From pose to activity: Surveying datasets and introducing CONVERSE

Michael Edwards, Jingjing Deng, Xianghua Xie

Department of Computer Science
Swansea University
Faraday Tower, Singleton Park
Swansea, SA2 8PP
United Kingdom
Email: x.xie@swansea.ac.uk

Abstract

We present a review on the current state of publicly available datasets within the human action recognition community; highlighting the revival of pose based methods and recent progress of understanding person-person interaction modeling. We categorize datasets regarding several key properties for usage as a benchmark dataset; including the number of class labels, ground truths provided, and application domain they occupy. We also consider the level of abstraction of each dataset; grouping those that present actions, interactions and higher level semantic activities. The survey identifies key appearance and pose based datasets, noting a tendency for simplistic, emphasized, or scripted action classes that are often readily definable by a stable collection of sub-action gestures. There is a clear lack of datasets that provide closely related actions, those that are not implicitly identified via a series of poses and gestures, but rather a dynamic set of interactions. We therefore propose a novel dataset that represents complex conversational interactions between two individuals via 3D pose. 8 pairwise interactions describing 7 separate conversation based scenarios were collected using two Kinect depth sensors. The intention is to provide events that are constructed from numerous primitive actions, interactions and motions, over a period of time; providing a set of subtle action classes that are more representative of the real world, and a challenge to currently developed recognition methodologies. We believe this is among one of the first datasets devoted to conversational interaction classification using 3D pose features and the attributed papers show this task is indeed possible. The full dataset is made publicly available to the research community at [1].

1. Introduction

Recent advances in human motion capture and action recognition have a range of applications including surveillance, synthesis of computer generated imagery, and human-computer interfaces. Despite this progress there are still several problems that require solving, including the understanding of complex classes and maintaining accuracy rates on significantly large datasets. The field has moved fluidly between the use

of both appearance and pose based features since its conception, with datasets being produced for both modalities that can be used for cross comparison between developed methods. The release of a commercial depth sensor has revived the use of pose based features in recent years, however the datasets have yet to represent the complexity of classes that are provided by appearance based sets. We therefore intend to highlight datasets within the field and then introduce the proposed dataset to build on the current state.

The beginning of human action recognition

In 1973 and 1975, the authors in [2, 3] presented a model for the representation of the human form that closely followed the biological interpretation of human movement, the human skeleton representational model, based in Gestalt principles that provide key interest points in the movement. Model representations were then expanded by the authors in [4–9] to develop systems that are able to identify human walking actions. A review of the field was reported in [10], focusing on the recognition of the articulated movement by the human body and acknowledging the benefit of *a priori* shape models in Human Action Recognition (HAR). Campbell and Bobick [11] used 3D coordinates of 14 joints to perform event recognition from a continuous sequence of ballet moves. In following years the use of pose estimation was reduced in favor of video sequence analysis, due in part to their ease of acquisition and relatively lower cost compared to the use of marker capture systems at the time. Aggarwal and Cai [12] formed another review of the field, discussing the use of both body part representation and the global motion of the body, recognizing the need for accurate tracking of body parts when undertaking 3D estimation from 2D projections, noting the difficulty in estimating the position of joints in the scene when using appearance based pose extraction methods. The review then draws light on the use of tracking motion without needing to directly identify body parts; making use of image processing methods for appearance based tracking such as bounding box locality [13] and mesh features [14–18]. This use of motion lead to the use of appearance features in the recognition of activities, with use of image features including motion fields [19, 20], motion histories [20] and space-time interest points [21–23]. Around 2004/2005 the KTH and Weizmann action recognition datasets were publicly released to the field, providing a collection of sequences with which to evaluate developed methodologies [22, 24]. Despite their huge success as a comparison dataset, both sets were representative of the time of their release, containing single camera recordings of individual subjects performing discrete actions. Since the release of the KTH and Weizmann action sets, recent appearance based HAR has moved towards understanding complex interactions between multiple individuals. Contextual understanding of the scene as a whole has been explored in recent years, with Choi *et al.* [25] utilizing the behaviors of multiple subjects in the scene to help obtain accurate classification of a given individual's action. Further appearance datasets are reviewed in [26] with identification of sets that provide classes for specific domains and describing complex scenarios; including meta-source sets, multiview recordings, and repositories of long observations.

The use of pose

The uses of low-level and high-level featuresin HAR have been established. Low-level features typically limit recognition of actions and interactions to those of distinct or exaggerated classes which can be distinguished via strong spatio-temporal gestures or poses; such as the jumping jack, handshake and high-five. The use of higher level temporal tracking can often out-perform low-level features in HAR, and Yao *et al.* [27] suggested the consideration of 3D pose features as a benefit over lower level appearance features, acknowledging previous difficulties in obtaining accurate 3D pose features. The recent advances within pose capture and estimation methodologies has helped to reduce the difficulty in collecting 3D human pose from an observed scene, thus increasing the prominence of 3D pose in HAR. Various features have been developed from the body pose domain; including joint-joint/joint-plane distances, motion velocities, and histograms of joint orientations [27–29]. Recent work has moved into the application of fusing multiple modalities for recognition, with particular highlight on the benefit of audio-visual fusion [30–32]. With this resurgence of 3D pose it is worthwhile reviewing the datasets that are available to the HAR community in order to facilitate comparable evaluation of research methods. Discussing these datasets in terms of their reflectance of real world scenarios and ability to provide challenges to a rapidly moving field highlights the difference between the appearance and pose based areas of the community, with challenges that have been explored in the image modality being relatively untouched in the pose domain.

Human action recognition methods

Methods in classifying individual actions have been well studied in both the image processing and depth based methods. Relatively simplistic pose rich actions such as waving, walking and clapping have been the focus of research for decades, with numerous datasets providing standard benchmarks with which evaluate the performance of new methodologies. HAR has often focused on the analysis of spatio-temporal features that are extracted from data collected in the raw domain. Schuldt *et al.* [22] makes use of local space-time features to identify key interest points of motion; these points are then used to develop a vocabulary of action primitives that train a Support Vector Machine (SVM) classifier. Blank *et al.* [24] presented the action event as an XYT volume, extracting local saliency and orientation combined with global space-time features to perform spectral clustering based classification. Methods designed for action representation, segmentation and recognition via appearance information has been reported in [33]; identifying the spatial features, temporal model, temporal segmentation, and view invariance provided by each method for appearance based recognition. For pose based recognition the depth sensor has become an efficient method of tracking and extracting human skeletal model representations of subjects during experimental recordings [34–36]. This has lead to a renaissance in pose estimation techniques [27, 37–39], and to the production of numerous public datasets for pose estimation method validation [37, 40, 41]. In addition, many recognition methods have been developed which are more generic in their ability to use both appearance and skeletal model derived features; focusing on the learning of similar representative sub-action primitives, which are then verified using both appearance and skeletal features [42–45].

Human action recognition problem

Over the development of the field some main problems have revealed themselves, namely, variation in execution style and appearance. Appearance effects are reduced by considering the individual as a human skeleton model, removing all external stimuli except for pose articulation. Despite the benefit of removing anecdotal image domain information by considering pose, it is argued that this lack of appearance data may remove higher level contextual information [27]. Temporal execution variation has a large impact on the ability to recognize events, not only in execution speed but also the order in which action primitives are executed. Some actions can be subject to more variation than others; some even have a definitive order in which primitives must be executed in order to fulfill the higher level contextual semantics of the action, often described as a sequence of key poses [36, 46–48]. Execution length variance became a large part of the HAR problem, with actions being executed at differing speeds. As such, Dynamic Time Warping (DTW) has been used to align two sequences of actions, adapted from [49] in [50, 51]. This method of sequence alignment has since been used to compare sequences of differing execution length, [44, 52–54]; however the use of DTW has also been criticized, especially when aligning highly periodic actions [41] or actions where the time taken to execute of a key feature, such as walking and running [48]. Exemplar based methods make use of key poses almost as a series of checkpoints frames which make up an action, and therefore are believed to not require a time warping alignment phase [55, 56]. These developed methodologies seem to provide reliable accuracy for the publicly available datasets on which they are often validated, despite their variance in execution rates and styles.

Another issue in the community is the lack of methods which are able to extend beyond the recognition of simplistic action classes. Weinland *et al.* [33] reports upon the predictive accuracy of methods that are evaluated on the KTH, Weizmann and IX-MAS datasets; showing that in recent years the level of accuracy can often reach over 90%. State of the art performance accuracy is also reported within [57], with older datasets often reporting the highest number of correct classifications. Hassner [57] also shows that those datasets which are more representative of real world observations tend to challenge the current methods within the community; such as Hollywood1/2, HMDB51 and Olympic Sports. This suggests that current HAR methods are able to easily classify the relatively simplistic classes presented in established datasets, but that the community requires challenging with complex scenarios.

Contributions

This study aims to first consider a large selection of the current datasets that are available for human action recognition, evaluating properties that facilitate comparable testing of developed HAR methods. The survey identifies the growth of the field from consideration of generic emphasized actions towards the understanding of interactions between numerous individuals. Datasets are analyzed based on a variety of key properties that influences their use for various HAR techniques, including number of action classes, complexity of events and their application domain. Differing levels of abstraction within the understanding of human behavior are described, detailing the nature between pose, gesture, action, interaction, and activity. Despite the progression

of the field towards higher levels of behavior abstraction there is still need for a dataset that provides interaction classes that contain complex person to person activities, representing actions and interactions that are not readily identifiable by the presence of a given gesture or pose. The second part of this study then aims to help occupy this gap in the community with a dataset describing subtle conversational interaction classes. CONVERSE provides a collection of interactions in which the activity develops over a long period of time, with realistic representations of behaviors that have contextual differences that are difficult to define by motion.

The rest of this paper will outline the current state of data available to the community and outline the requirement for a novel conversational set. In section 2 we present the evaluation of a plethora of available datasets, evaluating each one based on the provided set of criteria and highlighting the need for a dataset which introduces subtle interactions between individuals. Section 3 then describes the surveyed sets, providing key information regarding their composition and usage. Section 4 then draws on these findings to present our novel dataset, describing the composition of the data and its usage in HAR, and providing a baseline set of classification results for comparison.

2. Discussion of current state

Numerous HAR datasets have been produced and publicly released in the last decade for the purpose of detecting and identifying action events in an observed scene. Many of these sets have the added benefit of allowing cross-verification of methodologies developed in the field of computer vision; specifically those of action detection and classification. Available datasets contain a variety of traits which require consideration when deciding upon their appropriate usage. Sets differ in the data collection modality; including RGB videos, depth maps, accelerometers and marker based motion capture. They also differ in the actions carried out; including simple gestures, discrete actions, and continuous sequences of actions, multi-user interactions and person-object interactions. Some datasets make use of original data collection, allowing a degree of control over certain parameters within the data collection methodologies. Others use meta-data collected from video clips that are publicly available from media such as films and online video clips; these tend to have large amounts of variation between individual sequences, however they are also among the largest of the datasets, with some meta-sets containing thousands of sequences [87, 131]. Numerous sets have ground truth labels for an entire sequence; however many are either manually segmented out of a continuous sequence of multiple actions, or are left for users to perform labeling before their use. Ground truth labeling on a frame-by-frame basis is rare, due to the complexity in determining the exact frame at which an action begins.

Datasets, such as KTH, Weizmann and MSR Action3D [22, 24, 40], provide the common examples of well annotated and discrete action executions; including kicking, walking, and shaking. Others, such as the CMU Motion Capture set [74], expand the complexity further by containing sequences of multiple actions executed in a continuous manner. Recently, sets have moved towards recognizing interaction between two people, including SBU Kinect, BIT-Interaction and K3HI [37, 64, 101]; however, these sets still provide interactions using the classic simplistic actions of pushing, punching

Table 1: Comparisons of key action recognition datasets, detailing the download location, associated descriptive publications, and number of simultaneous viewpoints.

Name	Modality	URL	Description	Views
50 Salads	RGB-D, IMU	[58]	[59]	1
BEHAVE	RGB	[60]	[61]	2
Berkeley MHAD	RGB-D, IMU, Audio, MoCap	[62]	[30]	14
BIT Interaction	RGB	[63]	[64]	1
CAD120	RGB-D	[65]	[66]	2
CAD60	RGB-D	[65]	[67]	2
CASIA	RGB	[68]	[69]	3
CAVIAR	RGB	[70]	[71]	1, 2
CMU MMAC	RGB, MoCap, IMU	[72]	[73]	6
CMU MoCap	MoCap	[74]	-	1
CONVERSE	RGB-D	[1]	[75–77]	1
Drinking/Smoking	RGB	[78]	[79]	1
ETISEO	RGB	[80]	[81]	1, 3, 4
G3D	RGB-D	[82]	[83]	1
G3Di	RGB-D	[84]	[85]	1
HMDB51	RGB	[86]	[87]	1
Hollywood	RGB	[88]	[89]	1
Hollywood-2	RGB	[90]	[91]	1
Hollywood3D	RGB-D	[92]	[93]	1
HumanEVA-I	RGB, MoCap	[94]	[95]	7
HumanEVA-II	RGB, MoCap	[94]	[95]	4
IXMAS	RGB, Silhouette	[96]	[97]	5
JPL	RGB	[98]	[99]	1
K3HI	RGB-D	[100]	[101]	1
KTH	RGB	[102]	[22]	1
LIRIS	RGB-D	[103]	[104]	1
MPI08	RGB, IMU, Laser Scan	[105]	[106, 107]	8
MPII Cooking	RGB	[108]	[109]	1
MPII Composite	RGB	[110]	[111]	1
MSR Action-I	RGB	[112]	[113]	1
MSR Action-II	RGB	[112]	[114]	1
MSR Action3D	RGB-D	[112]	[40]	1
MSR DA3D	RGB-D	[112]	[41]	1
MSR Gesture3D	RGB-D	[112]	[115]	1
MuHAVi	RGB, Silhouette	[116]	[117]	8
Olympic Sports	RGB	[118]	[119]	1
POETICON	RGB, MoCap	[120]	[121]	7
Rochester AoDL	RGB	[122]	[123]	1
SBU Kinect Interaction	RGB-D	[124]	[125]	1
Stanford 40 Actions	Image	[126]	[127]	1
TUM Kitchen	RGB, Markerless MoCap, RFID	[128]	[129]	4
UCF101	RGB	[130]	[131]	1
UCF11	RGB	[132]	[133]	1
UCF50	RGB	[134]	[135]	1
UCF Sport	RGB	[136]	[137]	1
UMPM	RGB, MoCap	[138]	[139]	1
UT Interaction	RGB	[140]	[141]	1
ViHASi	RGB, Silhouette	[142]	[143]	40
VIRAT	RGB	[144]	[145]	-
Weizmann	RGB, Silhouette	[146]	[24, 147]	1
WVU MultiView	RGB	[148]	[149, 150]	8

and kicking. A few studies, including MSR DailyActivity3D and the TUM Kitchen [41, 129], have made steps towards the recognition of so-called 'daily activities', natural actions which may be more representative of the real world executions.

Despite this abundance of datasets, there is still a lack of sets that make use of subtle interaction classes, representing loosely defined actions such as those in natural

Table 2: Comparison of provided data and presence of dedicated validation sets.

	Datasets	
Data		
RGB/Greyscale	All sets except CMU MoCap, K3HI,	
MoCap	Berkeley MHAD, CMU MMAC, CMU MoCap, HumanEVA-I, HumanEVA-II, POETICON, TUM Kitchen, UMPM	
Depth	50 Salads, Berkeley MHAD, CAD120, CAD60, G3D, G3Di, Hollywood3D, LIRIS, MSR Action3D, MSR DA3D, MSR Gesture3D, SBU Kinect Interaction, **CONVERSE**	
Skeleton	Berkeley MHAD, CAD120, CAD60, G3D, G3Di, K3HI, MSR Action3D, MSR DA3D, SBU Kinect Interaction, **CONVERSE**	
IMU	50 Salads, Berkeley MHAD, CMU MMAC, MPI08, TUM Kitchen	
Audio	Berkeley MHAD, POETICON	
Laser Scan	MP108	
	Appearance sets	Pose sets
Train/Test split		
Yes	Drinking/Smoking, ETISEO, Hollywood, Hollywood 2, IXMAS[1], KTH, Olympic Sports, Rochester AoDL[1], Stanford 40 Actions, UCF101, UCF11[1], UCF50[1], UCF Sport[1], UT Interaction, ViHASi[1], VIRAT[1], Weizmann[1], WVU MultiView-I, WVU MultiView-II	Hollywood3D, HumanEVA-I, HumanEVA-II, LIRIS, MSR Action3D, SBU Kinect Interaction, TUM Kitchen[1], **CONVERSE**[1]
No	BEHAVE, BIT-Interaction, CASIA, CAVIAR, HMDB51, JPL, MPII Cooking, MPII Composite, MSR Action-I, MSR Action-I, MuHAVi	50 Salads, Berkeley MHAD, CAD120, CAD60, CMU MMAC, CMU MoCap, G3D, G3Di, K3HI, MPI08, MSR DA3D, MSR Gesture3D, POETICON, UMPM

[1] provided in description paper via Leave Out cross validation methodology

conversational styles, or in context dependent situations. With [76, 77], we have presented methodology, using a dataset of subtle conversational interactions, which is able to classify such subtle action events, based upon 3D pose features.

The following section will evaluate the public datasets detailed within section 3 and summarized in Table 1, identifying key features for their usage in the HAR community. Several parameters that require consideration when developing and evaluating action recognition methodologies using publicly available data are identified; including the modality of data acquisition, data provided by the set, and consistent training and testing subsets. The complexity of each dataset is also evaluated, based upon the number of individual classes they present, the number of samples provided, and the presence of complex and realistic class scenarios. Summaries are provided in Tables 1, 2, 4, 5, 6, 7, 8, 9. The proposed CONVERSE dataset [1] is included within the evaluations to highlight the necessity for such a set and identify where it resides amongst the currently available data. A detailed explanation of the proposed dataset is given in section 4.

2.1. Modality

In Table 2 we cluster the datasets based on their method of data capture; from video, depth maps, skeletal tracking, Motion Capture (MoCap) marker tracking, IMU, and audio. The majority of sets in HAR make use of vision; however recent progress has been made towards the use of 3D pose estimation via depth sensors; therefore

understanding the modality provided by a dataset will often impact on the choice of features used to describe each sequence.

Video

Appearance based HAR makes use of datasets that are often collected via still images or video, as cameras can provide a relatively cost effective method of obtaining both real-world and staged execution samples from both a laboratory or real-world environment. In Table 1 it can be seen that all of the datasets presented contain some form of video or appearance based data (except CMU MoCap, K3HI and UCF iPhone); therefore in Table 2 we omit the video data. The quality of the recordings varies greatly between sets, with some specializing in evaluating action detection and recognition in low quality or small scale recordings. High intra-set and inter-sequence variation in image quality, camera motion, scale and viewpoint are common in meta-data sets that collect observations from multiple sources, such as UCF101, UCF50, UCF11, Hollywood, Hollywood-2 and HMDB51, and these pose a more realistic problem to the community. Visual based HAR can provide an intuitive representation of the scene, however there can often be superfluous information contained within an observation that negatively impacts on the reliable global recognition of a given action; therefore, appearance based modalities can often make use of subject localization and background removal, coupled with the extraction of descriptors such as Space-Time Interest Point (STIP)s, Histogram of Oriented Gradients (HOG), Histograms of Optical Flow (HOF) or local regions of motion features to enable the global recognition of actions regardless of background information or subject-specific appearance. Many depth based datasets also provide simultaneously captured video representations of their data; this appearance data can either be omitted from the learning, or combined to form a multi-modal system. Of the appearance based datasets, the KTH and Weizmann datasets have been cited the most for single action recognition method evaluation. For appearance based interaction recognition the CAVIAR, Hollywood and UT Interaction datasets have been used frequently by the community.

MoCap

Motion capture concerns the recording of numerous markers placed upon the body by multi-camera systems, providing accurate tracking of the markers within a volume over time. MoCap often provides a method of capturing a spatial ground truth for the marker locations within the scene, being used as a stand-alone modality or augmenting datasets captured through other methods. MoCap systems are often calibrated using built in software and a calibration tool, allowing all cameras to be spatially and temporally synchronized, increasing confidence in the marker tracking. Placement of the markers varies between datasets and as such datasets which make use of MoCap provide details of the marker placement on the body, allowing semantic affordance to be applied to each marker. MoCap can be seen as a cost-expensive method of data collection, often requiring dedicated systems; however the generation of a spacial ground truth and reliable pose tracking method is of great benefit when developing pose from appearance or pose based action recognition methodologies. Despite this, an implementation of marker based MoCap systems in a real world environment is impractical, requiring individuals to wear a motion capture suit to be detected by the system would

provide little benefit to the user; as such there has been some effort that has also been made to produce human skeletal tracking without the use of markers from simple RGB image recording [129] and from depth maps [40].

Of the HAR datasets that utilize MoCap, the HumanEVA, Berkeley Multimodal Human Action Database (MHAD) and Carnegie Mellon University (CMU) MoCap datasets are most commonly used. The HumanEVA dataset provides a set of evaluation metrics for the purpose of action recognition, Berkeley MHAD provides a detailed dataset containing multiple modalities for fusion based action recognition, and the CMU MoCap dataset contains a vast number of continuous sequences which can be used for action detection and sequence segmentation.

Depth

The production of a consumer level depth sensor, most notably the Microsoft Kinect, coupled with efficient and accurate joint tracking software has provided the HAR community with an inexpensive method of collecting 3D poses of a subject performing actions within a scene [34–36]. This has allowed for the development of methods that represent the action as a series of key poses or bag of words model [46, 47, 77], extracting the key frames that describe the overall action event. Datasets such as 50 Salads, Berkeley MHAD, CAD120, CAD60, G3D, G3Di, K3HI, LIRIS, MSR Action3D, MSR DA3D, MSR Gesture3D, and SBU Kinect Interaction all make use of the Kinect depth sensor to collect data providing the depth map of the scene. The Hollywood3D set utilizes commercial films that have been recorded using a 3D stereo camera system to provide depth maps. By obtaining a 3D pose estimation of the subjects within the scene users are able to, given accurate tracking, generate pose, scale, and appearance invariant features for the purpose of HAR that include joint trajectories, joint-joint distances, joint-plane distances, and joint motion histories. Many of the depth datasets captured using the Kinect provide the associated estimated skeleton representation of the individual, tracking a number of joints across the scene. The number of joints tracked and the position of the provided markers often depends on the method used to extract the skeleton; those using the Microsoft Kinect SDK often provide 20 points, whilst those using the OpenNI standard track 15 joints on the body. The selection of joints often aligns with the major joints of the human body, and so provides an estimation of limb motion. Currently the use of depth sensors are limited to a viewpoint that is in a roughly front-on position due to the method of estimating depth, using distortions of infra-red projections into the scene which is then captured by a receiving sensor. This method has little ability to handle scene occlusions which can cause shadowed regions in the depth map, resulting in lost or noisy tracking in the extracted skeletons.

The most prominent depth datasets for single person actions include those presented by the Microsoft Research group, namely the Action3D and DA3D datasets. Despite the small number of samples and action classes provided by the MSR Action3D dataset there has been a vast number of citations for its use as an evaluation dataset. For person-person interactions there are few datasets available which make use of depth based data; the K3HI and SBU Kinect Interaction datasets provide sequences of single executions of a given interaction, analogous to those provided by the BIT Interaction and UT Interaction appearance datasets, however their recent release may reflect their low citation and usage for evaluation of pose based methods.

Other

Various other methods of data capture have been used for HAR purposes, including the use of audio recordings [30, 151] and IMUs [30, 152, 153]. These methods can provide reasonable classification results on their own, however they are often used in a multi-modality system to improve the accuracy rates of single modality methods. These datasets are beyond the scope of this survey and omitted for brevity.

2.2. Action class types

Human behaviors are often a set of events with differing levels of abstraction and complexity, therefore to aid comparison between HAR class types we shall first define assumptions made about terminology we wish to use. Many class labels provided within HAR datasets can often be re-labeled to fit within a different level of abstraction, however we attempt to use common terminology found across the community, with an overview provided in Figure 1 and a summary of the datasets in Table 4. Example images from datasets that describe differing levels of abstraction are given in Table 3.

Pose An atomic observation of the spatial arrangement of a human body at a single temporal instance, e.g. 'Arm above head'.

Gesture A temporal series of poses on a sub-action scale, sometimes described as action primitives e.g. 'Arm moves left'.

Action A series of gestures which form a contextual event, e.g. Repeated gestures of arm moving left and then right can be contextual described as an 'overhead wave action'. These are the most commonly used class labels found within current datasets, describing single actions executed by a subject including 'run','jump', and 'wave'.

Interaction A pairwise or reciprocal action is committed by two entities on each other. Each entity therefore has a single action that reflects it's state compared to the other entity, i.e. consider the action of person A shaking the hand of person B; A executes the action of shaking the hand of B, B executes the action of having their hand shaken by A, together this pairwise action execution can be described as that of a 'handshake' interaction. For the purpose of action recognition interactions are often further divided into differing interaction types based on if the entities include people, objects or groups. For this study we have omitted group interaction datasets due to space limitations.

- ***Person-Person*** An action is committed directly by one individual upon another. This definition does not include crowded scenes in which an individual performs a single person action with other subjects in the environment. The class labels in a P-P interaction treats the interaction as a single entity, rather than two separate single person actions, e.g. we consider the class 'punching' as an interaction between person A, the puncher, and person B, the individual being punched.

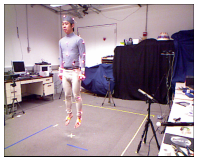
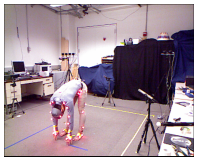
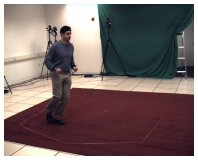
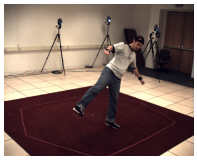
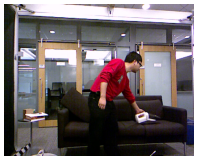
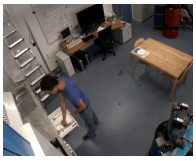
Action Type	Dataset	Example frames		
Action	Berkeley MHAD			
Action	HumanEva			
P-P Interaction	SBU Kinect Interaction			
P-O Interaction	50 Salads			
Activity	MSR DA3D			
Activity	TUM Kitchen			

Table 3: Example frames of currently available depth based human action recognition datasets. Images are provided here to give insight into the types of classes provided by pose based data.

Person-Object An action is committed directly by one individual upon an object. This includes the manipulation of objects. We consider class labels such as 'lift chair' and 'open box' as person-object interactions as the actions 'lift' and 'open' are performed on the objects 'chair' and 'box' respectively.

Groups Characterized as interactions carried out between a collected entity of more than two individuals. Group interactions can include inter- and intra-group behaviors and the interaction of the group on other objects, individuals, or even other groups. These often form their own subsets of group behaviors.

Activity A collection of actions and/or interactions that compound to describe a high level event. These are common within the sets that describe daily behaviors, e.g. 'cook a meal' and 'tidy room' can often include numerous actions and interactions that are executed. Each action and interaction can therefore be thought of a sub-activity event in such scenarios. Activity is also used to describe the daily activities, a more realistic observation execution than the exaggerated instances such as 'punch' and 'kick'.

A common scenario presented within HAR instances is that of a single person executing a singular action, in which an individual actor performs an action with no interaction to other individuals or objects, such as within KTH, Weizmann, MSR Action, and MSR Action3D. In recent years, interaction datasets have become more prominent, often displaying actions where one actor performs an action upon which another actor is the recipient. These interaction sets can still exhibit behaviors that are quite well defined, with a single instigator and a single recipient, such as punching, pushing and move towards. The most notable interaction sets include BIT Interaction, UT Interaction, K3HI, and SBU Kinect Interact datasets. There also exists interaction classes that are more complex in their composition, involving multiple entities, object manipulation or requiring higher level semantics; these are prominent in the TUM, BEHAVE, VIRAT, ETISEO, and POETICON datasets. The higher level activity datasets often provide observations of an entire task being carried out and require the understanding of the sub-activity actions and interactions being carried out over the course of the recording. In the current sets there are often annotations of lower level actions which are encompassed within a higher level activity context, with sets such as MPII Composite, 50 Salads and TUM Kitchen providing annotations of both levels of abstraction and the objects that are subject to interactions during the course of the activity.

The choice of classes that are performed by the actors is a key motivation in the generation and usage of the proposed dataset. Often the actions executed are those of a visually definable nature, comprising single executions of a discrete action which contain key poses and gestures. The complexity of the problem can then be increased by observing multiple executions of actions in a sequence, either with distinct boundaries between the classes or with a natural flow between different classes. These are all complex issues that are the focus of the community, with segmentation methods often utilized to separate out actions from a continuous sequence. Judging the difference in complexity between two classes can be subjective, depending upon the subtlety of

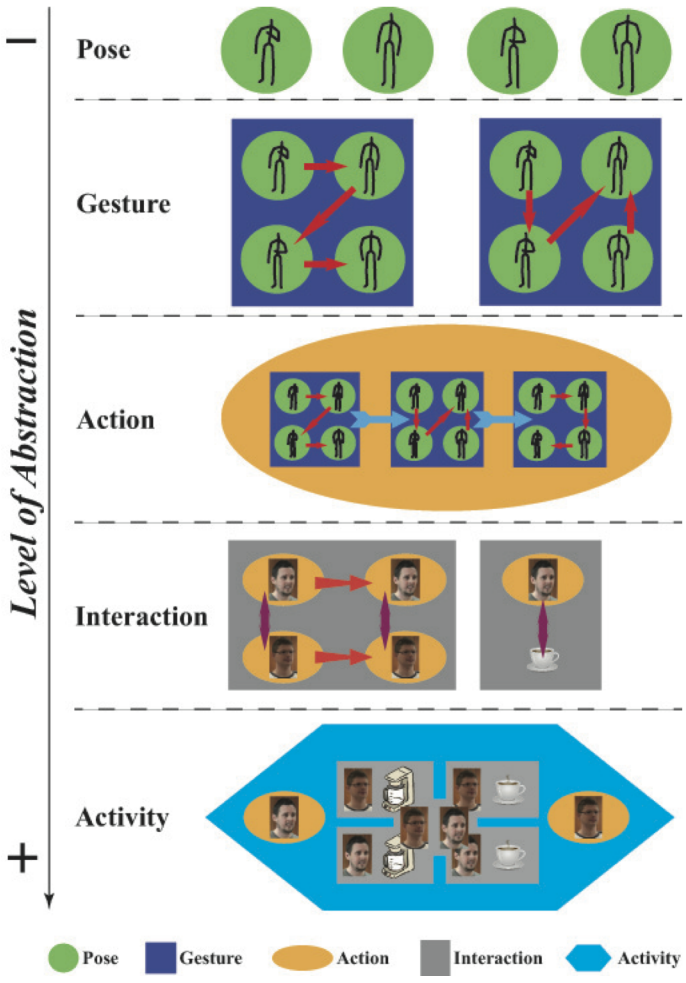


Figure 1: Levels of abstraction within human action recognition.

gestures, the context of any interactions, and the spatio-temporal rigidity of the executions; subtle gestures, for example, may well present a more complex recognition problem than the simplest of activity classes. We can however make some generalized assumptions about the complexity within the different abstraction levels. Lower levels of abstraction such as pose and gesture should provide less challenges to the field in its current state, while higher levels of abstraction, especially those involving interactions between two or more entities, still remain a challenging issue.

Obviously with the definitions of the action types presented there can be some overlap in how to handle events in which an entity is not only interacted with, but also pivotal to the context of the label. Consider the class label 'smoking', this event can fit both into the definition of a singular action in which the object is explicit to the action, a person-object interaction between the person and cigarette, and also into its own activity class in which smoking is the task executed. Consider also the class label of 'pushing', this may be a class label that can be readily classified as a single action, person-person interaction, or person-object interaction depending upon the entities present, and also as an activity if there is a contextual background to the event. This highlights the complexity in describing class labels and requires the careful consideration of overlaps that appear to be presented between datasets with similar action classes. To further this point, we ask should the community consider an interaction as its own complete class, or should the system understand the states occupied by all entities within the interaction, i.e. the class label of 'pushing' may be deconstructed into sub-classes that describe the action of the instigator and the reaction of the recipient. Many interaction datasets handle the class labeling as a single complete unit of interaction, often reliant on the action committed by the instigator, e.g. K3HI, SBU Kinect Interaction, and UT Interaction. However the TUM Kitchen, 50 Salads and MPII Composite sets explicitly annotate the states of both entities to define the person-object interactions for the purpose of activity recognition. The use of a single interaction class that encompasses all sub-divisions of that interaction may provide learning that is broad and resistant to variation of intra-class behaviors; however by learning the sub-divisions of an interaction class, considering the different actions and reactions as their own states, there may be an ability to learn more effective boundaries for execution variations. For this study we have considered and evaluated upon the class labels provided by the original datasets; however we invite the community towards potentially defining multi-scale class labeling for the purpose of action and activity recognition.

2.3. Size

The size of a dataset, not just in the number of sequences but also in the range of different action classes and participants, can impact on it's suitability for method evaluation. Testing on a small-scale dataset can provide misleading results during analysis which may not be replicated when introducing more class labels or observations, due in part to the highly variable nature of inter- and intra-instance executions. Contrarily there are implications in the usage of large datasets; not only the collection and storage of data, but also in the processing of features, class learning and validation. Due to the inherent issues in obtaining a large number of participants, action classes, and sequences, the largest sets tend to be meta-sets, which collect action sequences from various sources, such as YouTube and films, containing large variation between

Table 4: Comparison of dataset interaction types. Note that datasets can contain instances of several types of behaviors based on the labeling it provides.

	Appearance sets	Pose sets
Event type		
Action	CASIA, CAVIAR, Drinking/Smoking, ETISEO, HMDB51, Hollywood, Hollywood-2, IXMAS, KTH, MSR Action-I, MSR Action-II, MuHAVi, UCF11, UCF Sports, ViHASi, VIRAT, Weizmann, WVU MultiView-I, WVU MultiView-II	50 Salads, Berkeley MHAD, CAD120, CAD60, CMU MoCap, G3D, Hollywood3D, HumanEVA-I, HumanEVA-II, LIRIS, MPI08, MSR Action3D, MSR Gesture3D, POETICON, TUM Kitchen, UMPM
Interaction: Person - Person	BEHAVE, BIT Interaction, CASIA, CAVIAR, ETISEO, Hollywood, Hollywood-2, JPL, UT Interaction	CMU MoCap, G3Di, Hollywood3D, K3HI, LIRIS, POETICON, SBU Kinect Interaction, UMPM, **CONVERSE**
Interaction: Person - Object	ETISEO, MPII Cooking, MPII Composite, VIRAT	50 Salads, CAD120, CMU MMAC, LIRIS, POETICON, TUM Kitchen, UMPM
Activity	CASIA, MPII Composite, MuHAVi, Olympic Sports, Rochester AoDL, Stanford 40 Actions, UCF101, UCF11, UCF50, UCF Sports, ViHASi	50 Salads, MSR DA3D, CAD60, LIRIS, TUM Kitchen, **CONVERSE**

sequences; this often makes meta-sets highly variable and challenging problems to be solved. A summary of dataset sizes is given in Table 5

Number of classes

Datasets with a small number of action classes, such as MSR Action-I, MSR Action-II, and Drinking/Smoking, can often provide strong recognition results in part due to the low number of partitions needed to divide the actions provided within the set. Those sets that contain a large number of action classes, namely HMDB51, UCF101, and UCF50, provide a difficult challenge to HAR methods due to the need to find partitioning information within each class that allows for inter-class partitioning, while preserving intra-class similarity. Due to the inconceivable number of possible actions and interactions that can exist in the real world it can be beneficial to evaluate methodologies on datasets with a large number of distinct action classes.

Number of subjects

Datasets that are able to provide more individual subjects performing an action are able to portray the variability in inter- and intra-subject execution of a given class. Observations of the same action class can often differ greatly in both their temporal rate and spatial occupancy, leading to complexity in learning the action for recognition purposes. Methods that are able to provide subject invariant action recognition should provide consistent results on a dataset which contains a large number of subjects. Again, the meta-sets tend to provide the highest number of subjects, almost capturing a new subject per sequence, representing a large range of inter-subject variation.

Number of samples per class

The number of observations per class can impact on the ability of a system to suitable learn a given class. A low number of observed instances of a class can result in weak recognition of unobserved instances of the same class. HMDB51 provides over 100 instances of each action class it contains, providing a range of observations across differing viewpoints, quality and executions, as such it can provide a useful benchmark for the recognition of actions from a subject and observation invariant methodology. Current pose based datasets contain few repeated instances of an action class, often with 3-5 repetitions per subject per class. To increase the number of instances per class it is possible to segment those datasets which contain continuous recordings of multiple executions into discrete single execution clips, this includes the KTH dataset.

Number of sequences

The total number of sequences within a dataset should be a factor of the number of subjects, classes, and number of class executions, and as such can impact on the reliability of the results produced. Larger datasets can provide larger testing sets for which to evaluate a system, allowing for more confidence in the results of the validation. Size alone however is only one parameter in the selection of evaluation benchmark, with domain, class complexity and modality impacting on the application of methodologies to real world implementations.

2.4. Application domain

The intended application domain of a dataset can provide certain intrinsic features in the data collection methodology and action classes captured, from low resolution images of CCTV surveillance footage to more complex action sequences of daily living. Some actions are representative of the domain from which they are intended; for example the UCF-Sports dataset, [137], makes use of numerous actions from various sports, such as javelin throws and long jumps. We classify the datasets into 4 action class domains; generic actions, daily living, surveillance, and sport. Generic action datasets have no overall theme, instead providing classes that are pan-domain; these include the classes 'running', 'jumping', 'punching', and also more complex interactions such as 'handshake' or 'play guitar'. Daily living datasets often include actions and activities that are more natural in their execution and environment, this includes classes based on assisted living and household tasks. Surveillance datasets often make use of elevated view points and lower resolution images, mirroring the common camera setups in the security industry [145, 154]. Sports based action recognition often makes use of previously captured data from multiple sources, often containing varying image quality and varying levels of camera motion. A summary of the domains for each of the datasets is provided in Table 6.

Generic

Many action recognition datasets often contain generic action classes that are observable in numerous domains. The intention is to cover a wide variety of actions to allow domain invariant action recognition, with generic datasets being the most widely used for validation purposes, including the KTH [22], Weizmann [24] and MSR Action3D [41] sets. Many generic datasets are collected in a laboratory environment; with

Table 5: Comparison of dataset sizes.

	Appearance sets	Pose sets
# Actions		
≤ 5	Drinking/Smoking, MSR Action-I, MSR Action-II	
6 - 10	BEHAVE, BIT Interaction, CAVIAR, Hollywood, Hollywood-2, JPL, KTH, Rochester AoDL, UCF Sport, UT Interaction, Weizmann, WVU MultiView-II	CMU MMAC, HumanEva-I, HumanEva-II, K3HI, LIRIS, MPI08, POETICON, SBU Kinect Interaction, UMPM, **CONVERSE**
11 - 15	CASIA, ETISEO, IXMAS, UCF11, VIRAT, WVU MultiView-I	Berkeley MHAD, CAD60, G3Di, Hollywood3D, MSR Gesture3D, TUM Kitchen
16 - 20	MuHAVi, Olympic Sports, ViHASi	50 Salads, CAD120, G3D, MSR Action3D, MSR DA3D
≥ 21	HMDB51, MPII Cooking, MPII Composite, Stanford 40 Actions, UCF101, UCF50	CMU MoCap
# Subjects		
≤ 5	Rochester AoDL	CAD120, CAD60, HumanEVA-I, HumanEVA-II, MPI08, POETICON, TUM Kitchen
6 - 10	MSR Action-I, MSR Action-II, UT Interaction, ViHASi, Weizmann	G3D, MSR Action3D, MSR DA3D, MSR Gesture3D, SBU Kinect Interaction
11 - 20	IXMAS, MPII Cooking, MuHAVi	Berkeley MHAD, G3Di, K3HI, **CONVERSE**
≥ 21	CASIA, KTH, MPII Composite	50 Salads, CMU MMAC, CMU MoCap, UMPM
Undefined	BEHAVE, BIT Interaction, CAVIAR, Drinking/Smoking, ETISEO, HMDB51, Hollywood, Hollywood-2, JPL, Olympic Sports, Stanford 40 Actions, UCF101, UCF11, UCF50, UCF Sport, VIRAT, WVU MultiView-I, WVU MultiView-II	Hollywood3D, LIRIS
# Sequences		
≤ 20	BEHAVE, CAVIAR, MSR Action-I,UT Interaction, WVU MultiView-II	HumanEVA-II, TUM Kitchen, **CONVERSE**
21 - 100	ETISEO, JPL, MPII Cooking, MSR Action-II, Weizmann	50 Salads, CAD60, CMU MMAC, G3Di, HumanEVA-I, MPI08, POETICON, UMPM
101 - 500	BIT Interaction, Drinking/Smoking, Hollywood, MPII Composite, Rochester AoDL, UCF Sport, ViHASi	CAD120, G3D, K3HI, MSR DA3D, MSR Gesture3D, SBU Kinect Interaction
501 - 1000	KTH, Olympic Sports, WVU MultiView-I	Berkeley MHAD, Hollywood3D, LIRIS, MSR Action3D
≥ 1001	CASIA, Hollywood2, HMDB51, IXMAS, MuHAVi, Stanford 40 Actions, UCF101, UCF11, UCF50, VIRAT	CMU MoCap

static cameras, static backgrounds and calibrated data-capture setups, including Berkeley MHAD and CMU MoCap. Others may be collected outdoors with a controlled clutter free setting, such as Weizmann and KTH. Others are collected within cluttered environments, featuring non-participatory subjects that complicate the scene, such as MSR Action-I and Action-II. Pose based datasets which make use of a depth sensor and the pose estimation technique of extracting the 3D skeleton are often captured in a relatively clutter free scene due to the limitations of the skeletal tracking methodology used.

Daily living

Daily living sets are designed to closely represent the natural world in both the environmental surroundings and the natural style of action classes executed. The Tum Kitchen [129], MSR DA3D [40, 41], MPII Cooking [155], and Rochester AoDL [156] sets are commonly used for the analysis of methodology in the recognition of day-to-day activities. Activities include 'having a conversation', 'phone calls', 'laying down', 'drinking' and 'eating', but may also include sub-actions within a higher level task, such as 'setting a table' or 'cooking a meal'. The executions may be allowed to occur naturally as in the 50 Salads, MPII Cooking, and MPII Composite datasets; or the observations may be more scripted, such as in the POETICON and the robotic class of the TUM Kitchen set [59, 121, 129]. By understanding the actions and interactions within a daily activity dataset the field is moving towards learning higher level semantics of human behavior via natural representations.

Surveillance

Surveillance is a domain concerned with detecting and identifying activity within a continuous observation of a scene, often making use of video-based action recognition samples that are taken from a distance, prone to crowding, and contain poor resolution recordings. A surveillance domain sequence may contain more frames of empty or redundant information, sporadically interspersed with temporally short regions of interest. Datasets such as UT-Interaction, CASIA, and BEHAVE make use of surveillance style setups to capture emphasized person-person interaction classes such as 'come together' and 'fight'. The CAVIAR, ETISEO, and VIRAT datasets all make use of detailed ground truth annotations to provide information regarding persons and objects within the scene, enabling the evaluation of methods in detecting varies entities and their interactions within a scene for higher semantic understanding of the events.

Sport

The UCF-Sports, [137], and Olympic Sports, [119], datasets are focused explicitly on sports related action examples. These sets contain samples that are collected from various sources of TV and online recordings, providing samples that vary in their recording quality and containing both static and dynamic camera movements. As such these can often be challenging datasets. In both cases the intent of the dataset is to be able to recognize the sport being performed, this can be more challenging than in the case of learning sports related actions, such as in the case of 'tennis serve' and 'boxing' from some of the generic action datasets. A sport as a high level class can contain numerous action and interaction actions that make up the overall activity and learning a sporting class may require learning vastly different observations that belong to the same class. 3D pose based HAR in the sports domain has few datasets due to the complexity in capturing a large volume in which the activity can be played. The G3Di dataset provides interactions between two people in the context of a sporting game played through a console, however we treat the provided classes as being generic actions rather than true sporting based actions.

Table 6: Comparison of dataset domain applications.

	Appearance sets	Pose sets
Domain		
Generic	BIT Interaction, HMDB51, Hollywood, Hollywood-2, IXMAS, JPL, KTH, MSR Action-I, MSR Action-II, MuHAVi, Stanford 40 Actions, UCF101, UCF50, UCF11, ViHASi, Weizmann, WVU MultiView	Berkeley MHAD, CMU MoCap, G3D, G3Di, Hollywood3D, HumanEVA, K3HI, MPI08, MSR Action3D, MSR Gesture3D, SBU Kinect Interaction, UMPM
Daily Living	Drinking/Smoking, MPII Cooking, MPII Composite, Rochester AoDL	50 Salads, CAD120, CAD60, CMU MMAC, LIRIS, MSR DA3D, POETICON, TUM Kitchen, **CONVERSE**
Surveillance	BEHAVE, CASIA, CAVIAR, ETISEO, UT-Interaction, VIRAT	
Sport	Olympic Sports, UCF Sports	

2.5. Ground truth

Table 7 outlines various ground truths provided with each dataset, both for spatial ground truths and labeling of action classes. Providing consistent ground truth with which to evaluate results is important for developing benchmarks against which to test developed methodologies, aiding in the generation of a metric score that can be used to compare implementations.

Class label ground truths and scene annotations of a dataset can provide a clear benchmark for quantifying the performance of a developed methodology. Some datasets provide frame-by-frame labeling of the scene, whilst others label an entire sequence as containing a given class label. These annotations allow quantification of results obtained from various methodologies, with predicted class labels and detections being compared against the ground truth. The collection of the class ground truth can be either manually annotated by the author or produced via some form of machine learning. Manual annotation can provide detailed descriptions of the entire scene, with locations and affordances being given to persons and objects within the scene, as can be seen with the ETISEO and HMDB51 datasets. These can be extremely useful when tracking the states of multiple entities within the scene, or for the understanding of a high level abstracted class; however the manual labeling of individual frames can produce observation bias into the dataset, requiring strict objective criterion to gain consistent ground truths. Machine based annotations can combined machine learning with data labeling to rapidly provide ground truths to large datasets, e.g. the Hollywood and Hollywood-2 datasets are partially annotated by learning textual descriptions within the film's scripts. An automated ground truth annotation may require subsequent manual verification to ensure the false labeling is minimized. The simplest form of ground truth labeling provided by HAR datasets is by attributing the entire sequence to a specific label, acknowledging that a given action occurs at some point within the observation, as is the case with CASIA, CMU MMAC, MSR Action3D, and many more. Having simplistic whole sequence labeling can make it hard to use such datasets for detection purposes, as evaluating the beginning and end frames of an action can be problematic to determine manually. For action recognition purposes the learning of background frames

from a sequence may also provide some level of noise to the partitioning of that class.

Spatial truth can be provided by explicitly locating the subjects and objects within the environment or by highlighting regions of interest in which the the subject, object or event resides by using bounding boxes or silhouette masks. Calibrated ground truth methods can be used to determine the spatial locations of the subjects within a scene, often using motion capture suits and markers to explicitly track the body through a capture volume, providing either a raw point cloud or the predicted skeletal frame of the body. The accuracy of motion capture systems can vary from method to method, however the resolution accuracy is often within a range of a few millimeters, providing superior body tracking than using machine learning based pose extraction. Marker based motion capture systems, such as those used in CMU MoCap and Berkeley MHAD, require the application of each marker to the individual at certain predetermined locations, and variation in placement of the markers on the body from sequence to sequence can introduce small errors in obtaining truly explicit spatial truths. The use of depth maps to extract an estimated 3D pose of the subject in the scene has become a prominent inclusion in depth based HAR datasets such as MSR Action3D, K3HI, SBU Kinect Interaction, CAD120, and CAD60. The observation is fed into a skeleton extractor, such as the OpenNI, Microsoft Kinect SDK softwares, or custom methods [157–159], in which a subject is located and a human skeleton model is fitted, predicting the 3D coordinates for a number of joints. Although an approximation of true 3D spatial orientation of the joints, depth sensors and joint tracking has been shown to be relatively accurate in the tracking of humans [34, 36]. The use of bounding boxes to describe regions of interest in a scene are common within appearance based datasets, such as BEHAVE, CAVIAR, ETISEO and MSR Action, especially those that consider person-object interactions or belong to the surveillance domain. They simply provide an area of focus that contains relevant annotated information, such as object and subject location. The use of silhouette masks also provide a region of interest, whilst simultaneously removing external and internal appearance information, representing the subject as a binary classification as either belonging to the background or foreground. These regions of interest can also be utilized to validate action detection and localization methodologies, removing the unwanted information from the overall observation.

2.6. Viewpoint

Camera based methods can also make use of various viewpoints, from single camera to multi-camera simultaneous viewpoint capture. Viewpoints can also differ greatly, capturing events from roughly a parallel plane with the ground, elevated above head height, or from an almost top-down viewpoint. Often events are captured from a viewpoint that is roughly parallel to the ground, producing observations that are almost representative of a human-eye view of the event, examples can be found in MSR Action3D, K3HI, and CMU MoCap. A summary of dataset viewpoint representation is given in Table 8. Sets such as BEHAVE, UT Interaction and CASIA contain events recorded from an elevated angle; these viewpoints are common within the surveillance domain due to the positioning of surveillance cameras for capturing a large scene at once. Recently there has been work towards the recognition of actions from a first person perspective, with data captured from the viewpoint of the observer [99, 160, 161].

Table 7: Description of ground truths provided by datasets.

Name	Spatial ground truth labels	Class ground truth labels
50 Salads	-	Frame labeling
BEHAVE	Bounding boxes	Frame annotation
Berkley MHAD	MoCap tracking	File labeling
BIT Interaction	-	File labeling
CAD120	Extracted skeleton, bounding boxes	Frame labeling
CAD60	Extracted skeleton	File labeling
CASIA	-	File labeling
CAVIAR	Bounding box	Frame labeling
CMU MMAC	MoCap tracking	File labeling
CMU MoCap	MoCap tracking	File labeling
CONVERSE	Extracted skeleton	Frame labeling
Drinking/Smoking	Bounding box	Frame labeling
ETISEO	Bounding box	Frame labeling including calibration parameters, scene descriptions, object affordance
G3D	Extracted skeleton	File labeling
G3Di	Extracted skeleton	File labeling
HMDB51	Bounding boxes	File labeling including view, camera motion, visible body parts, quality, and number of subjects
Hollywood	-	Frame labeling
Hollywood-2	-	Frame labeling
Hollywood 3D	-	File labeling
HumanEVA-I	MoCap tracking	File labeling
HumanEVA-II	MoCap tracking	File labeling
IXMAS	Silhouette masks	Frame labeling
JPL	-	Frame labeling
K3HI	Extracted skeleton	File labeling
KTH	-	Frame labeling including scenario labeling
LIRIS	Bounding boxes	Frame labeling
MPI08	MoCap tracking and 3D scan	File labeling
MPII Cooking	-	Frame labeling
MPII Composite	-	Frame labeling
MSR Action-I	Bounding box	Frame labeling
MSR Action-II	Bounding box	Frame labeling
MSR Action3D	Extracted skeleton	File labeling
MSR DA3D	Extracted skeleton	File labeling
MSR Gesture3D	Extracted skeleton	File labeling
MuHAVi	Silhouette masks	Frame labeling
Olympic Sports	-	File labeling
POETICON	MoCap tracking	File labeling
Rochester AoDL	-	File labeling
SBU Kinect Interaction	Extracted skeleton	File labeling
Stanford 40 Actions	Bounding box	File labeling
TUM Kitchen	Markerless MoCap tracking	Frame labeling including body trunk, left arm, right arm, and object affordance
UCF101	-	Frame labeling
UCF11	-	Frame labeling
UCF50	-	Frame labeling
UCF Sport	-	File labeling
UMPM	MoCap tracking	File labeling
UT Interaction	Bounding box	Frame labeling
ViHASi	Silhouette masks	File labeling
VIRAT	Bounding box	Frame labeling including object affordance
Weizmann	Silhouette masks	File labeling
WVU MultiView-I	-	File labeling
WVU MultiView-II	-	File labeling

This field is often working towards the understanding of interactions by robots for the purpose of human-robot interaction. Such a viewpoint is believed to provide more meaningful information when the observer has an active role in the interaction rather

Table 8: Comparison of dataset viewpoints and scenario control.

	Appearance sets	Pose sets
Simultaneous Views		
Monocular	BIT Interaction, Drinking/Smoking, HMDB51, Hollywood, Hollywood-2, JPL, KTH, MPII Cooking, MPII Composite, MSR Action-I, MSR Action-II, Olympic Sports, Rochester AoDL, Stanford 40 Actions, UCF101, UCF11, UCF50, UCF Sport, UT Interaction, Weizmann	50 Salads, CMU MoCap, G3D, G3Di, Hollywood3D, K3HI, LIRIS, MSR Action3D, MSR DA3D, MSR Gesture3D, SBU Kinect, UMPM
Multi-view	BEHAVE, CASIA, CAVIAR, ETISEO, IXMAS, MuHAVi, TUM Kitchen, ViHASi, WVU MultiView-I, WVU MultiView-II	Berkeley MHAD, CAD120, CAD60, CMU MMAC, HumanEVA-I, HumanEVA-II, MPI08, POETICON, **CONVERSE**
Environment		
Interior Natural	CAVIAR, Drinking/Smoking, HMDB51, Hollywood, Hollywood-2, JPL, MuHAVi, Olympic Sports, Stanford 40 Actions, UCF101, UCF11, UCF50	Hollywood3D
Interior Controlled	IXMAS,MPII Cooking, MPII Composite, Rochester AoDL, ViHASi, WVU MultiView-I, WVU MultiView-II	50 Salads, Berkeley MHAD, CAD120, CAD60, CMU MMAC, CMU MoCap, G3D, G3Di, HumanEva-I, HumanEva-II, K3HI, LIRIS, MPI08, MSR DA3D, MSR Gesture3D, POETICON, SBU Kinect Interaction, TUM Kitchen, UMPM, **CONVERSE**
Exterior Natural	BEHAVE, BIT Interaction, Drinking/Smoking, ETISEO, HMDB51, Hollywood, Hollywood-2, MSR Action-I, MSR Action-II, Olympic Sports, Stanford 40 Actions, UCF101, UCF11, UCF50, UT Interaction, VIRAT	Hollywood3D
Exterior Controlled	BIT Interaction, KTH, Weizmann	

than simply observing a scene, as is the case in human-robotics interactions. There are also datasets which attempt to capture simultaneous multi-camera views of an event for the purpose of evaluating supposedly pose-invariant methodologies. Sets such as WVU MultiView, Berkeley MHAD and TUM Kitchen all contain numerous cameras located in differing positions capturing the same scene. Depth based data, such as tracked skeletons and motion capture marker coordinates, can be orientated arbitrarily about its three axes to develop multi-view methodology, with some pose alignment used to reduce the effect of orientation discrepancies, [162]. However this is dependent upon accurate pose estimation in order to provide data which has confident tracking. Due to the nature of extracting pose estimation from depth based methods there are limited numbers of datasets that utilize multiple depth sensors; however Berkeley MHAD provides multiple Kinect recordings alongside it's vast number of appearance views, with the sensors located in positions from which the infrared sensors are not causing occlusions.

2.7. Use in community

Popularity of a dataset within the community can be difficult to evaluate, however here we attempt to identify the number of citations that are made to the dataset's de-

Figure 2: Example images for Drinking/Smoking dataset

scription publication via Google Scholar. Using this count as a measure of how well adopted a given dataset has become, we rank each set in Table 9. Note that older sets can often show higher citation due in part to their steady accumulation of references over time. Similarly, the number of citations made may not explicitly reflect the use of dataset as a benchmark, as often the datasets are published in parallel with a novel methodology which may accrue its own citations. It can be seen from Table 9 that the pose based datasets show considerably fewer citations, most likely due to the relative age of the rapidly growing field.

3. Current datasets

The following section will now detail the datasets evaluated above, describing the composition of each dataset and a brief discussion of their usage in literature. We also report on some of the accuracy rates achieved using each dataset; however due to the multitude of evaluation criteria these are used as an indicative measure of the dataset complexity as opposed to a definitive survey of state-of-the-art results obtained. It would be unfair to directly compare results obtained between datasets, or even within datasets for differing purposes. Such a survey would require extensive analysis to ensure that cross comparison between results are fair and reflective of their achievement.

The section is divided into the appearance and pose based datasets, with further grouping into their respective abstraction levels as described by figure 1.

3.1. Appearance based datasets

Even though we wish to examine datasets that utilize pose estimation techniques for action recognition, we will briefly discuss availability and impact of video based sets. Video provides a relatively cheap method of obtaining sample sequences, with both real-world and staged executions being obtained. Collection methods can make use of single or multi-camera setups. Actions can be performed from a singular viewpoint, most often face-on, or from differing angles.

3.1.1. Action

Drinking/Smoking. The Drinking/Smoking dataset [78], Figure 2, contains 308 sequences of either drinking or smoking actions taken from 3 sources (two movies and one custom recorded set). The dataset can be used for detection, recognition and localization evaluation. There are 159 instances of drinking and 149 of smoking, from

Table 9: Citation count for dataset description paper. Correct at time of submission. Note: CMU MoCap has no attributed publication

Name	Year of Publication	Total Citations
Appearance		
KTH	2004	2013
Hollywood	2008	1772
Weizmann	2005	1182
UCF11	2009	602
IXMAS	2006	590
UCF Sport	2008	584
Hollywood-2	2009	580
Drinking/Smoking	2007	327
UT Interaction	2009	303
Olympic Sports	2010	283
Rochester AoDL	2009	266
HMDB51	2011	265
MSR Action-I	2009	189
UCF101	2012	155
VIRAT	2011	144
Stanford 40 Actions	2011	137
UCF50	2013	131
ETISEO	2007	103
CAVIAR	2004	90
MSR Action-II	2011	82
MPII Cooking	2012	67
MuHAVi	2010	60
MPI08	2010	48
JPL	2013	38
ViHASi	2008	33
BEHAVE	2010	33
MPII Composite	2012	32
BIT Interaction	2012	19
CASIA	2009	12
WVU MultiView	2011	0
Pose		
HumanEVA	2010	373
MSR Action3D	2010	333
MSR DA3D	2012	311
CAD120	2012	159
TUM Kitchen	2009	117
CAD60	2013	81
MSR Gesture3D	2012	75
Berkeley MHAD	2013	50
CMU MMAC	2008	48
SBU Kinect Interaction	2012	33
Hollywood3D	2013	32
G3D	2012	28
POETICON	2011	8
UMPM	2011	7
50 Salads	2013	6
LIRIS	2014	5
CONVERSE	2015	4
K3HI	2013	2
G3Di	2014	0
CMU MoCap	-	-

either a front or side viewpoint. Instances are taken from two movies and some custom lab recordings. The dataset provides the training and testing samples that were utilized for method evaluation in [79], allowing for direct comparison to the original methodology. The authors in [79] used singular key frames, coupled with a space-time action

Figure 3: Example images for HMDB51 dataset - punch, swing baseball, and hand shake

Figure 4: Example images for Hollywood dataset - sit down, answer phone, handshake

classifier to detect the events in a given scene. The dataset has been utilized for validation several times, notably in the evaluation of modeling person-object interactions via the object trajectory [163] and the recognition of an action class based on a single observation training instance [164].

HMDB51. The HMDB51 [86], Figure 3, is a dataset of 6849 video sequences of 51 different actions with a minimum of 101 executions per label. Videos are taken from a mixture of online clips, movies and television. The actions encompass 5 perceived top level classes; facial expressions, facial object interaction, body movement, body object interaction and person-to-person interaction. The dataset also provides detailed labeling of video quality, number of people in scene, viewpoint, visible body parts and camera motion. Instances of the same class can vary greatly in terms of the execution style, the subject appearance, the quality of the images and the camera view. As such the HMDB51 dataset is one of the more challenging appearance datasets for use as an evaluation tool. The original publication attempts to use the HOG/HOF feature combination to recognize action events within the scene [87], developing a collection of visual words to train an SVM classifier. It has since been used for the recognition of actions within natural settings and loosely controlled parameters [135, 165, 166].

Hollywood. The Hollywood dataset [88], Figure 4, intends to provide realistic human behaviors from unconstrained videos, namely those produced for purposes other than HAR, e.g. films and television. The set provides 5 action classes: answer phone, get out of car, sit down, sit up, and stand up and 3 interaction classes: handshake, hug, and kiss. The sequences are automatically annotated by forming alignments with the script, subtitle and time stamps of the sequence. A subsample of these have been manually corrected to provide 'clean' training and testing sets. In the associated publication [89] the videos are represented by STIPs at multiple spatio-temporal scales. Each STIP is then used to generate a set of HOG and HOF features which are then used to train a non-linear Support Vector Machine (SVM) for event classification. The Hollywood

Figure 5: Example images for Hollywood-2 dataset - hug, answer phone, stand up

dataset, and it's sister dataset Hollywood2 (see below), are often used for the validation of methods under realistic conditions; due to the high variation in the quality and examples of behaviors observed.

Hollywood-2. The Hollywood-2 dataset [90], Figure 5, is an extension on the Hollywood dataset, greatly increasing the number of observed sequences and action classes. The set contains 3669 sequences of 8 single person actions and 4 interactions. There is a large overlap with the Hollywood dataset in terms of the action classes provided, including answer phone, get out of car, handshake, hug, kiss, sit down, sit up, and stand up. The set also introduces 4 new classes; drive car, eat, fight person, and run. To explore the relationship between an action and the scene it occurs within the dataset provides 10 scenario locations, with a large focus on interior environments. The set takes scenes from 69 movies and automatically annotates them using the same script synchronization as with the Hollywood dataset. The set is divided into a training and testing set, selecting given films for each set. There is some intersection between the Hollywood and Hollywood-2 sets, with some films being included in the training set for Hollywood and the testing set for Hollywood-2, thus the two sets should be used independently of each other to avoid issues in training on samples that may be duplicated in the training sets. Marszalek [91] utilizes the set for the learning of both actions and scenes; locating space-time salient motion with a 3D-Harris detector, and static salient areas using 2D-Harris regions. They compute HOG/HOF descriptors from the 3D-Harris, and Scale Invariant Feature Transform (SIFT) descriptors from the 2D-Harris points. These features provide a vocabulary for a bag of words representation of the scene and action. Hollywood-2 has been used as an evaluation set for a number of studies, including multi-modal fusion of audio-visual cues for action recognition [31] and the use of action primitives for classification [167].

IXMAS. The INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset [96], Figure 6, is a multi-view dataset designed for view invariant HAR. 5 cameras capture simultaneous views of 12 actors performing 13 actions with 3 repetitions; check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, and throw. There is an additional labeling for the action class 'nothing', and the throwing action is divided into an over-head and underarm subclass. The ground truth is provided in the form of frame-by-frame annotation of the action class label present within the scene, subject silhouettes, and reconstructed volumes. The dataset is initially used for the recognition of actions regardless of viewpoint [168]

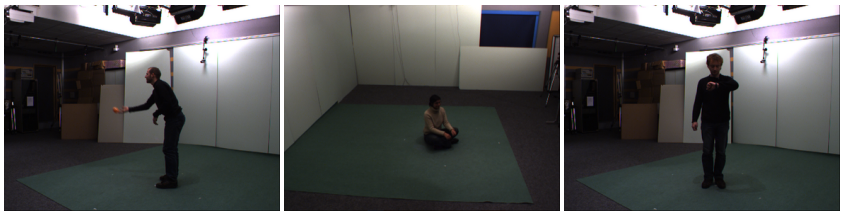


Figure 6: Example images for IXMAS dataset - throw, sit down, check watch

Figure 7: Example images for KTH dataset - punch, run, hand waving

and history volumes [97]. It has since been used widely to evaluate methodology on view-invariant recognition [169–171].

KTH Action. Presented in 2004 by Schuldt *et al.* [22], the KTH dataset [102], Figure 7, consists 25 subjects performing 6 actions in 4 scenario types, recorded via a static camera. Actions are performed with single subjects visible in a frame, with multiple executions of an action in a sequence. Actions performed were walking, jogging, running, boxing, hand waving, and hand clapping. Scenarios covered involved outdoor, scale variations, clothing variation and indoor recordings. For the original study, the 600 continuous recordings are divided to provide 2391 single execution sequences. Despite the simplistic nature of the actions performed, the set has become prominent within the appearance based HAR community, with hundreds of citations making use of the dataset for validation. The original study [22] used the dataset to extract local space-time features from the observation for classification. Of the vast number of subsequent uses of the KTH dataset there have been uses of individual sequences for classification methodologies [172–175], while several sequences are often appended to test segmentation methods [44].

MSR Action. The Microsoft Research group have provided a number of appearance based HAR datasets, including MSR Action-I and Action-II, Figure 8,. These sets are readily available to the research community at [112] and include actions, daily activities and gestures. The Microsoft Research (MSR) Action I dataset [112] contains 16 video sequences of 10 subjects performing 3 different action classes: clapping, waving and boxing. Each sequence contains continuous recording of different actions being carried out in series, often in a cluttered outdoor environment or with multiple subjects in

Figure 8: Example images for MSR Action-I dataset - boxing, boxing and waving, clapping

Figure 9: Example images for MuHAVi dataset - climb ladder, pick up and throw, punch

the observation. Manually provided ground truth labeling is given as a spatio-temporal bounding box over for each frame in which the action is present, and in [113] correct detection is determined by the overlap of ground truth and prediction areas. The dataset is used within [113] for the purpose of action detection and localization within the scene; it has since been used for evaluating a variety of HAR recognition and detection methods [176, 177] MSR Action II [112] is an expansion on the previous set, containing the same 3 action classes; clapping, waving and boxing. The set includes 54 continuous video sequences recorded in crowded environments, including multiple subjects and non-subject individuals. Both the MSR Action I and Action II datasets contain action classes that allow them to be overlapped with the KTH dataset, intending to promote cross-dataset action detection evaluation. The dataset has been used to validate methods in action detection, localization and recognition [114, 178, 179]

MuHAVi. The Multicamera Human Action Video (MuHAVi) dataset [116], Figure 9, is a large scale multi-view action recognition set, capturing 17 action classes performed by 14 subjects. The action classes are performed within the capture area and include punch, kick, run and stop, walk and turn, collapse, pull object, pick up and throw, walk and fall, look in car, crawl, wave, draw graffiti, jump over fence, drunk walk, climb ladder, smash object, and jump over gap. Many of these classes contain sub-action primitive action classes themselves, which can either be handled separately or as a compound action. The project is ongoing, and provides ground truth silhouette masks for a number of sequences and ground truth frame annotation for all sequences. A large number of publications use the MuHAVi set for evaluation of action recognition and view invariant methods, most of which are detailed in [116].

Figure 10: Example images for Olympic Sports dataset - diving springboard, snatch, tennis serve

Figure 11: Example images for Stanford 40 Actions dataset - applauding, fixing a bike, jumping

Olympic Sports. The Stanford Olympic Sports dataset [118], Figure 10, contains video clips of 16 sport actions taken from YouTube; high-jump, long-jump, triple-jump, pole-vault, basketball, bowling, tennis-serve, platform, discus, hammer, javelin, shot put, springboard, snatch, clean and jerk, and vault. The clips include cluttered scenes, dynamic camera movement, varying scales and execution styles. [119] uses the dataset to evaluate their method of modeling temporal structure motion for action recognition. The suggested testing and training split of the 50 video sequences is provided at [118] and the ground truth is provided as simple whole sequence labels.

Stanford 40 Actions. The Stanford 40 Actions dataset [180], Figure 11, is a collection of 9532 still images that represent naturally executed actions including riding a horse, rowing a boat, fishing, applauding, and smoking. There are between 180 and 300 images per action class and the dataset provides bounding box annotation for the subject in the observation for the purpose of action localization and recognition. The challenge of understanding human action from a singular instance is explored in [127], learning the context between actions and the objects contained within the image, with further study into the use of still image understanding being evaluated on the dataset [181, 182].

UCF. The UCF action datasets are a collection that make use of video to represent action sequences. UCF-11, UCF-50 and UCF-101 are all video sets taken from YouTube designed to provide an action recognition problem that focuses on the accurate recognition of observations in which there are highly variable training observation samples.

The UCF-11 dataset [132], Figure 12, was produced to enable the evaluation of recognition methods upon unconstrained observations of an action class. The collec-

Figure 12: Example images for UCF11 dataset - biking, basketball, tennis swing

Figure 13: Example images for UCF50 dataset - high jump, skiing, yo-yo

tion provides 1168 sequences from 11 different action classes, with over 100 instances in each action class. The actions presented include basketball shooting, cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. Liu [133] presents this dataset as an evaluation set for the understanding of action classes from natural observations that were produced for reasons other than for HAR, providing little knowledge about camera quality, viewpoint and motion. The samples are grouped into 25 categories, with each category containing numerous instances of the same action from similar scenarios.

UCF-50 [134], Figure 13, extends upon the UCF-11 dataset by introducing yet more action classes, increasing the total count to 50, including baseball pitch, basketball shooting, bench press, biking, billiards shot, breaststroke, clean and jerk, diving, drumming, fencing, golf swing, playing guitar, high jump, horse race, horse riding, hula hoop, javelin throw, juggling balls, jump rope, jumping jack, kayaking, lunges, military parade, mixing batter, nunchucks, playing piano, pizza tossing, pole vault, pommel horse, pull ups, punch, push ups, rock climbing indoor, rope climbing, rowing, salsa spins, skate boarding, skiing, skijet, soccer juggling, swing, playing tabla, tai chi, tennis swing, trampoline jumping, playing violin, volleyball spiking, walking with a dog, and yo yo. Again the initial use of the dataset is in the recognition of actions from an unconstrained set of recordings [135]. This dataset has since been superseded by the UCF-101 dataset.

UCF-101 [130], Figure 14, is the latest extension of the UCF appearance based action datasets, containing 101 separate action classes collected from various sources, which are grouped into 5 activity types; Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments and Sports. The sub-activity

Figure 14: Example images for UCF101 dataset - apply eye makeup, drumming, mopping floor

actions include apply eye makeup, apply lipstick, archery, baby crawling, balance beam, band marching, baseball pitch, basketball shooting, basketball dunk, bench press, biking, billiards shot, blow dry hair, blowing candles, body weight squats, bowling, boxing punching bag, boxing speed bag, breaststroke, brushing teeth, clean and jerk, cliff diving, cricket bowling, cricket shot, cutting in kitchen, diving, drumming, fencing, field hockey penalty, floor gymnastics, frisbee catch, front crawl, golf swing, haircut, hammer throw, hammering, handstand pushups, handstand walking, head massage, high jump, horse race, horse riding, hula hoop, ice dancing, javelin throw, juggling balls, jump rope, jumping jack, kayaking, knitting, long jump, lunges, military parade, mixing batter, mopping floor, nunchucks, parallel bars, pizza tossing, playing guitar, playing piano, playing tabla, playing violin, playing cello, playing daf, playing dhol, playing flute, playing sitar, pole vault, pommel horse, pull ups, punch, push ups, rafting, rock climbing indoor, rope climbing, rowing, salsa spins, shaving beard, shotput, skate boarding, skiing, skijet, sky diving, soccer juggling, soccer penalty, still rings, sumo wrestling, surfing, swing, table tennis shot, tai chi, tennis swing, throw discus, trampoline jumping, typing, uneven bars, volleyball spiking, walking with a dog, wall pushups, writing on board, and yo yo. Over 13,320 sequences are collected to provide over 100 instances of each action class, with each sequence containing variation in subject, scenario and camera parameters. The original publication [131] provides a baseline recognition score by extracting Harris3D corners from a clip and representing them via HOG/HOF descriptors. These descriptors were then used to generate a histogram of video words, utilizing the training and testing splits provided at [130] to evaluate the performance of an SVM developed on the histogram vectors. These baseline results allow for the evaluation of novel methods on the previously developed methods by utilizing the benchmark splits.

Overall the use of the UCF Action datasets for method evaluation has been reported within numerous publications, especially in the recognition of action classes from observations that contain little similarity in regards to the camera positioning and quality [165, 166, 183].

UCF Sport. The UCF Sport dataset [136], Figure 15, is similar in construction to the previous UCF Action sets, with sequences being collected from previously recorded events. The main difference is that the focal domain of the dataset is within the recognition of sporting activity domain, providing class labels such as diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging, and walking. The

Figure 15: Example images for UCF Sport dataset - skateboarding front, kicking side, golf swing side

Figure 16: Example images for ViHASi dataset - punch, running, jump kick

dataset collects 200 sequences and contains the same unconstrained camera parameters as the previous action sets. The dataset has since been used for the evaluation of action recognition methodologies both within the generic and sports specific domain, including [174, 184, 185].

ViHASi. The Virtual Human Action Silhouette (ViHASi) dataset [142], Figure 16, provides synthetic silhouette masks that have been produced by using 20 actions performed by 9 virtual actors. The use of a virtual environment has allowed for the generation of 40 virtual viewpoints from which the silhouettes are produced, creating a dataset that provides evaluation of view invariant silhouette based action recognition. The 20 action performances are generated using the same motion capture sequences, ensuring that each virtual actual performs the same action execution. Classes include hang on bar, jump on bar, jump over object, run and pull object, run and push object, run and turn left, run and turn right, hero smash, hero door slam, knockout spin, knockout, grenade, collapse, stand and look, punch, jump kick, walk, walk and turn back, and run. Differing subjects were developed that included not only differences in body proportions but also variation in clothing which impact upon the silhouettes produced. Despite being a niche dataset there have been several works that make use of the ViHASi dataset, evaluating the use of silhouette pose projection for action recognition [143, 186–188].

Weizmann. One of the three main appearance based action recognition datasets, the Weizmann dataset [24, 146], Figure 17, provides RGB recordings of 10 actions performed by 9 subjects, captured at 50fps. Actions performed were running, walking, skipping, jumping-jacks, jump forwards, jump in place, sideways gallop, two-handed

Figure 17: Example images for Weizmann dataset - bend, jump forwards, two-handed wave

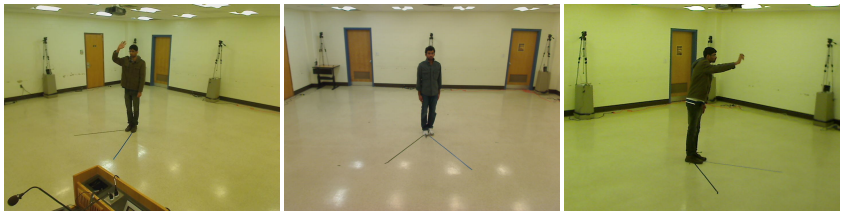


Figure 18: Example images for WVU MultiView dataset - waving, standing still, waving from alternate viewpoint

wave, one-handed wave and bend. Each recording uses a static camera to capture multiple executions of the same action against a solid wall background. Tangential to the main dataset are a series of samples deemed for method robustness testing; with recordings including occlusion, abnormal execution style and carrying objects. Despite providing readily definable action classes, the Weizmann dataset has been used repeatedly for HAR method validation since its creation [48, 173, 174, 189–192].

WVU MultiView. The WVU MultiView dataset [148], Figure 18, is comprised of two sets; with WVU MultiView-I containing sequences of a single action execution from one of 12 action classes, and WVU MultiView-II describing continuous combinations of 9 available actions in an interleaved fashion. The intention is to utilize the first dataset to perform action recognition, while the second requires detection and segmentation. WVU MultiView makes use of 8 cameras to collect data that can be used to test view-invariant methods. The classes included in the first dataset are standing still, nodding head, clapping, waving one hand, waving two hands, punching, jogging, jumping jack, kicking, picking up, throwing and bowling. In WVU-II the actions show slight overlap; including clapping hands, waving one arm, waving two arms, punching, jogging in place, jumping in place, kicking, bending, and underarm bowling. The two datasets are often used as a validation method for multiple distributed camera action recognition [149, 150, 193].

3.1.2. Interaction

BEHAVE. BEHAVE, Figure 19, is a human action recognition project that is comprised of two separate datasets, the Multiagent Interaction dataset [60], featuring multi-

Figure 19: Example images for BEHAVE dataset - in group, fight, following

Figure 20: Example images for BIT Interaction dataset - bow, handshake, push

person interactions captured from an elevated viewpoint, and the Optical Flow set [194], containing recordings of human flow in a train station exit under three scenarios. The Multiagent Interaction set contains 10 interaction classes being performed by multiple individuals in an outdoors environment, recorded using static RGB cameras from 2 non-simultaneous viewpoints. The set contains bounding box annotations of the individual and their action class, including; InGroup, Approach, WalkTogether, Meet, Split, Ignore, Chase, Fight, RunTogether, Following. There are 163 instances with varying number of instances per class. The class WalkTogether contains 43 instances at a total of 6694 frames, while the classes Meet and Following only contain a single instance each, comprising of less than 100 frames each. The majority of the sequences are annotated by providing the ground truth individual bounding box locations. Action labels are then provided along with start and end frames for the event; this is coupled with the identification labels for each individual involved in the event. Ref. [60] also provides image pixel position measurements for the computation of the ground plane homography. One viewpoint of the set is captured from inside a building, filmed through a window, and as such contains a large amount of noise in the illumination of the scene due to reflections from the glass. The BEHAVE Optical Flow dataset has been used several times for event detection in crowded scenes [195, 196], while the Multiagent Interaction set has been utilized for the recognition of actions and identification of individuals within the surveillance domain [197–200].

BIT-Interaction. The Beijing Institute of Technology (BIT)-Interaction dataset [63], Figure 20, consists of 400 AVI video clips capturing 8 interaction events with 50 videos per class. The dataset provides a further level of complexity by introducing varying occlusions, appearances, temporal and spatial scale. The classes include these which

Figure 21: Example images for CAVIAR dataset - fight, slump, leave bag unattended

Figure 22: Example images for ETISEO dataset

are definable by their respective poses, such as bow, boxing, handshake, high-five, hug, kick, pat, and push. The presence of pedestrians, occlusions and variable views results in a dataset that can be used for detection, localization and recognition. The original publication [64] utilized the BIT-Interaction set to develop a set of high-level phrases which describe the interaction in terms of the interdependencies of lower-level attributes belonging to each individual in the interaction.

CAVIAR. The Context Aware Vision using Image-based Active Recognition (CAVIAR) project [70], Figure 21, provides action recognition sets for the purpose of determining if local image descriptors guided by contextual knowledge of the scene can improve image-based action recognition. The project provides two RGB sets, one from an entrance lobby of the Institut National de Recherche en Informatique et en Automatique (INRIA) labs, the second utilizes 2 simultaneous views from within a shopping center. The INRIA subset provides 4 single person actions: walk, browse, rest/slump, and leave bag unattended; the set also provides 2 interaction classes: meet and fight. The shopping center subset provides the remaining 3 action classes enter shop, window shop, and leave shop. Both subsets provide the ground plane homography measurements for the scene. The ground truth labeling in the sets are XML hand labeled bounding boxes for each image in the sequence. The CAVIAR dataset is one of the most utilized appearance based datasets for human action recognition alongside KTH and Weizmann, being utilized for evaluation of numerous methodologies, including tracking, recognition and segmentation [201–203].

ETISEO. The ETISEO dataset [80], Figure 22, provides a methodology and accompanying dataset for video surveillance evaluation. 5 main scenarios are presented, con-

Figure 23: Example images for JPL dataset - throwing object, handshake, punch

taining instances of 15 action classes. 10 are single person actions: walk, run, sit, lying, crouching, holding, jumping, pick up, put down, and tailgate. 5 are person-person interactions: push, fight, meet, exchange, and queue. The dataset is annotated with bounding boxes detailing both events and physical objects within the frame, and the project provides evaluation metrics on a variety of problems including object detection, object localization, object tracking, object classification, and event recognition. The project is a multi-institute venture into developing a benchmark for evaluating security and surveillance domain observations, containing detailed information regarding it's use as a validation tool and the data structures provided [81]. ETISEO has been used as a benchmark for the detection, localization, and recognition of both pedestrians and behavioral actions in a surveillance domain [204–206].

JPL First-Person Interaction. The Jet Propulsion Laboratory (JPL), Figure 23, at the California Institute of Technology provide a first person viewpoint dataset into person-person interactions [98]. The project captures a mixture of positive and negative interactions from a camera positioned on a non-static subject as they traverse an office environment. During the sequences the subject encounters 7 interactions which are recorded from their perspective of recipient; including handshake, petting subject, wave at subject, conversation with pointing at subject, punching subject, and throwing objects at subject. In the associated publication [99] the use of local motion descriptors across space-time provides a bag of visual words representation for recognition of first person recipient view interactions. The egocentric domain is often used for determining the actions of the observer and of the subject observed [99, 207], with complications arising from the motion and perspective captured by the camera's location on the body [208, 209].

UT-Interaction. The UT-Interaction dataset [140], Figure 24, contains 20 continuous static camera recordings of multiple subjects performing multiple interactions within a scene. Each recording captures all action classes recorded from an elevated angle. The interactions between two individuals including handshake, hug, kick, point, punch, and push. Several subsets are present within the dataset; with static backgrounds, dynamic backgrounds, multiple events in the scene and crowded scenes. Ground truths are provided in terms of bounding box frame-by-frame annotation. The UT-Interaction set has often been used for evaluating interaction recognition within a surveillance domain [141, 210, 211].

Figure 24: Example images for UT Interaction dataset - kick, handshake, hug

Figure 25: Example images for VIRAT dataset

Figure 26: Example images for MPII Cooking dataset

VIRAT. The VIRAT dataset [144], Figure 25, provides action classes expected within a surveillance domain, describing natural scenes and interactions between individuals and the environment. The sequences within the collection are annotated with a high level of detail, providing information via bounding box annotations regarding the people, objects, vehicles and the interactions that occur between them. 12 activity classes are provided, including loading an object on a vehicle, unloading an object from a vehicle, opening the trunk of a vehicle, closing the trunk of a vehicle, getting in to a vehicle, getting out of a vehicle, gesturing, digging, carrying an object, running, entering a facility, and exiting a facility. The challenge of this dataset is in the natural executions of the interactions, and also in the cluttered scene that is observed, a common problem task for real world surveillance domain. Common use of the VIRAT dataset is in the analysis of surveillance domain action detection and localisation [212–215].

3.1.3. Activity

CASIA. The CASIA action database [68] provides a multi-view action and interaction dataset containing 8 single person actions and 7 person-person interactions. The single person actions include walk, run, bend, jump, crouch, faint, wander, and punch car. The interactions include rob, fight, follow, follow and gather, meet and part, meet and together, and overtake. The scene is captured from three simultaneous static viewpoints; horizontal/side on, top down and angle, although global locations of the cameras are not provided. The choice of viewpoint provides a surveillance style dataset with simultaneous viewing allowing the evaluation of view-invariant methods. Each AVI sequence is annotated as a whole clip by filename; detailing the viewing angle, action class, subject ID, and action repetition number. The CASIA dataset has been used to evaluate view-independent, surveillance based action recognition [69, 216–218].

MPII Cooking. The Max Planck Institut Informatik (MPII) Cooking datasets, Figure 26, are a pair of closely related datasets that concern the daily living activities of cooking and the action and interactions that are compounded into the higher level semantic classes. The MPII Cooking Activities dataset [108] contains 44 continuous recordings of naturally executed daily cooking activities, with 12 participants completing activities that included any number of 65 potential activities, such as chopping and pouring. These fine grained activities are recorded as part of a higher level semantic activity, such as preparing a salad or cake, allowing flow between each action to be natural. The focus of [109] is to detect and recognize the execution of the lower level actions within

Figure 27: Example images for Rochester AoDL dataset - answering phone, chopping banana, eat snack chips

an activity, with the dataset providing detailed frame annotation to facilitate evaluation. The MPII Composite set [110] is an expansion upon the MPII Cooking set, introducing more detailed information regarding the higher level activity classes. 14 activities, such as cake, omelet, mashed potato, and pancake, are provided as composite activities which are built from the finer actions provided by the MPII Cooking Activities set. These two sets provide a method of evaluating methods that are able to recognize events at differing levels of abstraction. The overall activity can be decomposed into a set of lower level actions and interactions.

Rochester AoDL. The Rochester Activities of Daily Living (AoDL), Figure 27, dataset contains 10 natural action classes performed 3 times by five subjects in front of a waist height desk. Actions performed include answering a phone, dialing a phone, looking up a phone number in a telephone directory, writing a phone number on a white-board, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware. The intention of the project is to perform HAR on more realistic executions of behavior classes, with [123, 219–221] using tracked key point trajectories for action recognition and [222] considering the pairwise spatio-temporal relationships of the interest points in the scene.

3.2. Pose Based sets

MoCap has allowed for highly accurate localization of body positioning, using markers to identify joints and bones in coordinates of a volume space. Motion capture techniques often utilize the pose of an individual during an action's execution. There are several purely MoCap datasets available, however most now use MoCap techniques as part of a multimodal collection. In recent years community focus has moved from traditional motion capture techniques to the collection of joint positioning via commercial depth sensors, such as the Microsoft Kinect. Depth data has become prolific in the community since the release of the Microsoft Kinect depth sensor; mostly due to its ability to accurately track a human, and provide a skeleton representation in 3D. Most depth sets also provide their corresponding skeleton representations so that the same skeletons are also part of the standard training and testing methods. It is not only the visual information that is used to identify action events. Often there can be use of accelerometers and gyroscopes to capture the kinematics of the body during the performance of an action. Sometimes these additional depth based modalities are captured in parallel with more conventional methods, sometimes they are the sole modality under focus.

Figure 28: Example images for Berkeley MHAD dataset - two handed wave RGB, two handed wave depth, sit down

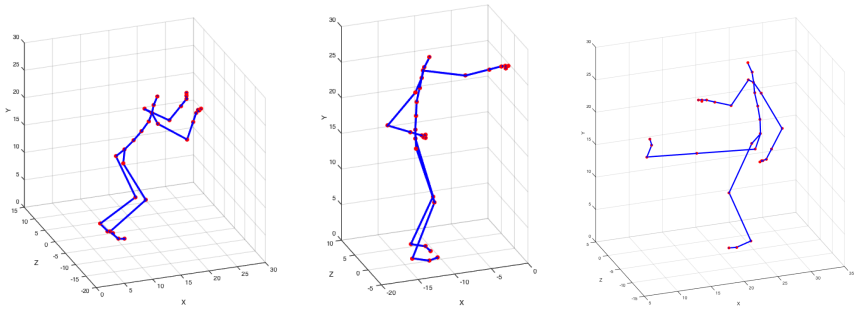


Figure 29: Example images for CMU MoCap dataset - jump, punch, kick

3.2.1. Action

Berkeley Multimodal Human Activity Database. The Berkeley MHAD [62], Figure 28, contains 660 sequences of 12 participants performing 11 actions, recorded using RGB video, depth sensors, marker based motion capture, accelerometers and microphones. Action classes include jumping jacks, bend, punch, two handed wave, one handed wave, clap hands, throw, sit down and stand, sit, and stand. Each class was recorded 5 times, with jumping jacks, bend, punch, two handed wave, one handed wave, clap hands, sit down and stand containing 5 continuous repetitions per recording. 3D coordinates for 43 markers were recorded via 8 MoCap cameras. 12 RGB cameras were grouped into 2 stereo vision clusters and 2 4-camera multi-view clusters. Two Kinect sensors captured RGB-D data. 6 tri-axial accelerometers were affixed to the wrist, ankles and hips to record limb dynamics during an action. Sensor recordings are geometrically and temporally synchronized to allow multimodal HAR. The motion capture system is first calibrated, with RGB and Kinect sensors being calibrated for both intrinsic and extrinsic parameters, referencing all sequences to a the motion capture world coordinate system. Due to the vast amount of data provided across numerous modalities the MHAD dataset has been used to evaluated methods that make use of modality fusion [39, 223], motion capture data [224, 225] and RGB-D joint tracking information for the purpose of action recognition [225, 226]

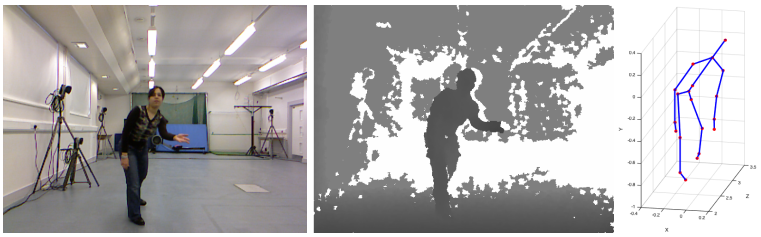


Figure 30: Example images for G3D dataset - bowling RGB, depth, skeleton

Figure 31: Example images for Hollywood3D dataset - run, hug, use phone

Carnegie Mellon University Motion Capture. The CMU Graphics Lab action dataset [74], Figure 29, contains marker-based MoCap sequences of subjects performing a large variety of actions. Sequences are grouped in 6 types; Human Interaction, Interaction with Environment, Locomotion, Physical Activities & Sports, Situations & Scenarios and Test Motions. Sequences can contain multiple action executions and can include person-to-person interactions. The set uses 40-60 markers to capture the full human skeleton of 109 subjects in 2605 sequences. The C3D format markers are not consistent from sequence to sequence, and thus require the user to determine the marker locations beforehand when using the C3D data. However the use of the AMC formatted joint angles are consistent between sequences. Evaluation on the CMU MoCap dataset often utilizes a subset of the overall dataset, due to the large number of sequences and action classes [41, 42, 227].

G3D. The G3D dataset [82], Figure 30, is an action set that focuses on the recognition of actions designed for gaming and computer interaction. 10 subjects perform 20 game based actions, with up to 3 repetitions, in front of a stationary Kinect sensor, capturing synchronized RGB-D data and 20 joint skeletons. Actions recorded by the dataset include punch right, punch left, kick right, kick left, defend, golf swing, tennis swing forehand, tennis swing backhand, tennis serve, throw bowling ball, aim and fire gun, walk, run, jump, climb, crouch, steer a car, wave, flap and clap. The purpose of the dataset is to develop a framework for the real time recognition of actions within a observed scene [83].

Hollywood3D. The Hollywood3D dataset [92], Figure 31, contains similar action classes to that of the appearance based Hollywood and Hollywood-2 datasets; null,

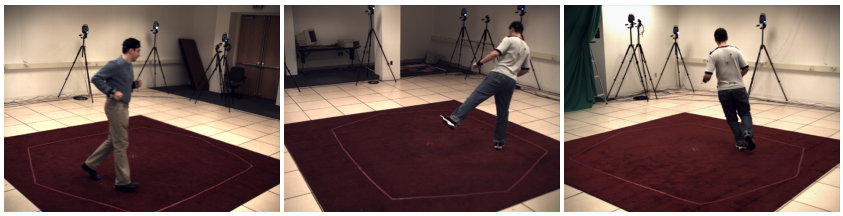


Figure 32: Example images for HumanEva dataset

run, punch, kick, shoot, eat, drive, use phone, kiss, hug, stand up, sit down, swim, and dance. However the data modality in this dataset consists of depth maps obtained from the production of commercial 3D movies. The purpose of [93] is to expand the complexity of using non-HAR based recordings for the purpose of action recognition by representing the observations as depth data. The dataset provided a significant challenge to the community, describing natural observations of action classes as depth information [228–230].

HumanEva. The HumanEVA-I dataset [95, 231] contains video sequences synchronized with motion capture poses, capturing 6 actions performed by 4 subjects. Actions include walking, jogging, gesturing, throwing & catching, boxing, and a combo action. Actions were repeated 3 times, once with MoCap and then twice with a combination of MoCap and video. The set contains separate training, testing and validation sets, detailed in [95], allowing comparative results. The MoCap markers were tracked using 6 ViconPeak cameras, while the video data was collected using 3 RGB cameras and 4 grayscale cameras. The grayscale cameras are located in the corners of the capture space, with the color cameras positioned to the front, left, and right viewpoints of the subject.

HumanEVA-II, Figure 32, then expands on the previous set by having 2 subjects perform combinations of the previous actions to develop a secondary testing set that is ten times smaller than that of HumanEva-I. The dataset is designed as a testing set for the methods developed on the HumanEva-I dataset, providing only complex continuous sequences; starting with walking a path, then jogging, concluding with the subject alternating balancing on each foot. The intent is to use the HumanEva-I set to train and validate the system, with testing be executed on the HumanEva-II set. The MoCap markers are tracked using 12 ViconPeak cameras, twice as many as the original dataset, and the video data is collected by 4 color cameras located in the corners of the capture space. The HumanEva sets have been used repeatedly to evaluate the performance of pose estimation and sequence segmentation algorithms due to it's continuous series of multiple action classes performed in the combo-action observations and the motion capture ground truth [232, 233].

Microsoft Research Action3D. The MSR Action3D dataset [40, 41, 112], Figure 33, provides the first example of a public depth map dataset for HAR, capturing both the depth data for 20 gaming related actions performed by 10 subjects, with up to

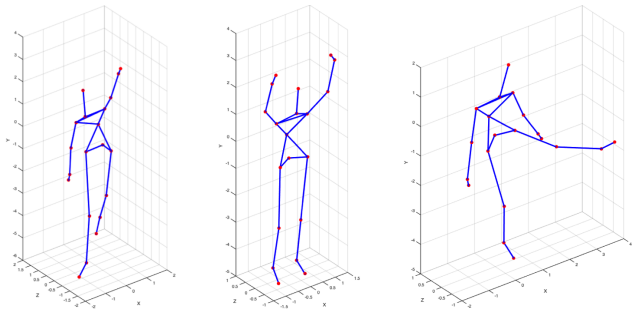


Figure 33: Example images for MSR Action3D dataset - pick up and throw, two hand wave, side kick

Figure 34: Example images for MSR Gesture3D dataset - where, J, finish

3 executions of an action by each subject. Action classes include 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 swing, tennis serve, pickup and throw, and golf swing. In recent years the dataset has been expanded to include the tracked joints in screen and real world coordinates, each skeleton consists of 20 joints captured with a device similar to the Kinect; head, shoulder center, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hand, right hand, spine, hip center, left hip, right hip, left knee, right knee, left ankle, right ankle, left foot, and right foot. Due to the computation involved in learning the 20 total actions the total dataset is divided into 3 subsets, each containing 8 of the 20 possible actions, dubbed Action Sets. Action Set 1 contains similar action classes such as high throw and tennis serve. Action Set 2 contains actions that involve subtle actions with the arms and hands, including draw tick and draw circle. Action Set 3 then aims to group complicated actions together, including the sporting actions. 10 samples are considered to be too noisy and are omitted from the use of the dataset for evaluation [112]. The MSR Action3D dataset is one of the most prominent action recognition depth based datasets available, with numerous action recognition methods utilizing the set for evaluation purposes [28, 39, 45, 162, 234].

Microsoft Research Gesture3D. The MSR Gesture3D dataset, Figure 34, contains 336 sequences of American Sign Language gestures. 10 subjects remain in a seated

Figure 35: Example images for MPI08 dataset

Figure 36: Example images for 50 Salads dataset

position and perform 12 different dynamic sign language gestures in up to three repetitions. The dataset provides the depth maps for each frame in the sequence. In the associated publication [115] it was possible to develop a real-time system to recognize input to the Kinect sensor at 10fps. The dataset is captured from a front-on view, with the lower half of the subject obscured by a table, with focus being on the body, head, arms and hands. Due to it's focus on the behavior displayed by the hands the MSR Gesture3D dataset has often been used in the hand pose estimation and action recognition community [235, 236].

MPI08. The MPI08 dataset [105], Figure 35, collects motion capture recordings of subjects performing tasks for the purpose of multi-modal body tracking fusion. Despite this primary purpose it provides several sequences of highly accurate spatial tracking whilst the subjects execute their actions. The use of modality fusion within this dataset could be exploited for the purpose of action recognition, utilizing the frame labeling of the files for action recognition [106, 107].

3.2.2. Interaction

50 Salads. The University of Dundee 50 Salads dataset [58], Figure 36, is a collection of birds-eye-view recordings of food preparations using an RGB-D sensor and accelerometers for the purpose of recognizing gestures and person-object interactions. 25 participants prepared 2 salads each, utilizing a variety of tools and ingredients, resulting in a total of 966 observed action instances, with an average of over 55 observations per lower level class. The sequences are annotated with a label being assigned to all frames between a given start and stop frame. There are two tiers of labeling, the first

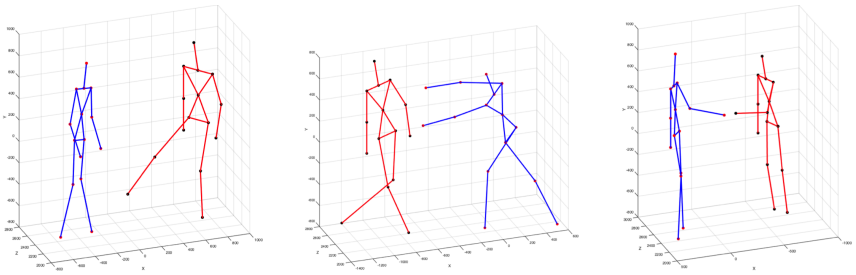


Figure 37: Example images for K3HI dataset - kicking, pushing, shaking

describing the higher level action as one of 3 tasks; cut and mix ingredients, prepare dressing, and serve salad. The second tier describes the frames in terms of a lower level actions such as peel cucumber, cut cucumber, place cucumber into bowl, cut tomato, place tomato into bowl, cut cheese, place cheese into bowl, cut lettuce, place lettuce into bowl, mix ingredients, add oil, add vinegar, add salt, add pepper, mix dressing, serve salad onto plate, and add dressing. Each of the lower level action labels are given a suffix of being either prep, core or post the action. Tri-axis accelerometer recordings are provided for 7 tools used in the sequences. Stein and McKenna [58] provide the RGB video recordings, depth maps, accelerometer sequences and the synchronization of all sequences. The sequences begin with an assistant making 4-5 sharp knocks to an IMU in the scene, allowing synchronization to that point. The 50 Salads set has been used to explore the impact of learning differing levels of abstraction, focusing on the information gained between higher and lower level behaviors [237] and the understanding of complex scenarios [238].

G3Di. A progression on G3D, the G3Di [84, 85] dataset makes use of a single Kinect depth sensor to track two individuals interacting within the scene. This dataset captures 6 pairs of subjects performing actions taken from 6 sports, with 14 action classes. The top level sports are boxing, volleyball, football, table tennis, sprint, and hurdles. The primitive actions include right punch, left punch, defend, overhand hit, underhand hit, jump hit, kick, block, save, serve, forehand hit, backhand hit, run, and jump. The action classes run and serve span two classes, whereas the remainder are top level action specific. The G3Di dataset presents interactions between two individuals who are side by side with both subjects facing the sensor. This makes it possible to separately recognize an individuals action, before then compounding this knowledge to recognize an interaction [85].

Kinect Based 3D Human Interaction. The K3HI dataset [101], Figure 37, contains 8 pairwise person-person interactions performed by 15 individuals. Each of the 320 sequences captures a single execution of one of the 8 action classes; approaching, departing, kicking, punching, pointing, pushing, exchanging an object, and shaking hands. Both individuals in the scene are tracked using the Kinect sensor, capturing the 15 joints of each subject. The OpenNI skeleton representation was extracted, tracking the head, neck, left shoulder, right shoulder, left elbow, right elbow, left hand, right

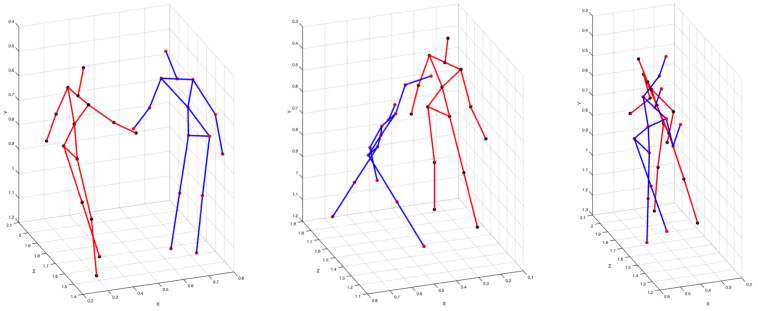


Figure 38: Example images for SBU Kinect Interaction dataset - handshake, punch, hug

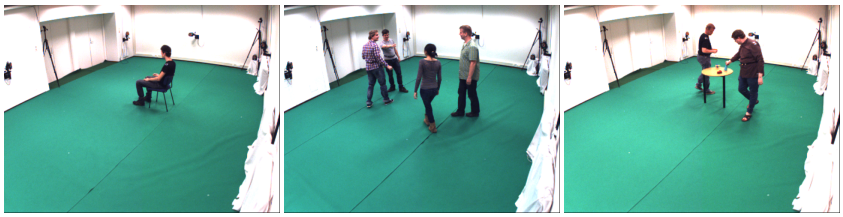


Figure 39: Example images for UMPM dataset - sit (one person), meet (four person), grab (two person)

hand, torso, left hip, right hip, left knee, right knee, left foot, and right foot. The use of approaching and departing classes are dismissed for method evaluation in [101], due to their simplistic nature and high recognition accuracy rates. [101] uses a 4-fold cross validation method for their initially reported experimentation on the dataset, however the training/testing splits are not provided in the dataset itself. The dataset has been used to evaluate positive action recognition in [101].

Stony Brook University Kinect Interaction. The SBU Kinect Interaction dataset [37, 124], Figure 38, presents person-person interaction recorded via synchronized video, depth maps and skeletal models of both actors. 7 individuals, in 21 pairings, performed 8 types of interaction. The interactions between the two individuals are captured from a side-on view and include approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. These interactions provide several classes that involve similar gestures in the arms, namely pushing, punching, shaking hands and exchanging objects. The dataset provides the RGB video and depth maps, alongside the OpenNI 15 joint skeleton tracking. The skeleton joints are head, neck, torso, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, left hip, right hip, left knee, right knee, left foot, right foot. In [37] the dataset was used to identify joint distance and velocity features that are coupled between the individuals in the scene.

Figure 40: Example images for CMU MMAC dataset

UMPM. The Utrecht Multi-Person Benchmark (UMPM) [138, 239], Figure 39, is a synchronized video and marker-based MoCap set that includes numerous subjects interacting and occluding one another. The set intends to provide a ground truth labeled standard dataset for the recognition of dense scenes, at risk of inter- and intra-subject occlusions. The dataset records 9 different scenarios using 4 RGB cameras and a 37 marker MoCap system; 1) walk,jog or run, 2) walk along circle or triangle shape, 3) walk around while another person hangs or sits on chair, 4) sit, lie, hang or stand on table, 5) grab object on table, 6) conversation with gestures, 7) throw or pass ball while moving, 8) stand still and 9) move around. These are all complex activities that contain actions, interactions and higher level activities that require recognition. This is coupled with a multi-view approach and the occurrence of a cluttered scene. UMPM has been used to explore tracking and action recognition that occurs in a scene that contains numerous complex occlusions [239].

3.2.3. Activity

Carnegie Mellon University Multimodal Activity. The CMU Multimodal Activity dataset, [73] Figure 40, commonly known as the CMU MMAC, aims to understand recognition of complicated human daily actions. The dataset utilizes RGB video, marker-based motion capture, audio, accelerometers and gyroscopes for the capture of 5 differing cooking recipes by 43 subjects. 6 RGB camera viewpoints record the scene with a variety of spatial and temporal resolutions, including a first person view from a head-mounted camera. 63 markers are tracked using a Vicon motion capture system of 12 cameras which provide the spatial ground truth for the body tracking. The CMU MMAC database has been used to evaluate the temporal segmentation of complex activities from a first person perspective [240], and the segmentation of joint gestures for classification [241].

Cornell Activity Dataset 60. The CAD-60 dataset [65], Figure 41, provides 60 RGB-D recordings of 4 subjects performing 12 activities across 5 different environments. Participants were captured executing more natural actions, including rinsing mouth, brushing teeth, wearing contact lens, talking on phone, drinking water, opening container, chopping, stirring, talking on couch, relaxing on couch, writing on white-board, and working on computer. The dataset is provided as a collection of RGB images, depth maps and the corresponding 15-joint tracked skeletons. [66] first introduced the

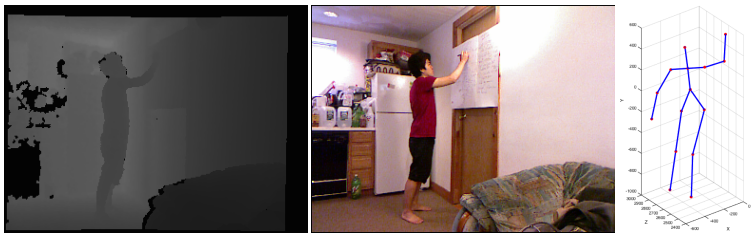


Figure 41: Example images for CAD60 dataset - Using whiteboard depth, RGB, skeleton

Figure 42: Example images for CAD120 dataset - taking medicine RGB, taking medicine depth, making cereal

dataset to classify unstructured human activity by constructing a graph of sub-activities that compound into the top level activities.

Cornell Activity Dataset 120. The CAD-120 set [65], Figure 42, focuses on the execution of long daily activities, capturing high and low level actions. 4 participants provide 120 sequences capturing 10 high level activities, which are each comprised of a number of 10 potential sub-activities. The compound actions include making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, and having a meal. Gestures are labeled as reaching, moving, pouring, eating, drinking, opening, placing, closing, scrubbing and null.

Figure 43: Example images for LIRIS dataset - enter door RGB, exchange object RGB, exchange object depth map

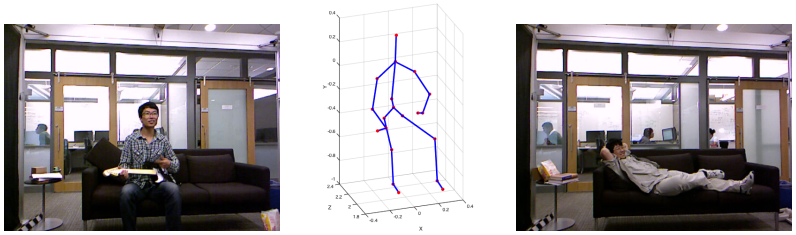


Figure 44: Example images for MSR DA3D dataset - play guitar RGB, play guitar skeleton, lay down on sofa RGB

LIRIS Human Activities. The LIRIS Human Activities dataset [104], Figure 43, provides RGB-D recordings of 21 subjects completing 10 behavioral classes; discussion, giving an item, picking up or putting down, entering or leave room, unsuccessful attempt to enter room, unlock room and enter, leaving an unattended bag, handshake, typing on a keyboard, and talking on a telephone. The dataset attempts to be purposefully difficult by introducing very little constraint in the execution of the behavior, relying more on the semantics of the behavior. To introduce a more realistic representation the use of different contexts and tools within an action class is provided, i.e. different types of phone are used, and discussions can occur seated or standing. Two different semi-independent sets are provided, the first represents the depth maps captured from a Kinect mounted on a joystick controlled robot, the other is a stationary mounted RGB camcorder.

Microsoft Research Daily Activities 3D. Using the Kinect sensor, 10 subjects were recorded performing 16 natural daily actions, [41], Figure 44; drink, eat, read book, call cellphone, write on paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, sit down. The actions were recorded, were possible, in both standing and seated positions in a living room environment. The majority of the action classes involve person-object interactions, such as toss paper and write on paper, thus provide a more real-world set of observations for the purpose of HAR. As with the MSR Action3D dataset, the RGB and depth map sequences are provided alongside the tracked 20-joint skeletons. No standard training/testing splits are provided either within the description paper [41] or the web location [112]. MSR Daily Activities 3D has provided evaluation for numerous methods in activity recognition via pose features, [29, 242, 243].

POETICON. The POETICON corpus [120] is a collection of scripted scenarios in which two individuals perform a daily living task such as cleaning the kitchen. The subjects are tracked using motion capture suits and recorded using 5 RGB camcorders. Certain tools and objects within the environment were also identified using marker based tracking. 4 pairs of actors learnt the associated script with 6 different high level scenarios, performing each activity in 3 repetitions per pair. Wallraven and Schultze [121] apply the dataset to identifying actions at differing levels of abstraction and granularity.

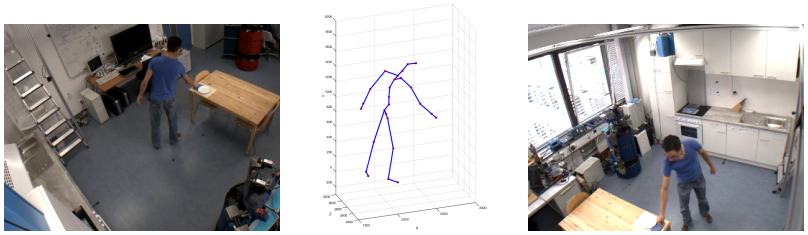


Figure 45: Example images for TUM Kitchen dataset - RGB, skeleton, alternate view RGB

TUM Kitchen. The Technische Universität München (TUM) Kitchen dataset [128, 129], Figure 45, provides a daily living set that describes the preparation of a table setting within a smart kitchen. The set makes use of video, marker-less skeletal tracking, RFID tagged objects and magnetic sensors to provide detailed information on the human action in the kitchen. The subjects collect items from cupboards and use them to set a place on the table. Actions were performed in one of two styles; natural or robotic, in which subjects could only carry one object at a time. There are also recordings of certain actions being repeatedly executed in order to train a classifier for recognition. The TUM Kitchen is designed to facilitate the learning of action events are differing levels of abstraction; recognizing not only the low level gestures and actions, but also the overall tasks completed. The TUM Kitchen has often been used for recognition of activities at various abstraction levels [129], and also for the detection, localization and segmentation problem [244, 245].

4. Proposed dataset

In the following section we draw upon the findings from the survey to present our own novel dataset for the recognition of complex conversational interactions between two individuals. We outline the necessity for the production of the set, the structure of the dataset and report on several previous publications that have utilized the dataset.

4.1. Requirement for the dataset

As can be seen from the previous sections, datasets that are able to capture human action using appearance based modalities, such as RGB videos, have developed from representing non-realistic emphasized actions to considering more complex interactions between individuals and their surrounding environment. The field has moved from actions which are easily distinguishable in the visual domain, e.g. 'waving' and 'jumping', to those of interactions, although still recognizable, e.g. 'hug' and 'kiss' [23, 246]. Due to the availability of these datasets many methods have been produced and evaluated for the purpose of action recognition and detection, including the use of SIFT [247], temporal Harris corner features [248] or STIPs [89].

Meanwhile the depth based methodology which has risen to prominence over the past decade has far fewer publicly available datasets which consider the problem of

person-person interactions, with most considering either emphasized actions or interactions. As such we believe that the publication of a dataset that represents highly complex person-person interactions is timely. We have chosen to capture conversational interactions between two individuals using the Kinect depth sensor, posing the challenge of recognizing subtle interaction classes.

The primitive action provided by many of the available datasets can be decomposed into a series of definable gestures and atomic poses. However we argue that real-world social interactions contain more complex and subtle class partitioning, being a product of multiple actions, semantics and the interplay between those involved. We therefore propose the problem of recognizing interactions in which the distinguishing features are containing within the temporal dynamics of the total event, such as that of a verbal interaction. We provide a dataset in which the interaction is labeled as a whole, rather than describing the event based on the primitive gestures within the scene. By providing such a dataset we hope to move the field towards the recognition of scenarios in which the defining descriptors are highly complex and context specific.

4.2. Apparatus setup

In this work, we choose seven conversational action categories and use a two-Kinect setup to capture 3D human pose during the interaction between two individuals. The collection environment consisted of a cleared space within a boardroom (Figure 46); in order to keep the dataset complex, no effort was made to homogenize the environment by use of any backdrops. Two Kinect sensors were located at opposite ends of the room, approximately two meters away from a marker on which a subject would be loosely located. Each person was recorded using a single Kinect Sensor at 30fps. The Kinect was offset to the front right of the subject in order to avoid occlusion from the opposing subject, which could occur if taking a frontal recording of the subject. Subjects were placed approximately one meter apart but not limited in their movement. Two PAL cameras (B cameras in Figure 46) were located to capture the full body of a single participant, with a third camera (M in Figure 46) located to capture the entire recording scene. These recordings are purely for the monitoring of the experiment and synchronization, thus are not provided within the dataset published in [1]. Cameras were also located to capture the face of each participant (F cameras in Figure 46), these provide the RGB recordings used to generate the gaze estimation provided. The recording devices were not located in the same place, and as such there is orientation variance between the depth maps and the RGB recordings.

4.3. Action descriptions

Participants were required to complete 7 different conversational tasks, outlined in Table 10. There was no time limitation on the execution of each task, and some tasks took naturally longer than others. Several tasks were given revealed to the participants before collection, to allow preparation; the actions that required preparation were describing work, story telling, debate, discussion and jokes (4.4 and 4.4). If the participants were given the problem or subjective question before the study then there may have been a reduction in interaction between the individuals.

Each task was performed and then there was a small break while the participants were reminded of the next task to carry out. The first task was to discuss an area of their

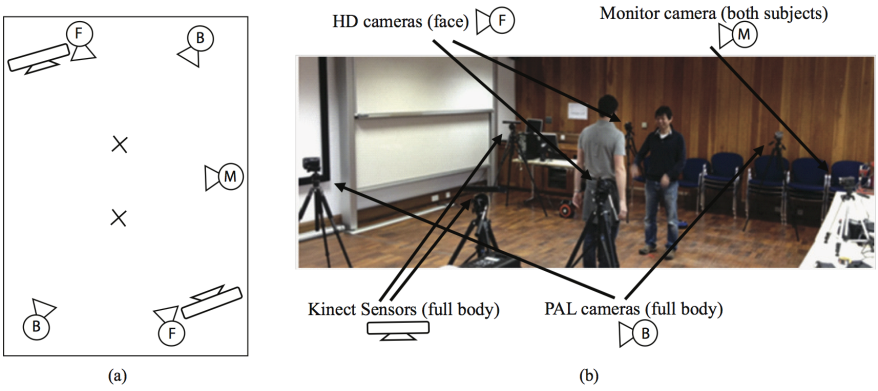


Figure 46: Layout of the CONVERSE data capture environment. a) Plan view of the capture environment b) Photo showing subjects within the cluttered environment.

current work. The second task was to prepare an interesting story to tell their partner, such as a holiday experience. The third task was to jointly find the answer to a problem. The fourth task was a debate, where the participants were asked to prepare arguments from opposing view points on an issue we gave to them. In the fifth task they were asked to discuss the issues surrounding a particular statement and come to agreement whether they believe the statement is true or not. The participants were asked to reach an agreement through discussion; hence, it is different to the debate task, which was based on conflicting views. The sixth task was to answer a subjective question, and the seventh task was to take it in turn telling jokes to one another.

4.4. Participants

16 subjects responded to a call for participants to take part in dataset collection and provided their consent for the collection. Participants were then organized into 8 pairs to record the person-person interaction during the following series of conversational styles. Interested individuals were asked to prepare for tasks 'Describing Work', 'Story Telling', 'Debate' and 'Joke' in advance, while the topics for 'Problem Solving', 'Discussion' and 'Subjective Question' were provided during collection. Participants were not subjected to time limitations or any execution styles.

4.5. Data provided

The main data in the collection is the skeletons extracted using the Microsoft Kinect SDK, providing the 20 tracked joints and the confidence of the tracking at each frame in the sequence. The raw depth and RGB recordings from the Kinect are also available alongside the RGB recordings from the separate camcorder. We also provide facial tracking features used for the tracking of gaze and facial dynamics which have been used for feature fusion in [249]. Despite the benefit that audio provides to action classification [30, 250–252], the audio has been stripped from all recordings due to the private natures of the conversations that occurred during the interactions. This allows

Figure 47: Example recordings from each of the 7 action classes, sampled at 2 second intervals and omitting the lower half of the body.

Figure 48: Example recordings and skeleton poses from each of the 7 action classes, sampled at 2 second intervals and omitting the lower half of the body.

Table 10: Description of each of the tasks given to the participants to perform. The rightmost column describes whether the participants were told about the task and asked to prepare before attending.

Task Name	Description	Prepared in advance
Describing Work	Each participant describes their current work or project to partner. The partner then repeats the description back, to confirm they had understood.	Yes
Story Telling	Participant were asked to think of an interesting story they could tell their partner.	Yes
Problem Solving	Participants were given the problem "Do candles burn in space and if so what shape and direction?", and asked to think of the solution of together.	No
Debate	Participants prepared arguments for a given point of view, pro or con, on the topic "Should University education be free?", and then debated this between them.	Yes
Discussion	Participants were asked to jointly discuss issues surrounding the statement "Social Networks have made the world a better place", and come to agreement whether they believe the statement is true or not.	No
Subjective Question	Participants responded to the subjective question "If you could be any animal, what animal and why?"	No
Telling jokes	Participants were asked to take it in turn telling three separate jokes.	Yes

the conversations to be natural, providing a more realistic representation of the scenarios than if each subject was given a script. Although this may be disappointing to those wishing to carry out audio-visual feature fusion, we believe that CONVERSE provides a more complex challenge to be solved when occluding the audio cues of conversation.

4.6. Results obtained on CONVERSE

To provide insight into the use of CONVERSE for interaction recognition we provide baseline results achieved using various state of the art methods for subject-specific classification, with results reported in Tables 11 and 12. To achieve this level of accuracy we followed the methods outlined in [249]; utilizing pose, face and head orientation features to provide a visual vocabulary of words and topics. Discriminative classifiers, SVM and Random Forest (RF), were trained to classify each CONVERSE task based on the discriminative power of the features. K-Nearest Neighbor (KNN) was selected as a baseline classification technique for comparison. First a Gaussian Mixture Model (GMM) was fitted to low level features (joint-joint/joint-plane distances and joint velocity) in order to obtain a vocabulary of 740 visual words consisting of the Gaussian components taken from 5 second clips, 370 words from facial features and 370 from pose features. Sequences were also sub-sampled into 20 second segments and Latent Dirichlet Allocation performed to obtain the 25 visual topics that made up each document. Both visual words and topics were used as temporal feature descriptors for each class. All sequences from the CONVERSE set were utilized, with 10 fold cross validation used to evaluate the performance. The RF classifier was produced using 100 trees with random sampling with replacement. The SVM was trained using a radial basis function kernel on the same training set.

Table 11: Classification results using visual words (%).

	Face&Pose			
	KNN	RF	SVM	SVM-R
Describing Work	81.2	90.6	88.4	100.0
Story Telling	59.7	51.0	70.6	80.2
Problem Solving	41.4	12.8	35.1	80.7
Debate	55.3	51.6	67.7	91.8
Discussion	50.0	62.7	69.5	61.1
Subjective Question	30.8	5.2	35.8	91.7
Jokes	36.3	14.2	47.7	80.0
Average	50.7	41.2	59.3	89.1

It was found that visual topics provide a generalization of the classes which benefit SVM and RF performance (Table 12), while KNN produced more accurate classification on data at the visual words level (Table 11). The importance of each feature was identified via novel use of particle swarm optimization (PSO) to generate a Ranked Feature SVM (SVM-R) classifier, reducing the dimensionality of the feature space and simultaneously performing optimal SVM model selection. The PSO method locates the optimal hyper-parameters that are used to subsequently train the SVM-R classifier by selecting towards correct identification of training samples, removal of redundant features, and the selection of compact feature vectors. This method significantly improved over the previous methods due to the selection of key partitioning features, increasing the accuracy on both visual word and topic feature sets. SVM-R optimization achieved 89.1% and 87.3% accuracy for word and topic respective levels of generalization due to its optimized feature set pruning. More detail regarding the use of the SVM-R classifier can be found in [249, 253].

Although these accuracy rates are relatively high, the results have been obtained on subject specific classification utilizing features extracted from long temporal segments of the observation. The main challenge we propose with CONVERSE is for the role of global recognition across multiple subjects for these complex interaction classes.

Table 12: Classification results using visual topics (%).

	Face&Pose			
	KNN	RF	SVM	SVM-R
Describing Work	63.5	91.7	76.4	100.0
Story Telling	35.1	73.2	68.3	80.2
Problem Solving	37.1	73.6	74.3	80.7
Debate	48.6	73.6	67.1	81.97
Discussion	38.4	78.7	63.5	61.11
Subjective Question	22.5	63.3	63.5	91.74
Jokes	27.5	70.3	66.3	80.0
Average	38.9	74.9	68.5	87.3

5. Conclusion

This paper presents the current state of the art in regards to the datasets that are available to the HAR community, highlighting the need for a dataset that presents subtle interactions between two individuals. The field has progressed over the previous decades, moving from the simplistic single action sequences towards a more natural representation of daily actions and interactions. We also provide clear definitions regarding the level of abstraction within the observations that are commonly encountered in the field, placing our proposed dataset within that of complex conversation interaction rich activities. By using pose based techniques we have shown that the recognition of top level action classes within the CONVERSE dataset is possible from using pose estimation output obtained from the Kinect sensor, [75–77, 249]. We have utilized current techniques, such as the Bag of Key Words, to describe the higher level event in terms of the composition of lower level action primitives. The full dataset is made publicly available for further research into the understanding of highly complex interactions at [1].

References

[1] Swansea University Computer Vision and Medical Image Analysis Group, "CONVERSE dataset," date accessed: 29/07/2015. [Online]. Available: http://csvision.swan.ac.uk/converse

[2] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[3] ——, "Visual motion perception," *Scientific American*, vol. 232, pp. 76–88, 1975.

[4] D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," in *Proc. of the Royal Society of London. Series B, Containing papers of a Biological character.*, vol. 200, no. 1140, 1978, pp. 269–94.

[5] R. Rashid, "Towards a system for the interpretation of moving light displays," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 2, no. 6, pp. 574–581, 1980.

[6] D. Hogg, "Model-based vision: a program to see a walking person," *Image and Vis. Computing*, vol. 1, pp. 5–20, 1983.

[7] H. Lee and Z. Chen, "Determination of 3D human body postures from a single view," *Comp. Vis., Graphics and Image Process.*, vol. 30, no. 2, pp. 148–168, 1984.

[8] Z. Chen and H. Lee, "Knowledge-guided visual perception of 3-D human gait from a single image sequence," *IEEE Trans. Syst., Man, and Cybern.*, vol. 22, no. 2, pp. 263–267, 1992.

[9] K. Rohr, "Towards model-based recognition of human movement in image sequences," *CVGIP: Image Understanding*, vol. 59, pp. 94–115, 1994.

[10] J. Aggarwal, Q. Cai, W. Liao, and B. Sabata, "Articulated and elastic non-rigid motion: A review," in *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994, pp. 2–14.

[11] L. Campbell and A. Bobick, "Recognition of human body motion using phase space constraints," in *Proc. Int. Conf. on Comp. Vis.*, 1995, pp. 624–630.

[12] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comp. Vis. Image Underst.*, vol. 73, no. 3, pp. 428–440, 1999.

[13] R. Polana and R. Nelson, "Low level recognition of human motion (or how to get your man without finding his body parts)," 1994.

[14] S. Osaka, "Recognition of human body motions by robots," in *Intelligent Robots and Syst.*, 1992, pp. 2139–2146.

[15] M. Rossi and A. Bozzoli, "Tracking and counting moving people," *IEEE Trans. Image Process.*, vol. 3, pp. 212–216, 1994.

[16] A. Azarbayejani and A. Pentland, "Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features," in *Proc. Int. Conf. Pat. Rec.*, vol. 3, 1996, pp. 627–632.

[17] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 19, no. 7, pp. 780–785, 1997.

[18] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 1992.

[19] O. Chomat and J. Crowley, "Probabilistic recognition of activity using local appearance," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 1999, pp. 0–5.

[20] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 23, no. 3, pp. 257–267, 2001.

[21] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. Int. Conf. on Comp. Vis.*, 2003, pp. 432–439 vol.1.

[22] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions : A local SVM approach," in *Pat. Rec.*, 2004, pp. 3–7.

[23] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65–72.

[24] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 29, no. 12, pp. 2247–53, 2007.

[25] W. Choi, S. Savarese, and S. Khuram, "What are they doing? : Collective Activity Classification Using Spatio-Temporal Relationship Among People," *Proc. Int. Conf. on Comp. Vis. Workshops*, vol. 24, no. October 2005, p. 2008, 2008.

[26] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comp. Vis. Image Underst.*, vol. 117, no. 6, pp. 633–659, 2013.

[27] A. Yao, J. Gall, G. Fanelli, and L. V. Gool, "Does human action recognition benefit from pose estimation?" in *Proc. British Conf. on Mach. Vis.*, 2011, pp. 67.1–67.11.

[28] L. Xia, C. Chen, and J. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 20–27.

[29] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2013, pp. 716–723.

[30] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive Multimodal Human Action Database," *Workshop on Applications of Computer Vision*, pp. 53–60, 2013.

[31] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, "Realistic Human Action Recognition With Multimodal Feature Selection and Fusion," *IEEE Trans. Syst., Man, and Cybern., Part A: Syst. and Humans*, vol. 43, no. 4, pp. 875–885, 2013.

[32] I. Lefter, G. J. Burghouts, and L. J. M. Rothkrantz, "An audio-visual dataset of human-human interactions in stressful situations," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 29–41, 2014.

[33] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comp. Vis. Image Underst.*, vol. 115, no. 2, pp. 224–241, 2011.

[34] M. Andersen, T. Jensen, P. Lisouski, A. Mortensen, M. Hansen, T. Gregersen, and P. Ahrendt, "Kinect depth sensor evaluation for computer vision applications," Aarhus University, Department of Engineering, Tech. Rep., 2012.

[35] K. Berger, S. Meister, R. Nair, and D. Kondermann, "A state of the art report on kinect sensor setups in computer vision," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, ser. Lecture Notes in Computer Science, M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, Eds. Springer Berlin Heidelberg, 2013, vol. 8200, pp. 257–272.

[36] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Transactions on Cybern.*, vol. 43, no. 5, pp. 1318–1334, 2013.

[37] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, D. Samaras, and S. Brook, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, 2012, pp. 28–35.

[38] L. Huynh, T. Ho, Q. Tran, T. B. Dinh, and T. Dinh, "Robust classification of human actions from 3D data," in *IEEE Int. Symp. on Signal Process. and Information Technology*, 2012, pp. 263–268.

[39] Y. Zhu, W. Chen, and G. Guo, "Fusing Spatiotemporal Features and Joints for 3D Action Recognition," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, 2013, pp. 486–491.

[40] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *IEEE Int. Workshop on CVPR for Human Communicative Behavior Analysis*, 2010.

[41] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 1290–1297.

[42] F. Zhou, F. D. Torre, and J. K. Hodgins, "Aligned cluster analysis for temporal segmentation of human motion," Carnegie Mellon University, Tech. Rep., 2008.

[43] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2013, pp. 915–922.

[44] F. Zhou, F. De la Torre, J. K. Hodgins, and F. D. la Torre, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 35, no. 3, pp. 582–596, 2013.

[45] A. A. Chaaraoui and F. Flórez-revuelta, "Adaptive human action recognition with an evolving bag of key poses," *Autonomous Mental Development*, vol. 6, no. 2, pp. 139–152, 2014.

[46] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, vol. 2, 2003, pp. 613–619.

[47] A. Chaaraoui, P. Climent-Prez, and F. Flrez-Revuelta, "An efficient approach for multi-view human action recognition based on bag-of-key-poses," in *Human Behavior Understanding*, ser. Lecture Notes in Computer Science, A. Salah, J. Ruiz-del Solar, . Merili, and P.-Y. Oudeyer, Eds., vol. 7559. Springer Berlin Heidelberg, 2012, pp. 29–40.

[48] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[49] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.

[50] H. Li and M. Greenspan, "Multi-scale gesture recognition from time-varying contours," in *Proc. Int. Conf. on Comp. Vis.*, vol. 1, 2005, pp. 236–243.

[51] Y. Chen, Q. Wu, and X. He, "Using dynamic programming to match human behavior sequences," *Control, Automation, Robotics and Vision*, pp. 17–20, 2008.

[52] T. Vajda, "Action recognition based on fast dynamic-time warping method," in *Int. Conf. on Intelligent Comp. Comm. and Process.*, 2009, pp. 127–131.

[53] Y. Shen and H. Foroosh, "View-invariant action recognition from point triplets," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 31, no. 10, pp. 1898–1905, 2009.

[54] S. Sempena, N. U. Maulidevi, and P. R. Aryan, "Human action recognition using dynamic time warping," in *Electrical Engineering and Informatics*, 2011.

[55] D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2008.

[56] T. Wang, S. Wang, and X. Ding, "Learning a similarity metric discriminatively for pose exemplar based action recognition," in *Int. Congress on Image and Signal Process.*, 2011, pp. 404–408.

[57] T. Hassner, "A critical review of action recognition benchmarks," *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, pp. 245–250, 2013.

[58] S. Stein and S. J. McKenna, "50 Salads dataset," date accessed: 29/07/2015. [Online]. Available: http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/

[59] ——, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Int. Joint Conf. on Pervasive and Ubiquitous Computing*, 2013.

[60] S. J. Blunsden and R. B. Fisher, "BEHAVE Interactions Test Case Scenarios," date accessed: 29/07/2015. [Online]. Available: http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/

[61] ——, "The BEHAVE video dataset : ground truthed video for multi-person behavior classification," *Annals of the BMVA*, vol. 2010, no. 4, pp. 1–11, 2010.

[62] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley Multimodal Human Action Database," date accessed: 29/07/2015. [Online]. Available: http://tele-immersion.citris-uc.org/berkeley_mhad

[63] Y. Kong, Y. Jia, and Y. Fu, "BIT-Interaction dataset," date accessed: 29/07/2015. [Online]. Available: https://sites.google.com/site/alexkongy/software

[64] ——, "Learning human interaction by interactive phrases," in *Proc. Euro. Conf. on Comp. Vis.*, vol. 7572, 2012, pp. 300–313.

[65] Cornell University, "Cornell Activity Datasets CAD-60, CAD-120," date accessed: 29/07/2015. [Online]. Available: http://pr.cs.cornell.edu/humanactivities/data.php

[66] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *Int. Conf. on Robotics and Automation*, 2012.

[67] H. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Research*, vol. 32, no. 8, pp. 951–970, 2013.

[68] Institute of Automation Chinese Academy of Sciences, "CASIA action database for recognition," date accessed: 29/07/2015. [Online]. Available: http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp

[69] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "View-independent behavior analysis." in *Proc. IEEE Int. Conf. Syst., Man and Cybern.*, vol. 39, no. 4, 2009, pp. 1028–35.

[70] R. Fisher, "CAVIAR test case scenarios," date accessed: 29/07/2015. [Online]. Available: http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

[71] ——, "The PETS04 surveillance ground-truth data sets," in *Performance Evaluation of Tracking and Surveillance*, 2004.

[72] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcadam, and J. Macey, "Quality of Life Technology Center. Grand Challenge Data Collection," date accessed: 29/07/2015. [Online]. Available: http://kitchen.cs.cmu.edu/

[73] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Tech. report CMU-RI-TR-08-22: Guide to the Carnegie Mellon University Multimodal Activity Database," Robotics Institute, Carnegie Mellon University, Tech. Rep., 2008.

[74] CMU Graphics Lab, "CMU Graphics Lab Motion Capture Database," date accessed: 29/07/2015. [Online]. Available: http://mocap.cs.cmu.edu/

[75] J. Deng, X. Xie, and B. Daubney, "A bag of words approach to 3D human pose interaction classification with random decision forests," in *Computational Visual Media Conference*, 2013.

[76] ——, "A bag of words approach to subject specific 3D human pose interaction classification with random decision forests," *Graphical Models*, 2013.

[77] J. Deng, X. Xie, B. Daubney, H. Fang, and P. W. Grant, "Recognizing conversational interaction based on 3D human pose," in *Advanced Concepts for Intelligent Vision Syst.*, 2013, pp. 138–149.

[78] I. Laptev and P. Pérez, "Drinking and Smoking action annotaion," date accessed: 29/07/2015. [Online]. Available: http://www.di.ens.fr/~laptev/download.html

[79] ——, "Retrieving actions in movies," in *Proc. Int. Conf. on Comp. Vis.*, 2007.

[80] Inria, "ETISEO: Video understanding Evaluation," date accessed: 29/07/2015. [Online]. Available: http://www-sop.inria.fr/orion/ETISEO/download.htm

[81] A. Nghiem and F. Bremond, "ETISEO, performance evaluation for video surveillance systems," in *Advanced Video and Signal-Based Surveillance*, 2007.

[82] V. Bloom, D. Makris, and V. Argyriou, "G3D gaming datasets," date accessed: 29/07/2015. [Online]. Available: http://dipersec.king.ac.uk/G3D/G3D.html

[83] ——, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, 2012, pp. 7–12.

[84] V. Bloom, V. Argyriou, and D. Makris, "G3Di gaming datasets," date accessed: 29/07/2015. [Online]. Available: http://dipersec.king.ac.uk/G3D/G3Di.html

[85] ——, "G3Di: A Gaming Interaction Dataset with a Real Time Detection and Evaluation Framework," in *Proc. Euro. Conf. on Comp. Vis.*, 2014.

[86] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large human motion database," date accessed: 29/07/2015. [Online]. Available: http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

[87] ——, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. on Comp. Vis.*, 2011, pp. 2556–2563.

[88] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld, "Learning human actions from movies," date accessed: 29/07/2015. [Online]. Available: http://www.di.ens.fr/~laptev/actions/

[89] I. Laptev and M. Marszalek, "Learning realistic human actions from movies," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2008.

[90] M. Marszaek, I. Laptev, and C. Schmid, "Human actions and scenes dataset," date accessed: 29/07/2015. [Online]. Available: http://www.di.ens.fr/~laptev/actions/hollywood2/

[91] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2009, pp. 2929 – 2936.

[92] S. Hadfield, "Hollywood3D," date accessed: 29/07/2015. [Online]. Available: http://cvssp.org/Hollywood3D/

[93] S. Hadfield and R. Bowden, "Hollywood 3D: Recognizing actions in 3D natural scenes," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2013, pp. 3398–3405.

[94] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva dataset," date accessed: 29/07/2015. [Online]. Available: http://humaneva.is.tue.mpg.de/

[95] L. Sigal, A. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comp. Vis.*, vol. 87, pp. 4–27, 2010.

[96] D. Weinland, R. Ronfard, and E. Boyer, "INRIA Xmas Motion Acquisition Sequences," date accessed: 29/07/2015. [Online]. Available: http://4drepository.inrialpes.fr/public/viewgroup/6

[97] ——, "Free viewpoint action recognition using motion history volumes," *Comp. Vis. Image Underst.*, vol. 104, no. 2-3, pp. 249–257, 2006.

[98] M. S. Ryoo and L. Matthies, "JPL First-Person Interaction dataset," date accessed: 29/07/2015. [Online]. Available: http://michaelryoo.com/jpl-interaction.html

[99] ——, "First-Person activity recognition: What are they doing to me?" in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2013.

[100] "K3HI dataset," date accessed: 05/06/2014. [Online]. Available: http://www.lmars.whu.edu.cn/profweb/zhuxinyan/DataSetPublish/dataset.html

[101] T. Hu, X. Zhu, W. Guo, and K. Su, "Efficient interaction recognition through positive action representation," *Mathematical Problems in Engineering*, vol. 2013, pp. 1–11, 2013.

[102] I. Laptev and T. Lindeberg, "Recognition of human actions," date accessed: 29/07/2015. [Online]. Available: http://www.nada.kth.se/cvap/actions/

[103] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur, "The LIRIS human activities dataset," date accessed: 29/07/2015. [Online]. Available: http://liris.cnrs.fr/voir/activities-dataset/

[104] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur, "Evaluation of video activity localizations integrating quality and quantity measurements," *Comp. Vis. Image Underst.*, vol. 127, pp. 14 – 30, 2014.

[105] M. P. I. for Informatics, "MPI08 dataset," date accessed: 29/07/2015. [Online]. Available: http://www.tnt.uni-hannover.de/project/MPI08_Database/

[106] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn, "Multisensor-fusion for 3D full-body human motion capture," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2010.

[107] A. Baak, T. Helten, M. Müller, G. Pons-Moll, B. Rosenhahn, and H.-P. Seidel, "Analyzing and evaluating markerless motion tracking using inertial sensors," in *Proc. Euro. Conf. on Comp. Vis.*, 2010.

[108] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "MPII cooking activities dataset," date accessed: 29/07/2015. [Online]. Available: https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-activities-dataset/

[109] M. Rohrbach and S. Amin, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012.

[110] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "MPII cooking composite activities," date accessed: 29/07/2015. [Online]. Available: https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-composite-activities/

[111] ——, "Script data for attribute-based recognition of composite activities," in *Proc. Euro. Conf. on Comp. Vis.*, 2012.

[112] Microsoft Research, "MSR action recognition datasets and codes," date accessed: 29/07/2015. [Online]. Available: http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/

[113] J. Yuan, "Discriminative subvolume search for efficient action detection," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2009, pp. 2442–2449.

[114] J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection." *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 33, no. 10, pp. 1728–1743, 2011.

[115] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Proc. Euro. Signal Process. Conf.*, 2012, pp. 1975–1979.

[116] S. Singh, S. Velastin, and H. Ragheb, "MuHAVi: Multicamera Human Action Video dataset," date accessed: 29/07/2015. [Online]. Available: dipersec.king.ac.uk/MuHAVi-MAS/

[117] S. Singh, S. A. Velastin, and H. Ragheb, "MuHAVi: A Multicamera Human Action Video dataset for the evaluation of action recognition methods," in *Workshop on activity Monitoring by Multi-camera Surveillance Syst.*, 2010, pp. 48—-55.

[118] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Olympic Sports dataset," date accessed: 29/07/2015. [Online]. Available: http://vision.stanford.edu/Datasets/OlympicSports/

[119] ——, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. Euro. Conf. on Comp. Vis.*, 2010, pp. 392–405.

[120] C. Wallraven, M. Schultze, B. Mohler, A. Vatakis, and K. Pastra, "POETICON Corpus," date accessed: 11/09/2014. [Online]. Available: http://poeticoncorpus.kyb.mpg.de

[121] C. Wallraven and M. Schultze, "The POETICON enacted scenario corpus: A tool for human and computational experiments on action understanding," in *Proc. IEEE Conf. Automatic Face and Gesture Rec. Workshops*, 2011.

[122] R. Messing, C. Pal, and H. Kautz, "University of Rochester Activities of Daily Living dataset," date accessed: 29/07/2015. [Online]. Available: http://www.cs.rochester.edu/~rmessing/uradl/

[123] ——, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. Int. Conf. on Comp. Vis.*, 2009.

[124] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," date accessed: 29/07/2015. [Online]. Available: http://www3.cs.stonybrook.edu/~kyun/research/kinect_interaction/

[125] ——, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, 2012.

[126] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei, "Stanford 40 Actions dataset," date accessed: 29/07/2015. [Online]. Available: http://vision.stanford.edu/Datasets/40actions.html

[127] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. Int. Conf. on Comp. Vis.*, 2011, pp. 1331–1338.

[128] M. Tenorth, J. Bandouch, and M. Beetz, "TUM Kitchen dataset," date accessed: 29/07/2015. [Online]. Available: https://ias.cs.tum.edu/software/kitchen-activity-data

[129] ——, "The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition," *Proc. Int. Conf. on Comp. Vis. Workshops*, pp. 1089–1096, 2009.

[130] K. Soomro, A. R. Zamir, and M. Shah, "UCF101 action recognition dataset," date accessed: 29/07/2015. [Online]. Available: http://crcv.ucf.edu/data/UCF101.php

[131] ——, "CRCV-TR-12-01: UCF101 : A dataset of 101 human actions classes from videos in the wild," University of Central Florida, Center for Research in Computer Vision, Tech. Rep., 2012.

[132] J. Liu, J. Luo, and M. Shah, "UCF YouTube action dataset," date accessed: 29/07/2015. [Online]. Available: http://crcv.ucf.edu/data/UCF_YouTube_Action.php

[133] J. Liu, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2009, pp. 1996–2003.

[134] K. K. Reddy and M. Shah, "UCF50 action recognition dataset," date accessed: 29/07/2015. [Online]. Available: http://crcv.ucf.edu/data/UCF50.php

[135] ——, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2012.

[136] M. D. Rodriguez, J. Ahmed, and M. Shah, "UCF Sports action dataset," date accessed: 29/07/2015. [Online]. Available: http://crcv.ucf.edu/data/UCF_Sports_Action.php

[137] ——, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2008, pp. 1–8.

[138] N. van der Aa, X. Luo, G. Giezeman, R. Tan, and R. Veltkamp, "Utrecht Multi-Person Motion Benchmark," date accessed: 29/07/2015. [Online]. Available: http://www.projects.science.uu.nl/umpm/

[139] ——, "Utrecht multi-person motion benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction," in *Proc. Workshop on Human Interaction in Comp. Vis.*, 2011.

[140] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities," 2010, date accessed: 29/07/2015. [Online]. Available: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

[141] ——, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. Int. Conf. on Comp. Vis.*, 2009, pp. 1593–1600.

[142] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "ViHASi: Virtual Human Action Silhouette data for the evaluation of silhouette-based action recognition methods," date accessed: 29/07/2015. [Online]. Available: http://dipersec.king.ac.uk/VIHASI/

[143] H. Ragheb and S. Velastin, "ViHASi: Virtual Human Action Silhouette data for the performance evaluation of silhouette-based action recognition methods," in *Int. Conf. on Distributed Smart Cameras*, 2008, pp. 1–10.

[144] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "VIRAT video dataset," date accessed: 29/07/2015. [Online]. Available: http://www.viratdata.org/

[145] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, no. 2, 2011.

[146] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," date accessed: 29/07/2015. [Online]. Available: http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

[147] ——, "Actions as space-time shapes," in *Proc. Int. Conf. on Comp. Vis.*, 2005, pp. 1395–1402.

[148] V. Kulathumani, "WVU Multi-View action recognition dataset," date accessed: 29/07/2015. [Online]. Available: http://csee.wvu.edu/~vkkulathumani/wvu-action.html

[149] S. Ramagiri, R. Kavi, and V. Kulathumani, "Real-time multi-view human action recognition using a wireless camera network," in *Int. Conf. on Distributed Smart Cameras*, 2011, pp. 1–6.

[150] R. Kavi and V. Kulathumani, "Real-time recognition of action sequences using a distributed video sensor network," *J. Sensor and Actuator Networks*, vol. 2, no. 3, pp. 486–508, 2013.

[151] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in *IEEE Int. Symp. on Robot and Human Interactive Communication*, 2012, pp. 509–514.

[152] M. Ermes, J. Parkka, J. Mantyjarvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Trans. on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[153] C. McCall, K. K. Reddy, and M. Shah, "Macro-class selection for hierarchical K-NN classification of inertial sensor data," in *Pervasive and Embedded Computing and Communication Syst.*, 2012.

[154] J. Ferryman, "PETS," in *Performance Evaluation of Tracking and Surveillance*, 2009.

[155] M. Rohrbach and S. Amin, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, no. 6, 2012.

[156] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. Int. Conf. on Comp. Vis.*, 2009.

[157] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. Int. Conf. on Comp. Vis.*, 2011, pp. 415–422.

[158] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, and Z. Tu, "Exemplar-based human action pose correction and tagging," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 1784–1791.

[159] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The Vitruvian Manifold : Inferring Dense Correspondences for One-Shot Human Pose Estimation," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 103–110.

[160] A. Fathi, "Social interactions: A first-person perspective," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, ser. CVPR '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1226–1233.

[161] S. Ryoo, M, J. Fuchs, T, L. Xia, K. Aggarwal, J, and L. Matthies, "Robot-Centric Activity Prediction from First-Person Videos: What Will They Do to Me?" in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2015, pp. 295–302.

[162] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices," *Expert Systems with Applications*, vol. 41, no. 3, pp. 786–794, 2014.

[163] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 35, no. 4, pp. 835–848, 2013.

[164] H. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 33, no. 5, pp. 867–882, 2011.

[165] B. Solmaz, S. M. Assari, and M. Shah, "Classifying web videos using a global video descriptor," *Machine Vision and Applications*, vol. 24, no. 7, pp. 1473–1485, 2012.

[166] O. Kliper-gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion Interchange Patterns for Action Recognition in Unconstrained Videos," in *Proc. Euro. Conf. on Comp. Vis.*, 2012, pp. 256–269.

[167] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 35, no. 11, pp. 2782–95, 2013.

[168] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. Int. Conf. on Comp. Vis.*, 2007, pp. 1–7.

[169] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2011, pp. 489–496.

[170] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using Hankelets," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 1362–1369.

[171] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *Proc. Int. Conf. on Comp. Vis.*, 2013, pp. 3176–3183.

[172] I. Laptev and T. Lindeberg, "Velocity adaptation of space-time interest points," in *Proc. Int. Conf. Pat. Rec.*, 2004, pp. 52–56 Vol.1.

[173] M. Bregonzio, G. S, and X. T, "Recognising action as clouds of space-time interest points," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2009, pp. 1948–1955.

[174] Z. Jiang, Z. Lin, and L. S. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees." *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 34, no. 3, pp. 533–47, 2012.

[175] T. Wang, S. Wang, and X. Ding, "Detecting human action as the spatio-temporal tube of maximum mutual information," *IEEE Trans. on Circuits and Syst. for Video Technology*, vol. 24, no. 2, pp. 277–290, 2014.

[176] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *Comp. Vis. Image Underst.*, vol. 116, no. 3, pp. 396–410, 2012.

[177] K. G. Derpanis and M. Sizintsev, "Action spotting and recognition based on a spatiotemporal orientation analysis," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 35, no. 3, pp. 527–540, 2013.

[178] G. Yu, N. Goussies, J. Yuan, and Z. Liu, "Fast action Detection via discriminative Random Forest voting and top-K subvolume search," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 507–517, 2011.

[179] M. Javan Roshtkhari and M. D. Levine, "Human activity recognition in videos using a single example," *Image and Vis. Computing*, vol. 31, no. 11, pp. 864–876, 2013.

[180] Stanford Vision Lab, "Stanford 40 Actions: A dataset for understanding human actions in still images," date accessed: 29/07/2015. [Online]. Available: http://vision.stanford.edu/Datasets/40actions.html

[181] F. Sener, C. Bas, and N. Ikizler-cinbis, "On Recognizing Actions in Still Images via Multiple Features," in *Proc. Euro. Conf. on Comp. Vis.*, 2012, pp. 263–272.

[182] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg, "Coloring Action Recognition in Still Images," *Int. J. Comp. Vis.*, vol. 105, no. 3, pp. 205–221, 2013.

[183] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2014, pp. 1725–1732.

[184] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 1234–1241.

[185] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," *IEEE Workshop on Applications of Comp. Vis.*, pp. 103–110, 2013.

[186] M. J. Marin-Jiménez, N. P. de la Blanca, and M. A. Mendoza, "RBM-based Silhouette Encoding for Human Action Modelling," *Proc. Int. Conf. Pat. Rec.*, no. 1, pp. 979–982, 2010.

[187] F. Martínez-Contreras, C. Orrite-Uruñuela, E. Herrero-Jaraba, H. Ragheb, and S. A. Velastin, "Recognizing Human Actions Using Silhouette-based HMM," in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, 2009, pp. 43–48.

[188] C.-S. Lee, Y. M. Lui, and S. Y. Chun, "Human action silhouette recognition based on tensor analysis using synthetic silhouette data," in *Proc. Int. Conf. on Comp. Vis. Workshops*, 2011, pp. 1318–1323.

[189] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao, "A unified framework for locating and recognizing human actions," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2011, pp. 25–32.

[190] I. Ar and Y. Akgul, "Action recognition using random forest prediction with combined pose-based and motion-based features," in *IEEE Int. Conf. on Electrical and Electronics Engineering*, 2013, pp. 315–319.

[191] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories," in *Proc. Int. Conf. on Comp. Vis.*, 2011, pp. 1419–1426.

[192] W. Bian, D. Tao, and Y. Rui, "Cross-domain human action recognition." *IEEE Trans. Syst., Man, and Cybern., Part B: Cybern.*, vol. 42, no. 2, pp. 298–307, 2012.

[193] E. A. Mosabbeb, K. Raahemifar, and M. Fathy, "Multi-view human activity recognition in distributed camera sensor networks." *Sensors*, vol. 13, no. 7, pp. 8750–8770, 2013.

[194] S. J. Blunsden and R. B. Fisher, "BEHAVE Optical Flow Data," date accessed: 29/07/2015. [Online]. Available: http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/CROWDS/index.html

[195] E. Andrade, S. Blunsden, and R. Fisher, "Modelling crowd scenes for event detection," in *Proc. Int. Conf. Pat. Rec.*, 2006, pp. 175–178.

[196] E. Andrade, R. Fisher, and S. Blunsden, "Detection of emergency events in crowded scenes," in *IEEE Int. Symp. on Imaging for Crime Detection and Prevention*, 2006, pp. 528–533.

[197] Y. Yin, G. Yang, J. Xu, and H. Man, "Small group human activity recognition," in *IEEE Int. Conf. Image Process*, 2012, pp. 2709–2712.

[198] D. Münch, S. Becker, W. Hübner, and M. Arens, "Towards a real-time situational awareness system for surveillance applications in unconstrained environments," *Future Security*, vol. 318, pp. 517–521, 2012.

[199] M. Elhamod and M. D. Levine, "Real-time semantics-based detection of suspicious activities in public spaces," in *Conf. on Comp. and Robot Vis.*, 2012, pp. 268–275.

[200] M. Elhamod and M. Levine, "Automated real-time detection of potentially suspicious behavior in public transport areas," *IEEE Trans. on Intelligent Transportation Syst.*, vol. 14, no. 2, pp. 688–699, 2013.

[201] J. Nascimento, M. Figueiredo, and J. Marques, "Segmentation and classification of human activities," in *Int. Workshop on Human Activity Rec. and Modelling*, 2005.

[202] ——, "Recognition of human activities using space dependent switched dynamical models," *IEEE Int. Conf. on Image Process.*, 2005.

[203] A. Fernández-Caballero, J. Castillo, and J. Rodríguez-Sánchez, "A proposal for local and global human activities identification," *Lecture Notes in Computer Science*, vol. 6169, pp. 78–87, 2010.

[204] S. Denman, C. Fookes, S. Sridharan, and V. Chandran, "Object tracking using multiple motion modalities," *International Conference on Signal Processing and Communication Systems*, pp. 1–10, 2008.

[205] A.-L. Ellis and J. Ferryman, "Benchmark datasets for detection and tracking," in *Visual Analysis of Humans*, 2011, pp. 109–128.

[206] R. Romdhane, C. F. Crispim, F. Bremond, and M. Thonnat, "Activity recognition and uncertain knowledge in video scenes," *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pp. 377–382, 2013.

[207] S. Narayan and M. S. Kankanhalli, "Action and Interaction Recognition in First-person videos," *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, pp. 526 – 532, 2014.

[208] C. Tan, H. Goh, and V. Chandrasekhar, "Understanding the Nature of First-Person Videos: Characterization and Classification using Low-Level Features," *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, pp. 535–542, 2014.

[209] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "An Overview of First Person Vision and Egocentric Video Analysis for Personal Mobile Wearable Devices," *Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 744–760, 2014.

[210] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Proc. Int. Conf. on Comp. Vis.*, 2011, pp. 1036–1043.

[211] W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," in *Proc. Int. Conf. on Comp. Vis.*, 2011, pp. 778–785.

[212] Y. Zhu, N. M. Nayak, and A. Roy-Chowdhury, "The role of spatial context in activity recognition," in *Indian Conference on Computer Vision, Graphics and Image Processing*, 2012, pp. 31:1–31:6.

[213] X. Wang and Q. Ji, "A hierarchical context model for event recognition in surveillance video," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2014, pp. 2561–2568.

[214] Y. S. Sefidgar, A. Vahdat, S. Se, and G. Mori, "Discriminative key-component models for interaction detection and recognition," *Comp. Vis. Image Underst.*, vol. 135, pp. 16 – 30, 2015.

[215] N. M. Nayak, Y. Zhu, and A. K. R. Chowdhury, "Hierarchical Graphical Models for Simultaneous Tracking and Recognition in Wide-Area Scenes," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2025–2036, 2015.

[216] K. Huang and S. Wang, "Human behavior analysis based on a new motion descriptor," *IEEE Trans. on Circuits and Syst. for Video Technology*, vol. 19, no. 12, pp. 1830–1840, 2009.

[217] S. Wang, K. Huang, and T. Tan, "A compact optical flow based motion representation for real-time action recognition in surveillance scenes," in *Int. Conf. on Image Process.*, 2009, pp. 1121–1124.

[218] K. Huang, Y. Zhang, and T. Tan, "A discriminative model of motion and cross ratio for view-invariant action recognition," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2187–2197, 2012.

[219] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," in *Proc. Euro. Conf. on Comp. Vis.*, vol. 6311, 2010, pp. 577–590.

[220] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2011.

[221] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *Proc. Euro. Conf. on Comp. Vis.*, 2010, pp. 536–548.

[222] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," in *Proc. Euro. Conf. on Comp. Vis.*, 2010, pp. 508–521.

[223] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving Human Action Recognition Using Fusion of Depth Camera and Inertial Sensors," *IEEE Trans. on Human-Machine Syst.*, vol. 45, no. 1, pp. 51–61, 2015.

[224] S. Vantigodi and R. Babu, "Real-time Human Action Recognition From Motion Capture Data," *Nat. Conf. on Comp. Vis., Pat. Rec., Image Process. and Graphics*, pp. 1–4, 2013.

[225] M. S. Cheema, A. Eweiwi, and C. Bauckhage, "Human activity recognition by separating style and content," *Pattern Recognition Letters*, vol. 50, pp. 130–138, 2014.

[226] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition," *J. Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014.

[227] S. Nayak, S. Sarkar, and B. Loeding, "Distribution-based dimensionality reduction applied to articulated motion recognition." *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 31, no. 5, pp. 795–810, 2009.

[228] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Human action recognition in stereoscopic videos based on bag of features and disparity pyramids," in *European Signal Processing Conference*, 2014, pp. 1317–1321.

[229] I. Mademlis, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Stereoscopic video description for human action recognition," in *Symposium on Computational Intelligence for Multimedia, Signal and Vis. Process.*, 2014, pp. 1–6.

[230] S. Hadfield, K. Lebeda, and R. Bowden, "Natural action recognition using invariant 3d motion encoding," in *Proc. Euro. Conf. on Comp. Vis.*, 2014, vol. 8690, pp. 758–771.

[231] L. Sigal and M. J. Black, "Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion," Brown University, Tech. Rep., 2006.

[232] S.-R. Ke, J.-N. Hwang, K.-M. Lan, and S.-Z. Wang, "View-invariant 3d human body pose reconstruction using a monocular video camera," in *ACM/IEEE Int. Conf. on Distributed Smart Cameras*, 2011, pp. 1–6.

[233] B. Daubney and X. Xie, "Tracking 3d human pose with large root node uncertainty," *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2011.

[234] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2012, pp. 14–19.

[235] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Proc. Euro. Conf. on Comp. Vis.* Springer Berlin Heidelberg, 2012, pp. 872–885.

[236] R. Slama, H. Wannous, and M. Daoudi, "Grassmannian representation of motion depth for 3d human gesture and action recognition," in *Proc. Int. Conf. Pat. Rec.*, 2014, pp. 3499–3504.

[237] Y. Yang, A. Guha, C. Fermuller, and Y. Aloimonos, "Manipulation action tree bank: A knowledge resource for humanoids," in *IEEE-RAS Int. Conf. on Humanoid Robots*, 2014, pp. 987–992.

[238] S. Stein and S. J. McKenna, "User-adaptive models for recognizing food preparation activities," in *Int. Workshop on Multimedia for Cooking and Eating Activities*, 2013, pp. 39–44.

[239] N. van der Aa, X. Luo, G. Giezeman, R. Tan, and R. Veltkamp, "UMPM benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction," in *Proc. Int. Conf. on Comp. Vis. Workshops*, 2011, pp. 1264–1269.

[240] E. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," *IEEE Comp. Soc. Conf. on Comp. Vis. and Pat. Rec. Workshops*, pp. 17–24, 2009.

[241] E. Borzeshi, O. Perez Concha, R. Xu, and M. Piccardi, "Joint action segmentation and classification by an extended hidden markov model," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1207–1210, 2013.

[242] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec. Workshops*, 2013, pp. 479–485.

[243] A.-A. Liu, Y.-T. Su, P.-P. Jia, Z. Gao, T. Hao, and Z.-X. Yang, "Multipe/single-view human action recognition via part-induced multitask structural learning," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1194–1208, 2015.

[244] S. H. Lee, I. H. Suh, S. Calinon, and R. Johansson, "Learning basis skills by autonomous segmentation of humanoid motion trajectories," in *IEEE-RAS Int. Conf. on Humanoid Robots*, 2012, pp. 112–119.

[245] S. Kwak, B. Han, and J. H. Han, "On-line video event detection by constraint flow," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 36, no. 6, pp. 1174–1186, 2014.

[246] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *Proc. Int. Conf. on Comp. Vis.*, 2009.

[247] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. Int. Conf. on Multimedia*. New York, NY, USA: ACM, 2007, pp. 357–360.

[248] I. Laptev, "On space-time interest points," *Int. J. Comp. Vis.*, vol. 64, no. 2-3, pp. 107–123, 2005.

[249] J. Deng, X. Xie, and S. Zhou, "Conversational Interaction Recognition based on Bodily and Facial Movement," in *Int. Conf. on Image Analysis and Rec.*, 2014.

[250] B. Peskin, L. Gillick, Y. Ito, S. Lowe, R. Roth, F. Scattone, J. Baker, J. Baker, J. Bridle, M. Hunt, and J. Orloff, "Topic and speaker identification via large vocabulary continuous speaker recognition," in *ARPA Human Language Technology*, 1993, pp. 119–124.

[251] I. R. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Dialogue speech recognition by combining hierarchical topic classification and language model switching," *IEICE Trans. on Information and Systems*, vol. 88, no. 3, pp. 446–454, 2005.

[252] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162 – 1181, 2006.

[253] S.-M. Zhou, R. Lyons, O. Bodger, J. Demmler, and M. Atkinson, "Svm with entropy regularization and particle swarm optimization for identifying children's health and socioeconomic determinants of education attainments using linked datasets," in *Int. Joint Conf. Neural Networks*, 2010, pp. 1–8.