Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Comprehensive Study of Different Dataset for Human Action and Activity to Understanding Human Behavior using Computer Vision Technique

[1]Rakesh Yashwant Gedam, [2]Rakesh Jagdish Ramteke

[1]Reaserch Scholar, School of Computer Sciences, North Maharashtra University, Jalgaon, India.
[2] Professor, School of Computer Sciences, North Maharashtra University, Jalgaon, India.

Email-Id :[1] mr.rakeshgedam1@gmail.com, [2] rakeshj.ramteke@gmail.com

Abstract :
The terms “action” and “activity” are frequently used inter-changeably in the vision literature
The terms “action” and “activity” are frequently used inter-changeably in the vision literature
The terms “action” and “activity” are frequently used inter-changeably in the vision literature
In the paper, understanding human behavior with the help of a machine on the human actions and activity attracts many researchers due to its wide range of application. Human behavior Analysis is dependent on both the temporal and spatial context. The datasets themselves specifically related to visual modalities, and their evolution towards modern datasets. This paper gives great detail of evolution of Datasets for Human Action Activity, Behavior Understanding from last two decade along with various evaluation parameter. Finally classification five type’s i.e atomic action level, behavior level, interaction level and group activities level to assist researcher for select as per requirement.
Keyword: Human Behavior Understand, Human Activity And Action Reorganization, Computer Vision, Dataset for activity and action.

1. Introduction :

To identifying actions or activities performed by a single person or a group of people or any human being under controlled or uncontrolled environment term as Human Activity Recognition (HAR) system. Effective and efficient HAR system play vital role in medical science, cognitive science , behavior science ,computer science & computer vision, pattern recognition, human computer interface, security & safety, intelligent surveillance, and endless area due computer and sensor are available everywhere Basically two approaches was focused handcraft traditional method and machine learning i.e deep learning. Some claims that machine learning approach have the advantage in adaptability, accuracy and more recently, greater speeds over traditional approaches [1,2]. There are many reviews who focus on the subject of human activity recognition on breakthrough algorithms rather we focus on the datasets themselves, specifically datasets related to visual modalities, and their evolution towards modern datasets[1,5]. Definition of action n varies according to the goals of researcher and their objective. According Herath el etc an “Action is the most elementary human-surrounding interaction with a meaning”[3]. Few distinguishes between activity and action according to them action sample is simple in lasting for few sec whereas activity is a complex sequential action carried for more time interval i.e more than 10 sec, or minute or lasting for hour with combination of interrelated simpler or primitive actions describe the interactions between people, subjects and their surroundings, usually having a beginning and an end [6,7,8]. According to Zelnik-Manor an event is an action anchored to a specific time and space, as well as well-defined rules and clear context [9].material, quality of video, quantity of labels, and complexity of annotation and generality of con tent. If the datasets able to test the capability of recognition systems for handling contextual cues, partial occlusion, intraclass variability, varying size .To fulfill our

requirement we try to analysis different available dataset and term them modern dataset for action /activity recognition which satisfies the following criteria considering that there *is no universal dataset*. I) Unconstrained inputs as close to real-world examples as possible. II) Large size, preferably high number of classes. III) Exhaustively annotated. Datasets Characteristics consider as *the Action, Group Activity , Behavior ,Human-object interaction, Human-human interaction* with consideration

a) One Person only: Activities such as *Walking* and *Running* involving a single person and usually atomic in nature and treated as 'low-level' for Action and Behavior dataset
b) Object Person Object: The classes in the form of ('verb+object') such as *Ride Bike* and *Spray Water*. 'object manipulation'-based activities where the object is essential in defining the activity for Human object Interaction and Group Activity [16,17,18]
c) Person to Person: complex activities which combine or involving interactions between people such as *Walk Together, Meet* and *handshaking.* Datasets such as BEHAVE [19]consideration of realistic Person–Person for Human- Human Interaction

II. Literature Review:

HRA dependent on both the temporal and spatial context which emerged as a distinct problem from static object recognition or classification[1,2]. Human action and activity Recognition (HAR). HAR datasets usually preserve the temporal dimension, unlike simpler image classification tasks, although there exist large datasets for image-based activity classification attract many researcher from last two decades hence continuous evolved computer vision and machine learning technique work has been carried out. Early datasets was carefully design under controlled conditions within labs or selective condition means very little variation in the content i.e number of actions, number of actors, lighting, occlusion, viewpoints, modalities and size of dataset. Further for more complex models and challenging datasets allow for the evaluation of the unconstrained action recognition from "real-world" videos[3]. Generally datasets consider activities performed during daily life or others may vary from specific domain focus. Some focus on very important sub problem in HAR i.e is sports-related and gaming activities. The Author focus on UCF sports [4], Olympic Sports [5], Sports-1M [6], Volleyball [7], and SoccerNet [8] are all dedicated towards sports activi- ties and they are important subclasses in datasets such as UCF101. Some author choose a classes for free hand gamming action G3D [9], G3Di [10] and MSR action [11].Few focus on Kitchen- based actions are also a popular dataset choice [12]. Application in autonomous driving , READ dataset [13]. For surveillance tasks such as Tailgating, Fighting, Shop Entering, and ShopExiting. such as CAVIAR [14] and ETISEO [15]

Data Captured Mechanism:

HRA is totally dependent on both the temporal and spatial context [1,2], which emerged as a distinct problem from static object recognition or classification (Fig. 2,3).The natural representation of datasets is in the form of clips of 2D images, and most datasets use this format exten sively. However, after the introduction of low-cost 3D sensors such as Microsoft Kinect, there has been great interest in using depth information [11]. A detailed description of RGB-D datasets can be found in [16]. Another totally different class of datasets is recorded by using non-visual sensor such as accelerometers and gyroscopes along with ambient and stationary sensors i.e RADAR. This class has a great diversity of datasets, such as the opportunity [17]dataset. Sensor-based datasets have been recently reviewed in [18]. Generally HAR datasets usually consisting the temporal dimension, simpler image classification tasks, although there exist large datasets for image-based activity classification [35,36]

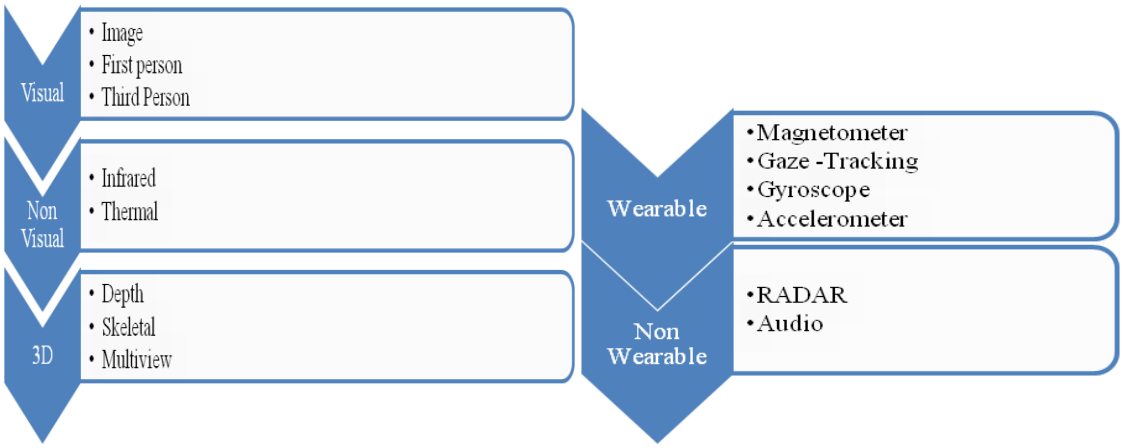

Fig. 1. Data Capture classification based on data Captured Methodology for Behavior Understanding

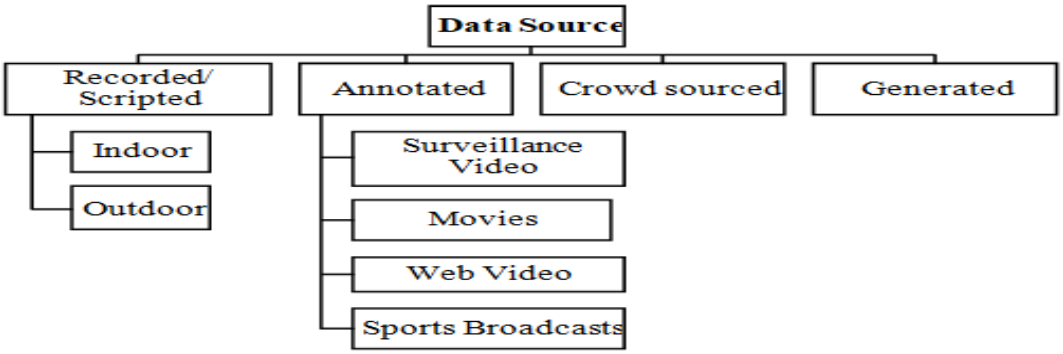

Fig. 2. Data Source for Human Action and Activity for Behavior Understanding

III. Evolution of Datasets for Human Action ,Activity ,Behavior Understanding :

We divide datasets for action and activity recognition for behavior understanding using Computer vision or Machine with 3 basic type 1) Early dataset 2)Video dataset 3)Image dataset based on characteristic of dataset i.e Action ,Behavior, Human-object -interaction, Human-human interaction and Group Activities.

Type I : Early Dataset :

Early datasets for action and activity understanding was very simple and were completely scripted datasets and filmed in very ideal conditions as per requirement . The performance on these datasets good but not work upto the mark over real data, especially as modern algorithms The few of them description of these early datasets listed below

A): Weizmann:

Weizmann Event dataset was introduced in 2001 which consist 4 simple class specifically focused on even and further redefined Weizmann institute of Science in 2005. 10 simple actions with static background, i.e., walk, run, skip, jack, jump forward or jump, jump in place or pjump, gallop-sideways or side, bend, wave1, and wave2 was present . It is considered as an honest benchmark for evaluation of algorithms proposed for recognition of straightforward actions**. [39]**

Fig 3. Sample action with Weizmann Dataset

B) KTH Human Action Dataset :

The KTH dataset was created by the Royal Institute of Technology, Sweden in2004. KTH having six types of human actions(walking, jogging, running, boxing, hand clapping and hand waving) performed by 25 actors with 4 different scenarios. Thus, it contains 25 × 6 × 4 = 600 video sequences. These videos were recorded with static camera and background; therefore, this dataset is additionally considered relatively simple for evaluation of act recognition algorithms[40] **.**

Fig 4. Sample action with KTH

C) IXMAS Dataset

IXMAS is knows as INRIA Xmas Motion Acquisition Sequences (IXMAS) a multiview dataset was developed for evaluation of view-invariant human action recognition algorithms in 2006. This dataset having 13 daily life actions performed by 11 actors 3 times with actions include crossing arms, stretching head, sitting down, checking watch, getting up, walking, turning around, punching, kicking, waving, picking, pointing, and throwing Basically, two types of methods have been proposed for multiview action recognition, i.e., 2D and 3D-based methods. The 3D based methods have reported higher accuracy than the 2D based methods on this dataset but at a higher computational[41].

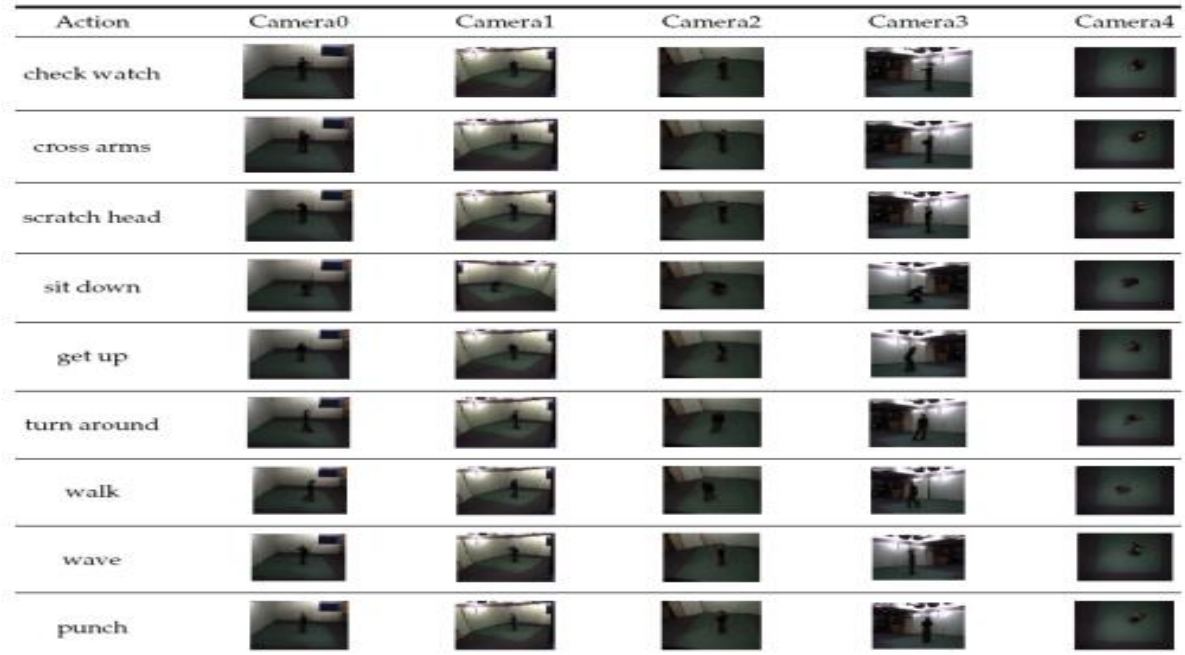

Fig 5. Sample action with IXMAS dataset

D) HMDB-51:

Its One of the largest datasets available for activity recognition developed by Serre lab, Brown University, USA in 2011. It consists of 51 types of daily life actions comprised of 6849 video clips collected from different sources such as movies, YouTube, and Google videos**[**42].

Fig 6. Sample action with HMDB-51

E) Hollywood2 [200] :

This dataset consists of 12 actions like get out of car, answer phone, kiss, hug, handshake, sit down, stand up, sit up, run, eat, fight, and drive car with dynamic background features was created by INRIA (Institut National de Recherche en Informatique et en Automatique), France in 2009. Dataset is extremely challenging, consists of short unconstrained movies with multiple persons, cluttered background, camera motion, and enormous intra class variations. This dataset is supposed for evaluation of HAR algorithms in real world scenarios[43].

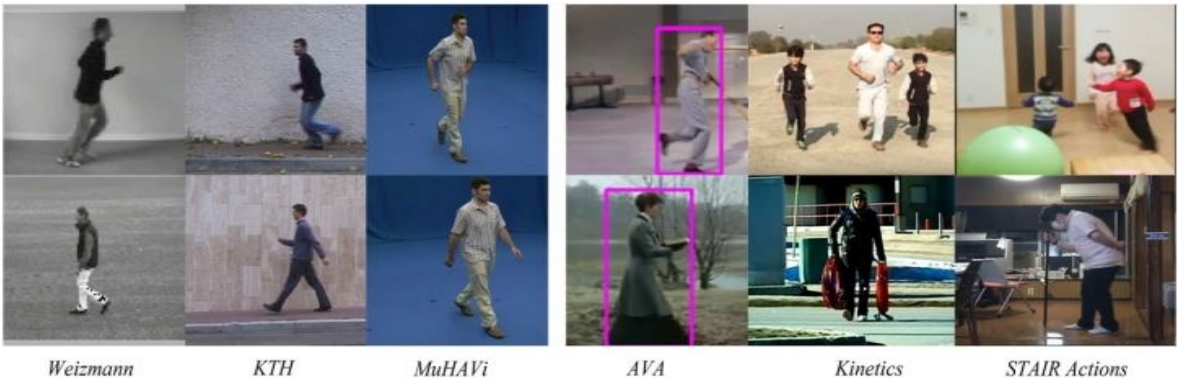

Fig7. Early Dataset may used in Human Action ,Activity for Behavior Understanding

Type II : Video datasets :

Those dataset are usful to understand action and activity from video or frame images. The few of them description of these early datasets listed below

I) a) Hollywood (2008), Hollywood2 (2009)and Hollywood2Tubes (2016) :

Another name of those dataset is HOHA was created at IRISA institute, France, for HAR in realistic natural video settings , which are annotated by automatic script-to-video alignment along with text-based script Classification in movie scripts. Hollywood has 8 activity classes Answer Phone, Get Out Car, Hand Shake, Hug Person, Kiss, Sit Down, Sit Up, and Stand Up collected from 32 movies such as The Butterfly Effect, Casablanca, and Lost Highway. Of these 20 form the train set and 12 the test set. The plaything is split into two: an automatic annotation set (233 clips) and a manually verified set (219 clips). The test set (211 clips) is manually verified. For the automated set, annotations correspond to sequence level, while for the manual sets, frame ranges are available. Hollywood2 is an extension of this and has 12 additional categories like Driving Car, Eat, and Run , there is 69 movies used here are split into 33 trains and 36 test clips. Many samples contain overlapping annotations[37]. Hollywood2Tubes provides action localization for Hollywood2 with bounding boxes and point annotations for all videos with an annotation stride of 10 s[38].

II) UCF11 (2009) UCF50 (2010) and UCF101 (2012)

The UCF datasets area unit a series of more and more massive datasets gathered from the net as a project by the Department of engineering and computing , University of Central American state. Since they use videos from unwritten sources, these datasets area unit extraordinarily difficult because of being collected in 'the wild'. All 3 datasets offer sequence level annotations. UCF11 or 'Actions at intervals the Wild' was one in every of the first datasets to believe free inputs, exploitation You-Tube videos as knowledge [44]. It contains eleven action categories like Basketball Shooting, Walking With A Dog, and Juggling associated with sports and daily actions. It contains 1168 videos sorted into twenty five comparatively freelance teams with variable background. UCF50 improved on UCF11 by including more categories (50 actions) and any reducing the interclass variation. As AN extension of UCF11, it includes the primary categories conjointly as actions like Jumping Rope, admixture Batter, and twine ascent. each UCF11 and UCF fifty use Leave One Out Cross validation for analysis. UCF101 [13] could also be a benchmark dataset of one zero one categories with thirteen,000 YouTube videos (27 h) (Fig. 7). it's AN extension of UCF50 and includes fifty one new categories like Brushing Teeth, dive , Floor athletics, and enjoying violoncello.

Fig.8. Sample action / Activity with HOHA

III. HMDB51 (2011) and JHMDB (2013)

The Human Motion info was created at Serre research laboratory, Brown, to raised capture the richness and quality of human actions. It contains fifty one actions sorted into five types: Facial (Smile, Chew), Facial + Object (Smoke, Eat), Body (Clap, Jump), Body + Object (Brush Hair, Draw Sword), and Body + Interaction (Hug, Shake Hands)[46]. every class contains a minimum of one zero one clips for a complete of 6766 clips extracted from sources like movies and YouTube. The videos area unit annotated manually with sequence level annotation and meta tags relating to visibility of the body components, camera motion and viewpoint, the amount of individuals and video quality. Train check splits (70,30) balanced with relevancy meta tags area unit provided. Joint-annotated HMDB (JHMDB) [128] could also be a set containing twenty one actions (928 clips) annotated using a 2nd human puppet pc model matched with every frame the video by AMT staff. From this ground truth, many options were computed: scale, pose, segmentation, coarse viewpoint, and dense optical flow for the humans in action. Recently, HMDB was used as bench- mark

IV. ASLAN dataset (2012) :

Action Similarity LAbeliNg (ASLAN) dataset having 1571 net videos with 432 complicated action categories. it absolutely was collected exploitation YouTube search queries supported the categories of the CMU dataset in conjunction with some new categories. 3631 action sequences area unit extracted with multiple categories allowed for a specific sequence. The dataset defines the action similarity metric that, not like action classification, focuses on the sameness of actions in 2 video and has been used as a benchmark for and Intersection over Union(IoU) similarity metric for tem poral detection [47].

V . Sports-1M (2014) :

Dataset which contains 1,133,158 YouTube sports videos which are annotated automatically with 487 sports labels using the YouTube Topics API. Example classes include Bowling, Cycling, Breast-stroke, Parasailing, Knife Throwing which are grouped in a manual taxonomy containing groups such as Aquatic Sports, Team Sports, Winter Sports, and Sports with Animals. The dataset contain videos which annotated automatically using text metadata 1000–3000 videos per class, but due to the large size and automatic annotation there are many irregularities [28]

Fig.9. Sample action / Activity with HMDB51 & JHDB

VI. SVW (2015):

Sports Videos in the Wild is a crowdsourced dataset of unconstrained sports videos. SVW consists of 4200 Smartphone camera videos recorded by users of Coach's Eye Smartphone app which allows users to compare their own videos with those of coaches or professionals. So, the dataset contains training videos by both professionals and amateur users. For genre classification, there are 30 categories of sports (Archery, Skating, Volleyball, etc.) and 44 actions which are manually labeled[[137]

VII) Charades (2016) and Charades ego (2018)

These datasets, created at the Allen Institute for AI, intro- duce the "Hollywood in Homes" crowdsourcing concept. Here, Amazon Mechanical Turk workers create scripts based on certain actions and objects provided to them by the researchers and then perform the scripts themselves making the dataset casual, creative as well as diverse. The original Charades dataset contains 157 classes of the form 'verb + preposition + noun' enacted by 267 people, with 1104 labels for 46 object classes and 27,847 textual descriptions of 9848 videos. Examples of the classes are *Put Down Laptop, Playing On Phone, Lying On Bed.* Each video is annotated with action class, action interval as well as object class. The dataset uses 80% train 20% test cross subject evaluation using mAP metric[47]

Fig.10 Sample action / Activity with Charades .

IX) MultiTHUMOS (2017)

MultiTHUMOS [145] is an extension of THUMOS'14 data- set with extensive frame level annotations of 30 h across 400 videos. the target of the dataset is to develop accu- rate localization of actions to encourage strong contextual modelling and multi-action reasoning. MultiTHUMOS features a long-tailed distribution of activities, as certain activities may occur very infrequently. Another challenge is that the inclusion of fine-grained actions also as having high intraclass vari- ation. Annotation was through with Datatang2, a billboard data annotation service, which provided manual action seg- ment annotation. Only the sets having temporal annotations within the original THUMOS'14 dataset were annotated.[48]

X) Something Something (2017)

The 20BN Something Something may be a dataset of 100,000 clips (each 2–6 s long), densely annotated with 174 fine-grained person–object categories of daily basic actions. These are grouped into 50 coarse-grained action groups. The videos like action templates like "Putting [something] into [something]" were collected employing a "Hollywood in Homes Approach" by AMT workers who selected the 'something'. The dataset features a cross subject train validation–test split within the ratio of 8:1:1. Something Something-V2 is larger, with 220, 847 videos of an equivalent 174 action categories. additionally , each video includes a caption that was authored and uploaded by the gang actor, that is, the captions mirror the action template, but with the generic placeholder Something replaced by the object(s) chosen by the actor [49]

XI) DALY (2017) :

Daily action localization in YouTube [50] may be a dataset created by THOTH, an INRIA research team and aimed toward temporaland spatial action localization. DALY contains 510 You- Tube Videos (31 h, 3.3 M frames) annotated with 10 daily actions (51 clips/class) having well-defined boundaries, e.g. Brushing Teeth, Cleaning Floor, Ironing, Drinking, etc. the gathering of videos is completed by direct search queries followed by manual annotation. Temporal annotation is completed using action segment of 8 s on the average . For spatial annotation, bounding boxes were annotated on subsample frames such a video may contain multiple different overlapping actions. It also contains upper body pose annotation, including a bounding box round the

head and any object(s) involved within the action. The train–test split is (31, 20) per class[50].

XII) AVA (2017)

Atomic Visual Actions may be a dataset released UC Berke- ley and Google with diverse environment and an outsized number of classes labelled using exhaustive frame level annotation of YouTube videos. 430 clips were sampled from the 15 to 30 min portion from movies and tv shows associated with famous actors from different countries. The annotations are spatiotemporally localized (Atomic Visual Actions) with a really fine granularity of 1 Hz, leading to 1.58 M action labels for 80 categories (e.g. Crawl, Dance, Paint, Smoke, Handshake, Play With Kids [51]

Fig.11 Sample action / Activity with AVA

XII) A2D (2017)

The Actor Actions Dataset is a YouTube video Dataset consisting of 3782 videos involving by 7 actors (Adult, Baby, Ball, Bird, Cat, Car, Dog) and 9 actions (Climbing, Crawl- ing, Eating, Flying, Jumping, Rolling, Running, Walking and No Action). One actor may perform many but not all actions (Adult-Flying etc. are not present), there are 43 valid actor–action classes.

XIII) Vlog (2017) :

Vlog is a large-scale YouTube dataset aimed at under- standing everyday human–object interaction through Life- style Vlogs, i.e. self-documenting videos. An important focus of the dataset is collecting implicitly tagged data instead explicitly querying for desired terms—since these invariably result in biased or staged data. Vlog- related queries ("daily routine 2013") were made in 14 European languages producing 216 K unique videos. Following automatic and manual filtering, the videos are manually labelled with sequence level annotation according to whether the person–object interaction is present, absent or inconclusive for 30 objects such as Food, Door, and Box. Additionally, scene category (bedroom, kitchen, bathroom, etc.), scene proxemics (personal, intimate, social, etc.), hand state (number of people and if they touch the object) and hand bounding box annotation is also provided. [53]

XIV) SoccerNet (2018)

The dataset contains 500 full games (764 h) with three classes (Goal, Yellow/Red Card, and Substitution). Annotations are mined automatically from commentary obtained from sports sites and manually refined to 1 s resolution by anchoring them to a single timestamp. Temporal annotation is done using 'spotting' with anchor times (similar to action points) which is the point in time that uniquely identifies an event. The evaluation is done using mAP metric. SoccerNet is a benchmark dataset for sparse action localization in soccer videos from six main European Leagues (2014–2017)[54].

XV) MLB-YouTube (2018) :

The MLB-YouTube (Major League Baseball) is a Fine- Grained Activity dataset consisting of 20 baseball games from the 2017 MLB postseason. The 9 activities (Swing, Foul, Ball, Strike, etc., and No Action) are not very distinct and there may be significant occlusion of activities. The multilabel and overlapping along

with pitch type (e.g. Fastball, Curveball, Slider, etc.) and the speed of the pitch also being given for each pitch. There are two separate sets: segmented video: sequence level annotation for activity recognition of 4290 video clips totalling 42 h; continuous video: dense Frame level annotation for activity classification of 2128 clips each 1–2 min long[55].

Fig. 12. Sample action / Activity with MLB-YouTube

XVI) **STAIR actions (2018) :**

Software Technology and Artificial Intelligence Research Laboratory (STAIR)Lab, University of Chiba, Japan created a dataset of 100 k + videos .It contains 100 categories each with 900–1200 trimmed video clips representing fine- grained everyday home activities. These were selected by sampling the Japanese Wiktionary verb list for verbs associ- ated with office, home, washroom, kitchen and living room. Action labels take the form of 'verb + object' (categorized into five types), e.g. Eating Meal (Kitchen), Brushing Teeth[56]

XVII) Kinetics400 (2017) and Kinetics600 (2018)

Kinetic datasets constitute a large-scale, high-quality You- Tube dataset, which includes human focused actions filmed in non-ideal conditions. Originally the dataset included 400 classes but was extended [153] to 600 classes (with some renaming and splitting of 32 original classes). The classes cover Person Actions (Drawing, Pumping Fist), Person–Person Actions (Hugging, Shaking Hands) and Per- son–Object Actions (Opening Present, Mowing Lawn). A nonexclusive parent–child grouping for action classes is also provided (Playing Games: Flying Kite, Hopscotch; Music: Beatboxing, Singing, etc.[14]

XVIII**) EPIC Kitchens (2018)** :

Egocentric perception, interaction and computing (EPIC) kitchens is a very large egocentric dataset of daily kitchen activities. It consists of 55 h of crowdsourced video recorded by 32 participants with a head-mounted GoPro camera in their own homes. Later, the participants also anno- tated their own footage with audio narration in their own language, which was then transcribed by AMT workers and further aligned to the video using the YouTube Closed Cap- tions Algorithm.. A total of 454.3 K object bounding boxes are also provided corresponding to the objects that take part in the action segment. [57]

XIX) Moments in time (2018) :

Created the Moments [21] very large-scale dataset created by MIT-IBM Watson AI Lab to help AI systems recognize and understand actions and events in 'moments'. Moments have both visual and auditory modalities. It contains 339 verb-based classes (from a set of most frequently used verbs) such as Chasing, Licking, Winking, Sewing, and Sliding. These were used to query multiple search engines to collect videos from which 3-s clips were randomly selected.[58].

Fig13. Sample action / Activity with Moments in time

XX) HACS (2019) :

Human Action Clips and Segments Dataset contains two subsets with sparse and dense annotation: "Clips" having sequence level annotations and "Segments" having action segment annotation both retrieved from about 504 K YouTube videos. HACS uses the ActivityNet-v1.3 classes to query for source videos where Clips contains 1.55 M sparsely sampled 2-s clips containing both positive (0.6 K) and negative (0.95 K) class examples[59].

Fig15. Sample action / Activity with HACS

Type III.: Image datasets

Even though the temporal context plays a pivotal role in the identification of images, it is still possible to recognize certain actions only from static images several datasets consisting of labelled still images have been constructed though this has received less interest than video datasets shown (Table 4) [60]

A) Willow (2010) :

Its a still image dataset of common human action in 968 consumer photographs collected from Flickr using search queries followed by manual filtering with seven classes such as Interacting With Computers, Riding Horse, and Walking. Manually annotated bounding boxes are provided for every person within the image. There is a train set with 70 images per class with the rest being in the test set. The metrics used are classification accuracy and mAP[60]

B) Stanford 40 actions (2011)

Its a **picture** dataset of 40 daily **act** obtained from Google, Bing and Flickr by search queries. There are a total of 9352 images with bound- ing boxes for the person performing in each image. For each class, there are about 180–300 images. Of these, 100 forming the train set and the rest forming the test set[61].

C)Tuhoi (2014) :

Trento Universal Human Object Interaction Dataset is an image only dataset of person–object interaction with 189 common objects in 10,805 images from the DET dataset in the ImageNet 2013 challenge [202]. 2974 unique actions in the form "verb + object" are annotated using Crowd- flower, a crowdsourcing service. Example objects include Dog, Watercraft, and ball and example verbs include Eat, Hit, Throw along side No Action. The images are split into 50–50 train and test sets with each action represented in either sets. [62]

D) HICO (2015) and HICO-DET (2018):

Humans Interacting with Common Objects(HICO) is a data- set focused on distinguishing a variety of

common sense- based interactions with the same object. It has a total of 47,774 images, with a total of 600 action categories in'verb + object' form. There are 80 (e.g. Bike, telephone , Apple) objects and 117 classes (e.g. Ride, Feed, Cut) are common to multiple objects. The ground truth is within the sort of multiple action labels per image along side a 'No Action' label (e.g. "person is near but not interacting with bicycle" = "Bike No Action"). The images are selected from Flickr according to queries relating to the classes and then verified manually by AMT worker. The evaluation metric is mAP per image with an 80–20 training–test split[63].

E) HICO- DET (2018)

HICO –DET Augments the dataset with instance level annotations consisting of bounding boxes of the objects and therefore the persons involved within the actions (explicitly ignoring unrelated people) along with a link between the person and relevant objec) [64]

F) BU-Action (2017)

The Boston University Action Datasets are three image datasets downloaded using search queries corresponding to the classes of large benchmark video datasets, namely UCF101 and ActivityNet. UCF101 classes and also includes 2769 images from the Stanford40 Dataset uses BU101-filtered. These are manually filtered to supply 23.8 K images. BU101-unfiltered also has the UCF101 classes but the images are not filtered after the queries thus producing a larger dataset of 204 K images. Similarly, BU203-unfiltered consists of unfiltered images queried from the 203 Activity[[65]

IV. The performance evaluation of HAR model.

Here we provide various evaluation metrics used in existing HAR models with the description of metrics along some key concept illustrate below and table indicate formula to calculate measure accuracy

- True positive (TP): no. of positive samples predicted correctly.
- False positive (FP): no. of actual negative samples predicted as positive.
- True negative (TN): no. of negative samples predicted correctly.
- False negative (FN): no. of actual positive samples predicted as negative

S. No	Metrics	Description
1	Accuracy = TP +TN / TP+TN+FP+FN	Ratio of number of correct prediction and total number of input samples
2	Precision = TP TP +FP	It is the no. of correct positives divided by the predicted positives
3	Recall = TP TP +FN	It is the no. of correct positives divided by total no. of true positives and false negatives
4	F1 − score = 2∗ [P × R P + R]	Harmonic mean between precision and recall
5	Specificity = TN TN +FP	The proportion of actual negatives predicted as positives
6	Sensitivity = TP TP +FN	The proportion of actual positives predicted as positives
7	Positive LHR = Sensitivity 100−Specificity Negative LHR = 100−Sensitivity Sepcificity	LHR assess the goodness of ft of two competing statistical models based on their likelihood

V. Conclusion & Discussion:

For HBA many public datasets used by researchers in order to validate their proposals and to evaluate their performance. These datasets / databases can be grouped into several classes depending on the types of action they contain, the viewpoint as well as the nature of data: databases relating to movie scenes, social networks, human behaviors, human poses, atomic actions or daily life activities. The most used datasets in the literature and categorize them according to activity types. We consider in this

classification only five types (levels): atomic action level, behavior level, interaction level and group activities level

Table 1. Sample Datasets with categorization.

Dataset	Action	Behavior	Human-object interaction	Human-human interaction	Group activities
KTH	YES				
Weizmann	YES				
XMAS	YES				
MSR Action 3	YES				
VISOR		YES			
Caviar		YES			
MCAD		YES			
MSR Daily Activity 3D	YES		YES		
50 Salads	YES		YES		
MuHAV	YES	YES	YES		
UCF Sports	YES		YES		
UCF50	YES		YES		
ActivityNet	YES	YES	YES	YES	YES
HMDB-51	YES	YES	YES	YES	YES
Hollywood & Hollywood2	YES	YES	YES	YES	YES
UCF-101	YES	YES	YES	YES	YES
YouTube Action	YES	YES	YES	YES	YES
Behave				YES	YES
Video Web			YES	YES	YES

Table 2 Evolution of Datasets (early)

Dataset	Year	Focus	Classes
Weizmann Event [19]	2001	Events, statistical methods	4
KTH [20]	2004	Activity benchmark	6
Weizmann [21]	2005	Silhouette based methods	10
CAVIAR [14]	2005	Surveillance	9
ETISEO [15]	2005	Video surveillance	15
ViSOR [22]	2005	Video surveillance	–
IXMAS	2006	Multi-camera	13
CASIA	2007	Behaviour analysis	8
UCF Aerial Action	2007	Aerial viewpoint	9
UCF-ARG	2008	Multiview (aerial–roof–ground)	10
UCF sports action [23]	2008	Sports, annotated	10
UIUC action [24]	2008	Sports, annotated	14
i3DPost multiview [25]	2009	3D volumes	13
URADL [26]	2009	Daily actions	10
Collective Activity Dataset [27]	2009	Collective activity, real-world data	5
BEHAVE [29]	2010	Activity benchmark	10
MuHAVi [30]	2010	Silhouette based methods, multiview	17
UT-interaction [31]	2010	Person–person actions	6
UT-tower [32]	2010	Action recognition at a distance	9
UCR-Videoweb [31]	2010	Non-verbal communication analysis, person–person actions	9
VIRAT video [33]	2011	Realistic video surveillance	12
MINTA [34]	2011	Kitchen activities, humanoid robot viewpoint, intention–activity–motion primitive distinction	9 (intention), 6 (motion)
KIT Robo-Kitchen Activity Dataset [12]	2011	Kitchen activities, humanoid robot viewpoint, multiview	14

Table 3. Evolution of modern datasets

Dataset	YearFocus	Method	Annotation	Classes
Hollywood	2008Daily actions	Movies, script alignment	Frame range/sequence level	8
Hollywood2	2009Daily actions	Movies, script alignment	Frame range/sequence level	20
UCF11	2009Actions in the wild	YouTube, manual annota-	Sequence level	11

High Five]	2010	Person–person actions	V Shows, manual annota-tion	Sequence level + upper interaction label	5
Olympic Sports]	2010	Sports actions, non-periodic	YouTube, AMT annotation	Sequence level	16
UCF50	2010	Actions in the wild	YouTube, manual annota-	Sequence level	50
HMDB	2011	Benchmark dataset	YouTube and movies, AMT	Sequence level + meta tags	51
UCF101]	2012	Benchmark dataset	YouTube, manual annota-	Sequence level	101
MPII Cooking]	2012	ooking actions, fine-grained actions, activities	Semi-scripted (recipe as	Frame range + pose	65
IPII Cooking Composite	2012	ooking actions, compositeactions	emi-scripted (recipe asscript)	Frame range + ingredient, script data	41
ASLAN]	2012	Action similarity metric	Search query	Sequence level	432
Hollywood2Tubes[a]	2013	ction localization, pointlocalization	Manual annotation	Point annotation, bounding	20
YouCook	2013	Cooking actions, summari-	YouTube, AMT annotation	Frame level (object, actor)	7

Continue Table 3. .....

Dataset	Year	Focus	Method	Annotation	Classes
THUMOS'13[a]	2013	Benchmark	Manual annotation	Sequence level + bounding	101

JHMDB[a]	2013	Joint annotated using 2D	AMT annotation	Joint annotations	21
Breakfast Actions	2014	reakfast actions, cookingactions	Semi-scripted (recipe as	Frame range	10
THUMOS'14[a]	2014	Benchmark dataset	YouTube, manual annota-	Sequence level	102
Sports1M	2014	Sports actions	YouTube ,YouTube topics	Sequence level	487
THUMOS'15	2015	Benchmark dataset	YouTube, manual annota-	Sequence level	102
Crêpe Dataset	2015	Cooking actions	Scripted	Dense frame level (action, +bounding boxes	9
SVW	2015	ports videos, smartphonecamera view	Crowdsourced	Sequence level/frame range	44
ActivityNet]	2015	rimmed and untrimmedactions, web	Search query collection, annotation	Sequence level/frame range	203
MPII Cooking 2	2015	ooking actions, fine-grained actions, compositeactions	Semi-scripted, AMT	rame range + pose, handtool, container ıbels,	67
MERL Shopping Dataset	2016	Shopping activities	Surveillance videos	Frame range	5

Continue…….
Table 3.

Something Something	2017	Fine-grained labeling, caption template, levels of label granularity	Hollywood in homes crowd-	Sequence level + object sourcing	labels
DALY	2017	aily actions, spatiotempo-ral localization	earch query, manual annotation	Frame range + bounding pose (upper body)	
AVA	2017	xhaustive annotation, atomic actions	lovies, hybrid, faster RCNN + manual annota-	Dense frame level + bound-	
A2D	2017	ctor–action correspond-ence	Search query	Dense frame level + pixel-	
Kinetics400	2017	Human focused, benchmark	YouTube, AMT annotation	Sequence level	
Kinetics600	2017	Human focused, benchmark	YouTube, AMT annotation	Sequence level	
Vlog	2017	ifestyle Vlogs, implicittagging, daily actions,	ouTube, search query,manual annotation	Sequence level + attribute	
YouCook2	2018	ooking actions, instruc- tional videos, procedure	ouTube, manual annota-tion	Frame range (sentences as	
SoccerNet	2018	Soccer actions	Sports broadcasts, anno- mining	Action points	
MLB YouTube annota-	018	ine-grained activity, base-		YSequence o level/dense frame u T u b e , m a n	

		ball videos, overlapping	tion	u a l	
STAIR Actions]	2018	ine-grained activity, pairedactions	YouTube home videos, annotation	Sequence Level	
Baseball (BBDB)	2018	ine-grained activity, base-ball videos	Sports broadcasts, anno- mining	Frame range	
Moments in Time	2018	vent detection, moments,benchmark	Veb videos, search query,AMT annotation	Sequence level (short	
HACS	2019	Temporal localization	ouTube, search query,manual annotation	Sequence level/ Frame	
COIN [	2019	nstructional video, hierar-chy of actions (domain-	YouTube, manual annota-	Frame range	
Mining YouTube	2019	Automatic extraction of	YouTube, search query, text	Frame range	

Reference:

[1] Liu, Y., Nie, L., Liu, L., Rosenblum, D.S.: From action to activ- ity: sensor-based activity recognition. Neurocomputing **181**, 108–115 (2016)

[2] Dai, X., Singh, B., Zhang, G., Davis, L.S. and Chen, Y.Q.: Tem- poral context network for activity localization in videos. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2017, pp. 5727–5736 (2017

[3] Fouhey, D.F., Kuo, W.C., Efros, A.A., Malik, J.: From lifestyle Vlogs to everyday interactions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4991–5000 (2018)

[4] Soomro, K., Zamir, A.R.: Computer Vision in Sports. Springer,Berlin (2014)

[5] Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH: a spatio- temporal maximum average correlation height filter for action recognition. In: 26th IEEE IEEE Conference on Computer Visions Pattern Recognition, CVPR (2008)

[6] Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal struc- ture of decomposable motion segments for activity classification. In: Lecture Notes in Computer Science (Including Subseries Lec- ture Notes in Artificial Intelligence and Lecture Notes in Bioin- formatics), vol. 6312, no. PART 2, pp. 392–405. LNCS (2010)

[7] Karpathy, A. et al.: Large-scale video classification with convo-lutional neural networks.

[8] Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.:A hierarchical deep temporal model for group activity recogni- tion. In: Proceedings of the IEEE Computer Society Conferenceon Computer

Vision and Pattern Recognition, vol. 2016, pp. 1971–1980 (2016)
[9] Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: a scalable dataset for action spotting in soccer videos. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work, vol. 2018, pp. 1792–1802 (2018)
[10] Bloom, V., Makris, D., Argyriou, V.: G3D: a gaming action data- set and real time action recognition evaluation framework. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work, pp. 7–12 (2012)
[11] Bloom, V., Argyriou, V., Makris, D.: G3di: a gaming interaction dataset with a real time detection and evaluation framework. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformat- ics), vol. 8925, pp. 698–712 (2015)
[12] Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work. CVPRW 2010, vol. 2010, pp. 9–14 (2010)
[13] Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: joint learning of gaze and actions in first person video. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Arti- ficial Intelligence and Lecture Notes in Bioinformatics), vol. 11209, pp. 639–655. LNCS (2018)
[14] Oxford Mobile Robotics Group, Oxford RobotCar Dataset. WWW (2016). [Online]. http://robotcar-dataset.robots.ox.ac.uk. Accessed 25 Mar 2019
[15] List, T., Bins, J., Vazquez, J., Fisher, R.B.: Performance evaluat- ing the evaluator. In: Proc.—2nd Jt. IEEE Int. Work. Vis. Sur- veill. Perform. Eval. Track. Surveillance, VS-PETS, vol. 2005, pp. 129–136, 2005
[16] Nghiem, A., et al.: ETISEO, performance evaluation for video surveillance systems. In: IEEE International Conference on Advanced Video and Signal based Surveillance. IEEE, London, UK, 5-7 Sept. 2007. https://doi.org/10.1109/AVSS.2007.44253 57
[17] Zhang, J., Li, W., Ogunbona, P.O., Wang, P., Tang, C.: RGB-D- based action recognition datasets: a survey. Pattern Recognit. **60**,86–105 (2016)
[18] Chavarriaga, R., et al.: The Opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recognit. Lett. **34**(15), 2033–2042 (2013)
[19] Piergiovanni, A.J., Ryoo, M.S.: Fine-grained activity recognition in baseball videos. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work, vol. 2018, pp. 1821–1830 (2018)
[20] Cao, L., Liu, Z., Huang, T.S.: Cross-dataset action detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1998–2005 (2010)
[21] Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE Transactions *on* PatternAnalysis *and Machine* Intelligence (2007)
[22] Nghiem, A.T., Bremond, F., Thonnat, M., Ma, R.: ETISEO data- set. Www, 2007. [Online]. https://www-sop.inria.fr/orion/ETISEO/. Accessed: 25 Mar 2019
[23] Gella, S., Keller, F.: An analysis of action recognition datasets for language and vision tasks. In: ACL 2017—55th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference (Long Paper), vol. 2, no. c, pp. 64–71 (2017)
[24] Khurram Soomro, A.R.Z.: UCF sports action data set, Www (2008). [Online]. http://crcv.ucf.edu/data/UCF_Sports_Actio n.php
[25] Tran, D, Sorokin, A.: Human activity recognition with metric learning, www, 2008.
[26] Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., Pitas, I.: i3DPostmulti-view human action datasets. (2009)
[27] Messing, R., Pal, C., Kautz, H.: University of Rochester activi- ties of daily living dataset, 2009.
[28] Le, D.-T., Uijlings, J., Bernardi, R.: TUHOI: Trento universal human object interaction dataset. In: Proceedings of the Third Workshop on Vision and Language (2014)
[29] Project, B.: Computer-assisted prescreening of video streams for unusual activities. Www, 2004. [Online]. http://homepages.inf. ed.ac.uk/rbf/BEHAVE/
[30] Singh, S., Velastin, S.A., Ragheb, H.: MuHAVi: multicamera human action video data. [Online]. http://dipersec.king.ac.uk/

[31] http:// MuHAVi-MAS/.
[32] Ryoo, M.S., Aggarwal, J.K., Chen, C., Roy-chowdhury, A.: ICPR 2010 contest on semantic description of human activities (SDHA2010), Www (2010). [Online]. http://cvrc.ece.utexas.edu/SDHA2010/
[33] Chen, C.-C., Ryoo, M.S., Aggarwal, J.K.: UT-tower dataset: aerial view activity classification challenge (2010). http://cvrc. ece.utexas.edu/SDHA2010/Aerial_View_Activity.html
[34] Oh, S. et al.: VIRAT video dataset, Www (2011). [Online]. http:// www.viratdata.org/
[35] Yoshikawa, Y., Lin, J., Takeuchi, A.: STAIR actions: a video dataset of everyday home actions (2018). arXiv:1804.04326
[36] Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: HICO: A benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2015 Inter, pp. 1017–1025 (2015)
[37] Kliper-gross, O., Hassner, T.: The action similarity labeling chal-lenge. Www (2012).
[38] Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work CVPR Work. 2009, vol. 2009 IEEE, no. i, pp. 2929–2936 (2009)
[39] Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247
[40] Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: 2004. ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition, vol 3. IEEE, pp 32–36
[41] Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Understand 104(2-3):249
[42] Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV)
[43] Marszałek M, Laptev I, Schmid C (2009) Actions in context. In: IEEE Conference on computer vision & pattern recognition
[44] Ivan, L., Marszałek, M., Schmid, C., Rozenfeld, B.: IRISA/ INRIA Rennes France: learning human actions from movies, Www (2008). [Online].
[45] "Olympic sports dataset," Standford University (2010). [Online].
[46] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2556–2563 (2011)
[47] Kliper-Gross, O., Hassner, T., Wolf, L.: The action similarity labeling challenge. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 615–621 (2012) Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: crowdsourcing data collection for activity understanding. In: Lecture Notes in Computer Sci- ence (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9905, pp. 510–526. LNCS (2016)
[48] Mori, G., Andriluka, M., Russakovsky, O., Jin, N., Fei-Fei, L., Yeung, S.: Every moment counts: dense detailed labeling of actions in complex videos. Int. J. Comput. Vis. 126(2–4), 375– 389 (2017)
[49] Goyal, R. et al.: The 'something something' video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision, 2017, vol. 2017, pp. 5843–5851 (2018)
[50] Weinzaepfel, P., Martin, X., Schmid, C.: Human action localization with sparse spatial sup
[51] Gu, C. et al.: AVA: a video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6047–6056 (2018)
[52] Kay, W. et al.: The kinetics human action video dataset (2017) arXiv:1705.06950 [cs.CV
[53] Fouhey, D.F., Kuo, W.C., Efros, A.A., Malik, J.: From lifestyle Vlogs to everyday interactions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4991–5000 (2018)
[54] Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: a scalable dataset for action spotting in soccer videos. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition

Work, vol. 2018, pp. 1792–1802 (2018

[55] Piergiovanni, A.J., Ryoo, M.S.: Fine-grained activity recognition in baseball videos. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work, vol. 2018, pp. 1821–1830 (2018)

[56] Yoshikawa, Y., Lin, J., Takeuchi, A.: STAIR actions: a video dataset of everyday home actions (2018). arXiv:1804.04326

[57] Moltisanti, D. et al.: Scaling egocentric vision: the dataset. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artifcial Intelligence and Lecture Notes in Bioinformatics), vol. 11208, pp. 753–771. LNCS (2018)

[58] rt, M. et al.: Moments in time dataset: one million videos for event understanding. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1 (2019)

[59] Zhao, H., Yan, Z., Torresani, L., Torralba, A.: HACS: human action clips and segments dataset for recognition and temporal localization (2017). arXiv:1712.09374

[60] Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: British Machine Vision Conference, BMVC 2010— Proceedings, pp. 97.1–97.11 (2010)

[61] Yao, B., Jiang X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: Proceedings of IEEE International Conference on Computer Vision, pp. 1331–1338 (2011)

[62] Le, D.-T., Uijlings, J., Bernardi, R.: TUHOI: Trento universal human object interaction dataset. In: Proceedings of the Third Workshop on Vision and Language (2014)

[63] Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: HICO: A benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2015 Inter, pp. 1017–1025 (2015)

[64] Ma, S., Bargal, S.A., Zhang, J., Sigal, L., Sclarof, S.: Do less and achieve more: training CNNs for action recognition utilizing action images from the web. Pattern Recognit. 68, 334–345 (2017).

[65] Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Front Robot AI 2:28