered Action Manifolds [86]	1. 90.5(Acc.); 87.94(F score) 2. 97.8 (Fighting activity) (F-score)	1. S 2. S	1. CS (4 subjects training, 1 subjects validation, 5 subjects test) 2. LOSubO CV
Falling Event [20]	structure-motion [20]	98(insufficient illumination) &100(sufficient illumination)	S	50 samples training, rest 100 test
UPCV [32]	DS-SRC+DTW dissimilarity on annotated UPCV [32]	89.25	S	LOSubO

DHA [18]	1. MMJRR [87] 2. CHCRF [33] 3. DMPP_PHOG [87] 4. DLRMPP_PHOG [87]	1. 98.2 2. 95.9 3. 95 4. 95.6	1. C,D 2. C,D 3. D 4. C	1. LOSubO CV 2. CS (10 training,11 test) 3. LOSubO CV 4. LOSubO CV
WorkoutSU-10 [23]	1. Graph Mining [88] 2. Hyper-graph [89]	1. 99.6 2. 99.5	1. S 2. S	1. CS(6 subjects training, 6 test) CV 2. CS(6 subjects training, 6 test) CV
IAS-lab [25, 26]	1. SUMFLOW+PCA [26] 2. Skeleton joint position [26]	1. 85.2 2. 76.7	1. C,D 2. S	1. LOSubO 2. LOSubO
Osaka [28]	Dynamic features [28]	77.5	S	LOSubO CV
Mivia [31]	1. Edit distance(HARED) [90] 2. Deep learning [91]	1. 85.2 2. 84.7	1. D 2. D	1. LOSubO CV 2. LOSubO CV
Concurrent Action [24]	1. COA [24] 2. MIP [24] 3. Actionlet Esemble [14]	1. 88 2. 86 3. 84	1. S 2. S 3. S	1. Not given 2. Not given 3. Not given
3D Online [36]	1. Orderlets+Boosting [36] 2. Orderlets [36] 3. Orderlets [36]	1. 71.4 2. 66.1 3. 56.4	1. S 2. S 3. S	1. Same-Environment (2-fold CV) 2. Cross-Environment (S1, S2 training, S3 test) 3. Continuous (S1, S2, S3 training, S4 test)
MAD [34]	Event transition [92]	85.0(Frame-level Prec.); 71.41(Frame-level Rec.); 77.41(Frame-level F-score); 74.4(Event-level Prec.); 85.02(Event-level Rec.); 78.83(Event-level F-score);	S	5-fold CV (8 groups training, 2 groups test)
Composable [35]	Hierarchical model [35]	85.7	S	LOSubO
RGBD-SAR [30]	1. LDP [30] 2. DLMC-STIPS [11]	1. 83.5 2. 80.2	1. C,D 2. D	1. Not given 2. Not given
SYSU [38]	DS+DCP+DDP+JOULE-SVM [38]	84.89 & 79.63	C,D,S	Half sample training, rest test & CS
RGB-D activity [37]	CaTM [37]	1. office: 30.6(OffSeg Acc.); 32.9(OnSeg Acc.); 33.1(OffSeg Average Prec.); 34.6(OnSeg Average Prec.); 39.9(OffFr Acc.); 38.5(OnFr Acc.); 41.5(Patching Acc.) 2. kitchen: 33.2(OffSeg Acc.); 29.0(OnSeg Acc.); 26.4(OffSeg Average Prec.); 25.5(OnSeg Average Prec.); 37.5(OffFr Acc.); 34.0(OffFr Acc.); 20.5(Patching Acc.)	C,D,S	Training and test sets are specified by the author
UTD-MHAD [39]	DMM+CRC [39]	1. 79.1 2. 67.2 3. 66.1	1. D,I 2. I 3. D	CS (odd indexed subjects training, rest test)

Morning-Routine [29]	HHMM [29]	77.01	D	Not given

Table 6: Summary of state-of-the-art results with corresponding methods and settings on single-view action/activity datasets. Notation for data format: C: Colour, D: Depth, S: Skeleton, I: Inertial sensor signal. Notation for evaluation protocol: CS: Cross subject, LOSubO: Leave one subject out, CV: Cross validation. Notation for evaluation metric: Acc.: Accuracy, Prec.: Precision, Rec.: Recall, OffSeg: Offline Segmentation, OnSeg: Online Segmentation, OffFr: Offline Frame, OnFr: Online Frame.

Dataset	State-of-the-art Methods	Acc.(%)	Data	Protocol
Berkeley MHAD [42]	1. Hierarchy of LDSs(28 joints used) [93] 2. HBRNN-L(35 joints used) [94] 3. CNN(3 joints used) [95] 4. Feature-Level-Fusion+SRC (Kinect+Acc1&Acc4) [96] 5. HACK [97]	1. 100 2. 100 3. 98.28 4. 99.54 5. 97.7	1. S 2. S 3. S 4. D, A 5. D	1. CS (First 7 training, last 5 test) 2. CS (First 7 training, last 5 test) 3. 5-fold group-wise CV 4. LOSubO 5. LOSubO
ATC4^2 [40]	1. Depth-VSFR(All-view) [67] 2. Depth-VSFR(Cross-view) [67] 3. SSM(All-view) [98] 4. SSM(Cross-view) [98]	1. 85.5 2. 82.0 3. 83.4 4. 81.2	1. D 2. D 3. D 4. D	1. LOSubO CV 2. Cross-View (Training on one viewpoint, test on other viewpoints) 3. CS (15 training,5 test,10 fold CV) 4. CS (15 training,5 test,10 fold CV)
Falling Detection [41]	Bayesian framework [41]	92.3(Prec.) 100(Rec.)	D	Cross-view
Multiview 3D Event [45]	4DHOI [45]	87	C, D, S	Not given
Multi-View TJU [48]	MTSL+LL/ML [48]	1. 93.9(multi view); 91.4(front view); 90.7(side view) 2. 95.8(multi view); 94.6(front view); 92.5(side view)	1. D, S 2. C, S	1. CS(6 subjects training, 6 validation, 8 test) 2. CS(6 subjects training, 6 validation, 8 test)
Northwestern-UCLA [46]	1. MST-AOG [46] 2. MST-AOG [46] 3. MST-AOG [46] 4. NKTM [99] 5. NKTM [99] 6. NKTM [99]	1. 81.6 2. 79.3 3. 73.3 4. 75.8 5. 73.3 6. 59.1	1. C, S 2. C, S 3. C, S 4. C 5. C 6. C	1. LOSubO 2. Cross-environment 3. Cross-view(1,2 Camera training,3 Camera test) 4. Cross-view(1,2 Camera training,3 Camera test) 5. Cross-view(1,3 Camera training,2 Camera test) 6. Cross-view(2,3 Camera training,1 Camera test)
UWA3D Multiview [47]	1. Holistic HOPC(Same view) [47] 2. MSO-SVM(Cross view) [47]	1. 84.93 2. 91.79(0°), 86.67(−25°), 88.89(+25°), 75.56(−50°), 77.78(+50°)	1. D 2. D	1. CS (Half training, half test) 2. 0° training
NJUST [49]	1. ToSP+SVM [49] 2. BSC+Spatial-Temporal [100]	1. 98.4 2. 94.7	1. C,D 2. D	1. LOSubO 2. LOSubO

| DMLSmart Actions [43] | 1. SVM-NNSC + Proposed Kernel [101]
2. Meta Learning [102] | 1. 79.9
2. 77.19 | 1. HDC
2. C,D | 1. LOSubO
2. LOSubO |
| ReadingAct [44] | 1. BoW+χ^2 SVM [44]
2. BoW+Linear SVM [44] | 1. 90.4
2. 82.1 | 1. C
2. C,D | 1. CS (15 training,5 test,4-fold CV)
2. CS (15 training,5 test,4-fold CV) |

Table 7: Summary of state-of-the-art results with corresponding methods and protocols on multi-view action /activity datasets. Notation for data format: C: Colour, D: Depth, S: Skeleton, A: Acceleration, HDC: High Definition Colour. Notation for evaluation protocol: CS: Cross subject, LOSubO: Leave one subject out, CV: Cross validation. Notation for evaluation metric: Acc.: Accuracy.

Dataset	State-of-the-art Methods	Acc.(%)	Data	Protocol
SBU [50]	1. MaxEnt IOC [103] 2. DMDP [103, 104]	1. 0.52 (AFD); 80 (NLL) 2. 0.51 (AFD); 113.5 (NLL)	1. S 2. S	1. LOSubO (7-fold CV) 2. LOSubO (7-fold CV)
K3HI [51]	Positive action (Joint motion) [51]	75.6	S	4-fold CV
LIRIS [52]	1. Pose+Appearance +Context+Scene (With Localization) [105, 52] 2. Pose+Appearance +Context+Scene (Without Localization) [105, 52]	1. 74(Rec.); 41(Prec.); 53(F-score) 2. 63(Rec.); 33(Prec.); 44(F-score)	1. G, D 2. G, D	1. 305 action samples training, 156 samples test 2. 305 action samples training, 156 samples test
G3Di [53]	1. Action segment [53] 2. Action points [17]	1. Action: 56.1(F1); Interaction: 57.1(F1) 2. Action: 42.6(F1); Interaction: 44.8(F1)	1. S (Boxing) 2. S (Boxing)	1. LOSubO 2. LOSubO
Office Activity [54]	Structured deep architecture [54]	60.1(OA1); 45.0(OA2)	D	5-fold CV
M^2I TJU [55]	-	-	-	-
ShakeFive [56]	1. Dyadic poselets [56] 2. Dyadic poselets [56]	1. 49.56 (Handshake) 34.85 (Highfive) 2. 47.87 (Handshake) 23.94 (Highfive)	1. C, S 2. C, S	1. 75% training (4-fold CV) 2. 25% training (4-fold CV)

Table 8: Summary of state-of-the-art results with corresponding methods and protocols on human-human interaction and multi-person action/activity datasets. Notation for data format: C: Colour, D: Depth, S: Skeleton, G: Grayscale. Notation for evaluation protocol: LOSubO: Leave one subject out, CV: Cross validation. Notation for evaluation metric: Acc.: Accuracy, Prec.: Precision, Rec.: Recall, AFD: Average image Feature Distance, NLL: Negative Log-Likelihood.

3.7. Recommendations

The intensity of research activity in human action/activity recognition has encouraged the development of new algorithms and possibly the generation of new datasets. Based on our review, some newly collected datasets share similar characteristics with existing ones and may not have expanded the variety of environ-

mental factors inherent in the dataset. Perhaps more importantly, comparisons between algorithms evaluated on different datasets are in many cases unfair and makes the progress achieved to date unclear. Here, we make some recommendations on the issues of dataset selection and evaluation protocols.

3.7.1. Datasets

It is clear that each of the datasets are matched to a specific application and aspect of action/activity recognition. Inherent in each dataset are factors that the algorithm under evaluation is meant to accommodate. These factors include variation of execution rate and style of performance, degree of clutter in background and occlusion, multi view points, camera motion, action detection, and online learning. All of these factors have been analysed in Section 3.4.

Based on the analysis, below, we provide the list of environmetal factors and applications, along with the datasets that incorporate/are suitable for them as a guide in their selection.

Execution rate and anthropomorphic variation: RGBD-HuDaAct, MSRC-12, Concurrent action, RGBD-SAR, composable, DMLSmart, Multiview 3D Event, LIRIS, and Office Activity.

Cluttered background and occlusion: UTKinect, RGBD-HuDaAct, MSRDaily Activity, CAD-120, RGBD-SAR, 3D Online, DMLSmartActions, Multiview 3D Event, LIRIS, and Office Activity.

Multi viewpoints: ATC4^2, Falling Detection, Berkeley MHAD, DMLSmartActions, ReadingAct, Multi-view 3D Event, Northwestern-UCLA Multiview, UWA3D Multiview, Multi-view TJU, NJUST, M^2I TJU, and Office Activity.

Camera motion: LIRIS.

Action detection: CAD-60, CAD-120, MAD, Human Morning Routine, Composable, Multiview 3D Event, and G3Di.

Falling detection: Falling Detection, Falling Event, and ATC4^2.

Online action recognition: 3D Online, Concurrent Action, RGB-D activity, DMLSmartActions.

Object detection: CAD-120, Human Morning Routine, 3D Online, and Multiview 3D Event.

3.7.2. Evaluation protocols

This review suggests that the most widely adopted experimental set up in the state-of-the-art results are "leave-one-subject-out cross validation" and "cross-subject test". The fact that several datasets are released without an accompanying *de facto* standard evaluation protocol results in controversial comparisons among algorithms. For example, the summaries of evaluation protocols given in section 3.5 shows that the most commonly used cross-subject scheme has different splitting methods. Some papers used odd indexed subjects as training and even indexed subjects as test, others may use first half of subjects as training data and the remainder as test data. Some have used cross-validation on the split data and some have only reported the results on one test. There are some papers that did not provide explicit information on the evaluation protocol used.

We recommend that any new release of dataset should be accompanied by "standard" and unified evaluation protocols, that future proposed algorithms can use for design and performance evaluation. Admittedly, some applications may require specific evaluation methods different from those published with a given dataset. New evaluation protocols should be clearly articulated and provided with informative justification.

4. Discussion

In this section, we point out the limitations of both current RGB-D action datasets and commonly used evaluation protocols on action recognition. Our aim is to provide guidance on future creation of datasets and establishment of standard evaluation protocols for specific purposes.

4.1. Limitations of current datasets

The review and analysis of current RGB-D action datasets have revealed some limitations including size, applicability, availability of ground truth labels and evaluation protocols. There is also the problem of dataset saturation, a phenomenon whereby algorithms reported have achieved a near-perfect performance. We now elaborate on these limitations.

Dataset size: The most obvious limitation of current dataset is the small number of action classes and sample size. Current RGB-D based action datasets typically contain 10 to 20 actions, which is not comparable to those of 2D video action datasets. A newly released 2D dataset [106] on action recognition contains 203 distinct action classes in 849 hours of video recording. Another 2D dataset [107] on sport activities contains 1 million YouTube videos aggregating 487 classes. A possible reason is that it is relatively easy to "harvest" 2D videos from the Internet, e.g., YouTube. In contrast, the RGB-D based videos have to be captured manually and, the time, financial and labour constraints limit the size of RGB-D datasets.

Applicability: The applications of current RGB-D-based action datasets are also very limited because of the restricted types of actions represented in each dataset. Most RGB-D datasets are collected within lab environment and the execution style of actions generally follow strict instructions. Thus, even with different subjects, the variations in performing style are subtle and indiscernible.

Ground truth: Some of current datasets are well constructed with many challenging factors, however, they provide poor ground truth labels, which limits their usability.

Evaluation protocols: As analysed in Sections 3.5 and 3.7, the controversy of evaluation protocols may lead to unfair comparison among algorithms; a situation largely due to lack of clarity on the protocols to be used with published datasets.

Saturation: Section 3.7 has provided recommendations on the selection of dataset for different purposes, suggesting that current datasets already represent the environmental factors required to rigorously test and evaluate different algorithms. However, based on the state-of-the-art results summarised in Section 3.6, it can be seen that algortihms have already achieved a near-perfect accuracy on some

30

modalities of these datasets. This suggests that these datasets are near saturated. This phenomenon obscures the fact that algorithms may not yet be suitable for deployment in real-world applications. It is necessary that the set of environmental factors and their level of complexities (Section 3.4), are matched to real-world applications and, guide the creation of new and challenging RGB-D action dataset.

4.2. Recommendations for future datasets

Based on the limitations identified above we provide some recommendations on creating future datasets. The number of samples and variety of action types needs to be increased so that a learning algorithm may generalize on the problem domain. Algorithms are destined for inclusion in some real-world applications and as such dataset creators may need to focus on specific applications and the inherent environmental factors. This will allow the creation of datasets with realistic and free-form performance of actions that properly model the problem. The proliferation of datasets has its advantage namely, opportunity to expand the test and evaluation suite. However, there is opportunity to create sequentially captured and randomly performed RGB-D action recognition dataset. The ground truth will then be the action segment points. Such dataset will be an all-in-one testing suite for different algorithms - action categorization, action detection and online recognition. Apart from the provision of action segement as ground truth, actor and object locations along with any other informative metadata should be provided along with the dataset.

Finally, a dataset should be published with a number of standard evaluation protocols for use in the design, testing and fair comparative evaluation of future algorithms. Perhaps more importantly, the evaluation protocols should match real-world applications expectation. For example, in video surveillance applications, the cross-subject scheme is more appropriate than leave-one-sequence-out scheme. However, in health monitoring applications, as the system only monitors specific subject without new subjects, the leave-one-sequence-out scheme is more appropriate.

4.3. Limitations of evaluation protocols

Incidentally, the limitations of evaluation protocols may impede the progress of action recognition algorithms towards maturity and robustness for real-world applications. Currently, the most widely adopted experimental settings are leave-one-subject-out cross validation set-up and cross-subject set-up. However, these settings are not without controversy from the real-world application perspectives. In most of the datasets, the cameras are fixed and background would not have changed during data capture. Furthermore, within a specific dataset the instructions for performing the actions are fixed and all subjects usually performed actions from a fixed location in a scene. These issues may limit the robustness of algorithms if cross-subject or leave-one-subject-out cross validation schemes are used. One reason adduced for this limitation is that algorithms may inadvertently rely on the background information or the position of actors. Hence, the algorithms tested using these protocols can only be used on particular real world applications where the background and camera are fixed.

To some degree, the cross-view and cross-environment protocols are more realistic than leave-one-subject-out and cross-subject versions. These protocols consider the variation of viewpoints and surrounding environments of the performed actions. However, those protocols can only be used with specific datasets having multi-view points or multiple capture environment. Moreover, these protocols retain the problem associated with similar performance styles between training and test set. They are limited to one dataset in which the actions are performed under identical instructions.

4.4. Recommendations for future evaluation protocols

As mentioned in Section 4.2, evaluation protocols should correspond to specific real-world applications. The cross-subject, leave-one-subject-out cross validation, cross-view, and cross-environment schemes can either only be used with specific datasets or for particular applications.

To overcome the drawbacks of current evaluation schemes, we advocate the use of cross-dataset evaluation scheme. In a cross-dataset set-up, the actors, view point, environment, and manners of performing actions in training and test data are all different. Furthermore it is not limited to a specific dataset, since any group of datasets that share similar actions and semantics can be used. Perhaps more importantly, the cross-dataset evaluation scheme is more akin to real-world applications where the system trained on particular scenario can be used in other similar scenarios without the need to retrain the whole system.

The cross-dataset scheme has already been adopted on some algorithms for action recognition in 2D videos [108] [109], however, to our best knowledge, there is no report of its usage on RGB-D video datasets. Such a protocol requires the algorithm to be robust and able to accommodate the various environmental factors in order to consistently perform well.

It is interesting to note that in the evaluation scheme it is common to report the average of several runs. While this is a good statistical practice, we notice that such avearges are compared straightforwardly with results from existing algorithms without a test of the statistical significance of the observed difference. Perhaps, in line with protocols of well designed statistical experiments, the results reported for action recogniton algorithms should also include statistical significance tests[110].

5. Conclusion

A comprehensive review of commonly used and publicly available RGB-D-based datasets for action recognition has been provided. The detailed descriptions and analysis, highlights of their characteristics and potential applications should be useful for researchers designing action recognition algorithms. This is especially so, when selecting datasets for algorithm development and evaluation as well as creating new datasets to fill identified gaps. Most of the datasets collected to date are meant for algorithms devised to solve specific action recognition problem. However, the simplicity of the datasets have resulted in a "saturated" state whereby algorithmic improvement has stalled. A more realistic collection of datasets representing a

broad selection of challenging enviromental factors is now required. We have advocated the use of cross-dataset evaluation set up to provide a more realistic testing scenario. Furthermore, we advoacted the use of evaluation protocol that include statistical significance test to ensure fair comparision amongst algorithms. Meanwhile, the state-of-the-art results over the datasets we reviewed have been provided in one place to help researchers when configuring their comparative evaluation schedule. We also summarise several commonly used evaluation and validation set-ups and address their drawbacks, resulting in a set of recommendations on future collection of datasets and use of evaluation protocols.

This review has highlighted the need for comprehensive statistically significant evaluation protocols as part of algorithm development and testing. We are working on publshing an open-source software suite that will enable easy evaluation of action recognition algorithms, especially with cross dataset schemes.

References

References

[1] S. Vishwakarma, A. Agrawal, A survey on activity recognition and behavior understanding in video surveillance, The Visual Computer 29 (10) (2013) 983–1009.

[2] C. H. Lim, E. Vats, C. S. Chan, Fuzzy human motion analysis: A review, Pattern Recognition 48 (5) (2015) 1773 – 1796.

[3] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, Pattern recognition 36 (3) (2003) 585–601.

[4] G. Guo, A. Lai, A survey on still image based human action recognition, Pattern Recognition 47 (10) (2014) 3343–3361.

[5] T. Hassner, A critical review of action recognition benchmarks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 245–250.

[6] J. M. Chaquet, E. J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Computer Vision and Image Understanding 117 (6) (2013) 633 – 659.

[7] S. Ruffieux, D. Lalanne, E. Mugellini, O. A. Khaled, A survey of datasets for human gesture recognition, in: M. Kurosu (Ed.), Human-Computer Interaction. Advanced Interaction Modalities and Techniques, Vol. 8511 of Lecture Notes in Computer Science, 2014, pp. 337–348.

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

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

[10] R. Lun, W. Zhao, A survey of applications and human motion recognition with microsoft kinect, International Journal of Pattern Recognition and Artificial Intelligence 29 (5).

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

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

[13] S. Fothergill, H. Mentis, P. Kohli, S. Nowozin, Instructing people for training gestural interactive systems, in: Proc. SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 1737–1746.

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

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

[16] V. Bloom, D. Makris, V. Argyriou, 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, 2012, pp. 7–12.

[17] V. Bloom, V. Argyriou, D. Makris, Dynamic feature selection for online action recognition, in: Human Behavior Understanding, Vol. 8212 of Lecture Notes in Computer Science, 2013, pp. 64–76.

[18] Y. C. Lin, M. C. Hu, W. H. Cheng, Y. H. Hsieh, H. M. Chen, Human action recognition and retrieval using sole depth information, in: Proc. ACM International Conference on Multimedia, 2012, pp. 1053–1056.

[19] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Proc. IEEE International Conference on Computer Vision, 2005, pp. 1395–1402.

[20] C. Zhang, Y. Tian, RGB-D camera-based daily living activity recognition, Journal of Computer Vision and Image Processing 2 (4).

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

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

[23] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, A. Erçil, A decision forest based feature selection framework for action recognition from RGB-depth cameras, in: Image Analysis and Recognition, Springer, 2013, pp. 648–657.

[24] P. Wei, N. Zheng, Y. Zhao, S.-C. Zhu, Concurrent action detection with structural prediction, in: Proc. IEEE International Conference on Computer Vision, 2013, pp. 3136–3143.

[25] M. Munaro, G. Ballin, S. Michieletto, E. Menegatti, 3D flow estimation for human action recognition from colored point clouds, Biologically Inspired Cognitive Architectures 5 (2013) 42 – 51, extended versions of selected papers from the Third Annual Meeting of the BICA Society.

[26] M. Munaro, S. Michieletto, E. Menegatti, An evaluation of 3D motion flow and 3D pose estimation for human action recognition, in: RSS Workshops: RGB-D: Advanced Reasoning with Depth Cameras, 2013.

[27] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola Jr, R. Sukthankar, Exploring the trade-off between accuracy and observational latency in action recognition, International Journal of Computer Vision 101 (3) (2013) 420–436.

[28] A. Mansur, Y. Makihara, Y. Yagi, Inverse dynamics for action recognition, IEEE Transactions on Cybernetics 43 (4) (2013) 1226–1236.

[29] M. Karg, A. Kirsch, Simultaneous plan recognition and monitoring (SPRAM) for robot assistants, in: Proc. Human Robot Collaboration Workshop at Robotics Science and Systems Conference, 2013.

[30] Y. Zhao, Z. Liu, H. Cheng, RGB-depth feature for 3D human activity recognition, Communications, China 10 (7) (2013) 93–103.

[31] V. Carletti, P. Foggia, G. Percannella, A. Saggese, M. Vento, Recognition of human actions from RGB-D videos using a reject option, in: New Trends in Image Analysis and Processing, Vol. 8158 of Lecture Notes in Computer Science, 2013, pp. 436–445.

[32] I. Theodorakopoulos, D. Kastaniotis, G. Economou, S. Fotopoulos, Pose-based human action recognition via sparse representation in dissimilarity space, Journal of Visual Communication and Image Representation 25 (1) (2014) 12 – 23, visual Understanding and Applications with RGB-D Cameras.

[33] A. Liu, W. Nie, Y. Su, L. Ma, T. Hao, Z. Yang, Coupled hidden conditional random fields for RGB-D human action recognition, Signal Processing.

[34] D. Huang, S. Yao, Y. Wang, F. D. L. Torre, Sequential max-margin event detectors, in: Computer Vision - ECCV 2014, Vol. 8691 of Lecture Notes in Computer Science, 2014, pp. 410–424.

[35] I. Lillo, A. Soto, J. C. Niebles, Discriminative hierarchical modeling of spatio-temporally composable human activities, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 812–819.

[36] G. Yu, Z. Liu, J. Yuan, Discriminative orderlet mining for real-time recognition of human-object interaction, in: Computer Vision–ACCV 2014, Vol. 9007 of Lecture Notes in Computer Science, Springer, 2015, pp. 50–65.

[37] C. Wu, J. Zhang, S. Savarese, A. Saxena, Watch-n-patch: Unsupervised understanding of actions and relations, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4362–4370.

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

[39] C. Chen, R. Jafari, N. Kehtarnavaz, 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.

[40] Z. Cheng, L. Qin, Y. Ye, Q. Huang, Q. Tian, Human daily action analysis with multi-view and color-depth data, in: Computer Vision - ECCV 2012. Workshops and Demonstrations, Vol. 7584 of Lecture Notes in Computer Science, 2012, pp. 52–61.

[41] Z. Zhang, W. Liu, V. Metsis, V. Athitsos, A viewpoint-independent statistical method for fall detection, in: Proc. International Conference on Pattern Recognition, IEEE, 2012, pp. 3626–3630.

[42] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Berkeley MHAD: A comprehensive multimodal human action database, in: Proc. IEEE Workshop on Applications of Computer Vision, 2013, pp. 53–60.

[43] S. M. Amiri, M. T. Pourazad, P. Nasiopoulos, V. C. M. Leung, Non-intrusive human activity monitoring in a smart home environment, in: Proc. International Conference on e-Health Networking, Applications Services (Healthcom), 2013, pp. 606–610.

[44] L. Chen, H. Wei, J. Ferryman, Readingact RGB-D action dataset and human action recognition from local features, Pattern Recognition Letters 50 (2013) 159 – 169, depth Image Analysis.

[45] P. Wei, Y. Zhao, N. Zheng, S. C. Zhu, Modeling 4D human-object interactions for event and object recognition, in: Proc. IEEE International Conference on Computer Vision, 2013, pp. 3272–3279.

[46] J. Wang, X. Nie, Y. Xia, Y. Wu, S. Zhu, Cross-view action modeling, learning and recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656.

[47] H. Rahmani, A. Mahmood, D. Q. Huynh, A. Mian, HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition, in: Computer Vision - ECCV 2014, Vol. 8690 of Lecture Notes in Computer Science, 2014, pp. 742–757.

[48] A. Liu, Y. Su, P. Jia, Z. Gao, T. Hao, Z. Yang, Multipe/single-view human action recognition via part-induced multitask structural learning, IEEE transactions on cybernetics 45 (6) (2015) 1194–1208.

[49] Y. Song, J. Tang, F. Liu, S. Yan, Body surface context: A new robust feature for action recognition from depth videos, IEEE Transactions on Circuits and Systems for Video Technology 24 (6) (2014) 952–964.

[50] K. Yun, J. H. D. C. T. L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 28–35.

[51] T. Hu, X. Zhu, W. Guo, K. Su, Efficient interaction recognition through positive action representation, Mathematical Problems in Engineering 2013 (2013) 1–13.

[52] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C. E. Bichot, C. Garcia, B. Sankur, Evaluation of video activity localizations integrating quality and quantity measurements, Computer Vision and Image Understanding 127 (2014) 14 – 30.

[53] V. Bloom, V. ArgyrV., D. Makris, G3Di: A gaming interaction dataset with a real time detection and evaluation framework, in: Computer Vision-ECCV 2014 Workshops, Vol. 8925 of Lecture Notes in Computer Science, 2014, pp. 698–712.

[54] K. Wang, X. Wang, L. Lin, M. Wang, W. Zuo, 3D human activity recognition with reconfigurable convolutional neural networks, in: Proc. ACM International Conference on Multimedia, 2014, pp. 97–106.

[55] N. Xu, A. Liu, W. Nie, Y. Wong, F. Li, Y. Su, Multi-modal & multi-view & interactive benchmark dataset for human action recognition, in: Proc. ACM International Conference on Multimedia, 2015.

[56] C. van Gemeren, R. T. Tan, R. Poppe, R. C. Veltkamp, Dyadic interaction detection from pose and flow, in: Human Behavior Understanding, Vol. 8749 of Lecture Notes in Computer Science, 2014, pp. 101–115.

[57] M. Ramanathan, W. Y. Yau, E. K. Teoh, Human action recognition with video data: Research and evaluation challenges, IEEE Transactions on Human-Machine Systems 44 (5) (2014) 650–663.

[58] A. Gupta, A. Kembhavi, L. S. Davis, Observing human-object interactions: Using spatial and functional compatibility for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1775–1789.

[59] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 17–24.

[60] W. Chen, G. Guo, Triviews: A general framework to use 3D depth data effectively for action recognition, Journal of Visual Communication and Image Representation 26 (0) (2015) 182 – 191.

[61] Y. Zhu, W. Chen, G. Guo, Fusing multiple features for depth-based action recognition, ACM Transactions on Intelligent Systems and Technology 6 (2) (2015) 18:1–18:20.

[62] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, P. O. Ogunbona, Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring, in: Proc. ACM international conference on Multimedia (ACM MM), 2015, pp. 1119–1122.

[63] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, Human-Machine Systems, IEEE Transactions on PP (99) (2015) 1–12. doi:10.1109/THMS.2015.2504550.

[64] C. Lu, J. Jia, C. K. Tang, Range-sample depth feature for action recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 772–779.

[65] J. Luo, W. Wang, H. Qi, Group sparsity and geometry constrained dictionary learning for action recognition from depth maps, in: Proc. International Conference on Computer Vision, 2013, pp. 1809–1816.

[66] J. Luo, W. Wang, H. Qi, Spatio-temporal feature extraction and representation for RGB-D human action recognition, Pattern Recognition Letters 50 (2014) 139 – 148, depth Image Analysis.

[67] S.-S. Cho, A-Reum Lee, H.-I. Suk, J.-S. Park, , S.-W. Lee, Volumetric spatial feature representation for view-invariant human action recognition using a depth camera, Optical Engineering 54 (3).

[68] J. Ye, K. Li, G.-J. Qi, K. A. Hua, Temporal-order preserving dynamic quantization for human action recognition from multimodal sensor streams, in: Proc. International Conference on Multimedia Retrieval, ACM, 2015.

[69] R. Slama, H. Wannous, M. Daoudi, Grassmannian representation of motion depth for 3D human gesture and action recognition, in: Proc. International Conference on Pattern Recognition, 2014, pp. 3499–3504.

[70] J. Shan, S. Akella, 3D human action segmentation and recognition using pose kinetic energy, in: Proc. IEEE Workshop on Advanced Robotics and its Social Impacts, 2014, pp. 69–75.

[71] Y. Zhu, W. Chen, G. Guo, Evaluating spatiotemporal interest point features for depth-based action recognition, Image and Vision Computing 32 (8) (2014) 453 – 464.

[72] J. S. Tsai, Y. P. Hsu, C. Liu, L. C. Fu, An efficient part-based approach to action recognition from RGB-D video with bow-pyramid representation, in: Proc. International Conference on Intelligent Robots and Systems, 2013, pp. 2234–2239.

[73] B. Ni, P. Moulin, S. Yan, Pose adaptive motion feature pooling for human action analysis, International Journal of Computer Vision (2014) 1–20.

[74] Y. Kong, Y. Fu, Bilinear heterogeneous information machine for RGB-D action recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1054–1062.

[75] A. Eweiwi, M. Cheema, C. Bauckhage, J. Gall, Efficient pose-based action recognition, in: Computer Vision – ACCV 2014, Vol. 9007 of Lecture Notes in Computer Science, Springer, 2015, pp. 428–443.

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

[77] M. E. Hussein, M. Torki, M. A. Gowayyed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2013, pp. 2466–2472.

[78] H.-J. Jung, K.-S. Hong, Enhanced sequence matching for action recognition from 3D skeletal data, in: Computer Vision – ACCV 2014, Vol. 9007 of Lecture Notes in Computer Science, Springer, 2015, pp. 226–240.

[79] J. Tayyub, A. Tavanai, Y. Gatsoulis, A. Cohn, D. Hogg, Qualitative and quantitative spatio-temporal relations in daily living activity recognition, in: Computer Vision – ACCV 2014, Vol. 9007 of Lecture Notes in Computer Science, 2015, pp. 115–130.

[80] A. Taha, H. H. Zayed, M. Khalifa, E.-S. M. El-Horbaty, Skeleton-based human activity recognition for video surveillance, International Journal of Scientific & Engineering Research 6 (1).

[81] H. S. Koppula, A. Saxena, Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation, in: Proc. International Conference on Machine Learning, Vol. 28, 2013, pp. 792–800.

[82] J. Beh, D. K. Han, R. Durasiwami, H. Ko, Hidden markov model on a unit hypersphere space for gesture trajectory recognition, Pattern Recognition Letters 36 (2014) 144–153.

[83] X. Jiang, F. Zhong, Q. Peng, X. Qin, Robust action recognition based on a hierarchical model, in: Proc. International Conference on Cyberworlds, IEEE, 2013, pp. 191–198.

[84] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection, in: Proc. IEEE International Conference on Computer Vision, 2013, pp. 2752–2759.

[85] S. Nie, Z. Wang, Q. Ji, A generative restricted boltzmann machine based method for high-dimensional motion data modeling, Computer Vision and Image Understanding.

[86] V. Bloom, D. Makris, V. Argyriou, Clustered spatio-temporal manifolds for online action recognition, in: Proc. International Conference on Pattern Recognition, IEEE, 2014, pp. 3963–3968.

[87] Z. Gao, H. Zhang, G. Xu, Y. Xue, Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition, Neurocomputing 151, Part 2 (2015) 554 – 564.

[88] O. Çeliktutan, C. B. Akgul, C. Wolf, B. Sankur, Graph-based analysis of physical exercise actions, in: Proc. ACM international workshop on Multimedia indexing and information retrieval for healthcare, 2013, pp. 23–32.

[89] O. Çeliktutan, C. Wolf, B. Sankur, E. Lombardi, Fast exact hyper-graph matching with dynamic programming for spatio-temporal data, Journal of Mathematical Imaging and Vision 51 (1) (2015) 1–21.

[90] L. Brun, P. Foggia, A. Saggese, M. Vento, Recognition of human actions using edit distance on aclet strings, in: Proc. International Conference on Computer Vision Theory and Applications, 2015, pp. 97–103.

[91] P. Foggia, A. Saggese, N. S. M. Vento, Exploiting the deep learning paradigm for recognizing human actions, in: Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance, 2014, pp. 93–98.

[92] Y. Kim, J. Chen, M.-C. Chang, X. Wang, E. Provost, S. Lyu, Modeling transition patterns between events for temporal human action segmentation and classification, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, 2015, pp. 1–8.

[93] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, R. Vidal, Bio-inspired dynamic 3D discriminative skeletal features for human action recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 471–478.

[94] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.

[95] E. P. Ijjina, C. K. Mohan, Human action recognition based on mocap information using convolution neural networks, in: Proc. International Conference on Machine Learning and Applications, IEEE, 2014, pp. 159–164.

[96] C. Chen, R. Jafari, N. Kehtarnavaz, Improving human action recognition using fusion of depth camera and inertial sensors, IEEE Transactions on Human-Machine Systems 45 (1) (2015) 51–61.

[97] L. Brun, G. Percannella, A. Saggese, M. Vento, HAck: A system for the recognition of human actions by kernels of visual strings, in: Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance, 2014, pp. 142–147.

[98] A. R. Lee, H. I. Suk, S. W. Lee, View-invariant 3D action recognition using spatiotemporal self-similarities from depth camera, in: Proc. International Conference on Pattern Recognition, 2014, pp. 501–505.

[99] H. Rahmani, A. Mian, Learning a non-linear knowledge transfer model for cross-view action recognition, in: Proc. the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2458–2466.

[100] Y. Song, S. Liu, J. Tang, Describing trajectory of surface patch for human action recognition on RGB and depth videos, Signal Processing Letters 22 (4) (2015) 426–429.

[101] S. Amiri, M. Pourazad, P. Nasiopoulos, V. Leung, A similarity measure for analyzing human activities using human-object interaction context, in: Proc. IEEE International Conference on Image Processing, 2014, pp. 2368–2372.

[102] S. M. Amiri, M. T. Pourazad, P. Nasiopoulos, V. C. M. Leung, Human action recognition using meta learning for RGB and depth information, in: Proc. International Conference on Computing, Networking and Communications, 2014, pp. 363–367.

[103] D. Huang, A. M. Farahmand, K. M. Kitani, J. A. Bagnell, Approximate maxent inverse optmal control and its application for mental simulation of human interactions, in: Proc. AAAI Conference on Artificial Intelligence, 2015.

[104] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, M. Hebert, Activity forecasting, in: Computer Vision - ECCV 2012, Vol. 7575 of Lecture Notes in Computer Science, 2012, pp. 201–214.

[105] B. Ni, Y. Pei, Z. Liang, L. Lin, P. Moulin, Integrating multi-stage depth-induced contextual information for human action recognition and localization, in: Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2013, pp. 1–8.

[106] F. C. Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

[107] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[108] L. Cao, Z. Liu, T. Huang, Cross-dataset action detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1998–2005.

[109] W. Sultani, I. Saleemi, Human action recognition across datasets by foreground-weighted histogram decomposition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 764–771.

[110] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, New York, 2014.