Towards Recognition of Human Activities and in depth Analysis of Evaluation Datasets

MUHAMMAD USMAN GHANI KHAN[1,2*], OMER IRSHAD[1,2], SALMAN ROUBIL[1]

[1]*Al-Khawarizmi Institute of Computer Sciences*
[2]*Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan*

**Corresponding author's e-mail: usman.ghani@uet.edu.pk*

Abstract

Human activity detection is one of the key areas of interest in all surveillance applications. This work has earlier been performed manually but recent technologies have successfully converted into automated systems to carry out this task. A variety of online dataset repositories are available for training the systems and testing them later. Some datasets are solely inclusive of single human actions, however, in some others, the interactions of humans with objects and other humans have also been accounted for. This paper focuses on single human actions alone and a survey of datasets with the dominance of single human actions.

Keywords: Human-object interaction; Human activity detection; Surveillance; dataset evaluation

INTRODUCTION

Human activity recognition and human to object interaction [1], [2] are considered to be the key objectives of surveillance systems to watch over any suspicious activity taking place. The purpose can be achieved by using human resources for monitoring the CCTV footage. However, human beings cannot be exempted from errors due to their inability to focus on a task for longer uninterrupted durations of time. Therefore, there is a need to develop automated vision-based computer systems to get the desired functionality more accurately and efficiently. A variety of online dataset repositories are available for training and testing such systems.

The increased rate of terror attacks coupled with crime activities in Pakistan demands development of comprehensive automated systems capable of detecting the human presence, identifying and recognizing their actions; detecting their emotions, age and gender and completely describing the scenario under observation. On the contrary, there is a need to predict these doubtful events prior to occurrence in order to prevent them from happening

altogether. System requires training for predicting an activity based on unit human actions leading to the complex activities.

In the context of single human action recognition, many online datasets are available, however, most of them consist of human to object interaction along with basic single human actions. This review has its main focus on datasets with the dominance of single human actions out of which KTH [3],[4] is one of the most frequently used datasets. This dataset may not be inclusive of larger set of actions but due to its uniform background, it is generally employed for training purposes. HMDB51[15] dataset has its significance in catering real life scenarios and is commonly used for testing of trained system. To account for actions captured from different view angles, datasets such as MuHAVi [6]-[7] dataset play their part. IXMAS [8-10] dataset is vital for training the system for dealing with both occluded and non-occluded actions. UT Tower action dataset [11] presents the searchers with action recorded from an aerial view.

Distinction in the dataset is made on the following parameters in the field of activity recognition or CV (Computer Vision) domain:

i. Presence of Human
ii. Human Skeleton/Structure
iii. Inner parts of body
iv. Arm Movements
v. Leg Movements
vi. Bending Curves of Human Skeleton
vii. Position of Human with respect to objects
viii. Indoor/Outdoor Environments
ix. Camera Angle

DATASET REVIEW

Following eleven datasets have been reviewed in this paper.

1. KTH
2. MuHAVi-MAS
3. UT Towers
4. IXMAS Action
5. UIUC
6. HMDB51
7. WVU
8. MPII
9. WEIZMANN
10. CASIA
11. Berkeley MHAD

KTH Dataset

The KTH dataset had been generated in 2004 by Royal Institute of Technology [12]. six human action classes have been focused in KTH dataset [4, 13]. It consists of 4 different conditions for recording actions i.e. outdoors, outdoors with scale variation, outdoors with different clothes and indoors. Each of the 25 subjects was made to perform these 6 actions, 4 times each. There are a total of 2391 video sequences in KTH dataset [13,3]. The resolution of these videos is 160x120 [13],[14]. It is considered to be one of the largest datasets available for single human action recognition [12]. The dataset has been divided into training set (8 subjects), validation set (8 subjects) and test set (9 actors) [13].

The action classes [1] included in this dataset are;

(i) Walking
(ii) Jogging
(iii) Running
(iv) Boxing
(v) Waving
(vi) Hands clapping

Apart from videos, the dataset also provides ground truths. The dataset consists of single action labels assigned to each action class [4]. This dataset is limited to clean, homogeneous and uniform backgrounds [2],[3]. This trait of KTH proves good for training reasons but some researchers demand rather complex and realistic backgrounds for testing purposes [2]. It is for this reason that KTH is mostly used for training purposes with action videos from other datasets used as testing data. Frames from a few instances of KTH dataset i.e. boxing, jogging and waving have been shown in Figure I

(a) Boxing **(b): Jogging** **(c): Waving**

Figure 1: Few instances of KTH Dataset

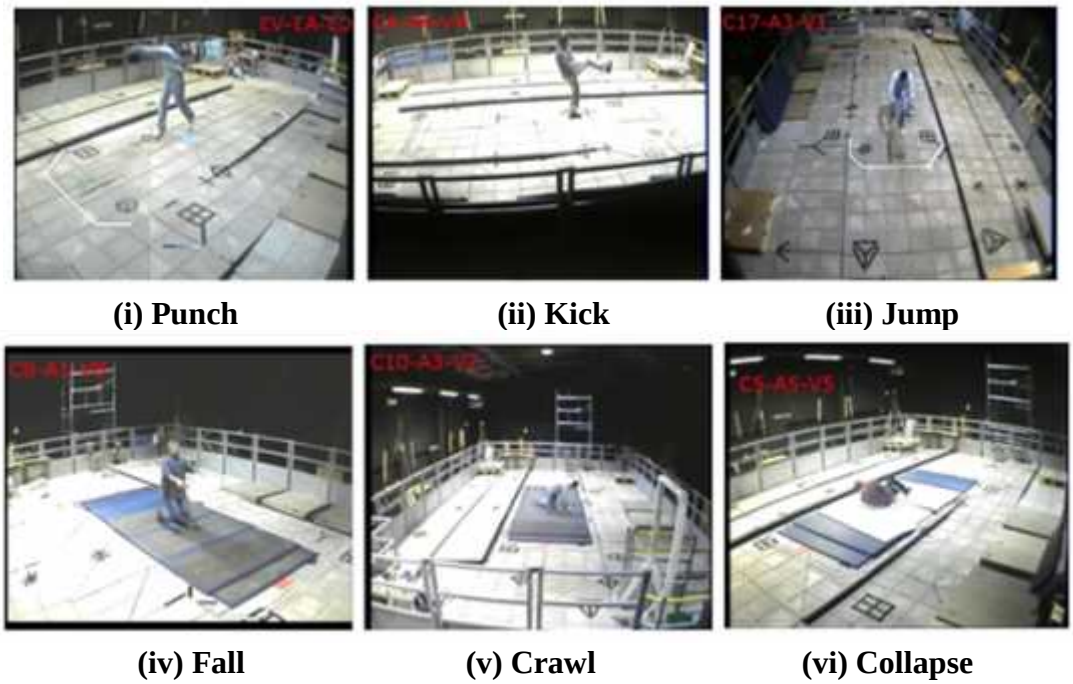
MuHAVi-MAS Dataset

This dataset has been originally generated by a team at Digital Imaging Research Centre, Kingston University in the year 2010. Each of the 17 different action classes, performed by 4 actors has been recorded from 8 different angles using eight CCTV cameras. Each action has been performed several times by actors in the action zone [6]. The resolution of these videos is 720x256.

The actions in this dataset are:

(i) Punch
(ii) Kick
(iii) Walk-fall
(iv) Walk-turn-back
(v) Run-stop
(vi) Shot gun-collapse
(vii) Pull heavy object
(viii) Pick up/Throw object
(xi) Walk-fall
(x) Look in car
(xi) Crawl on knees
(xii) Wave arms
(xiii) Raw graffiti
(xiv) Jump over fence
(xv) Drunk walk
(xvi) Climb ladder
(xvii)Smash object

The action videos with uneven backgrounds have been recorded in night street light illumination [6] and have been manually synchronized. Videos from 5 action classes have been manually annotated. The silhouettes of human presence and actions in various frames of videos has been provided for this subset of action classes [7]. It comprises of both unit actions and composite actions. For training purposes, often these composite actions have been cropped to obtain unit actions videos [7] e.g. walk-fall is a composite action and can be further divided into two unit actions i.e. walking and falling. Figure 2 (A) and Figure 2 (B) display a few samples of MuHAVi dataset and silhouettes of some actions of the same dataset respectively.



(i) Punch **(ii) Kick** **(iii) Jump**

(iv) Fall **(v) Crawl** **(vi) Collapse**

Figure 2A: A few instances of MuHAVi dataset

(i) Collapse (ii) Punch (iii) Kick

Figure 2B: Few instances of MAS silhouettes

UT Tower Dataset

Nine action classes have been targeted in UT tower dataset (pointing, standing, digging and others) [11]. This aerial dataset has been recorded using static camera setup at the top of the main tower of University of Austin. Each action has been performed 12 times by 6 individuals with 2 backgrounds i.e. floor and grassy lawn [15]. Five actions recorded in this dataset are:

(i) Carry
(ii) Run
(iii) Wave with one hand
(iv) Wave with two hands
(v) Jump

The remaining 4 actions i.e. point, stand, dig and walk have been recorded with floor background. There are 108 videos in total [16]. All videos have low resolution since the actions have been recorded from a distance. In these video frames, the person height is no more than 40 pixels. The figure centralization detector has been applied to provide figure-centric patches in the video frames [15] in order to focus on the person of interest only. These poses are common for recognizing multiple games action also [11]. Figure 3A presents a few cases of actions showing the person of interest only, however, Figure 3B shows the complete frames of actions and associated background taken from an aerial view.

(i) Pointing (ii) Walking (iii) Jumping (iv) Running

Figure 3A: Few instances of UT Tower dataset

(i) Bounding box **(ii) Corresponding Frame**

Figure 3B: UT Tower Dataset

IXMAS Action Dataset

IXMAS action dataset was first published in the year 2007. It comprises of 1148 videos, most of which are without occlusions except 698 video sequences in which objects partially occluding the actor. Eleven actions have been included in this dataset. These actions were performed by 10 actors; 5 males & 5 females. Each of these action instances has been recorded from five different view angles using five standard fire wire cameras [8].
The eleven actions include:

(i) Walking
(ii) Waving
(iii) Punching
(iv) Kicking
(v) Picking up
(vi) Sitting down
(vii) Checking Watch
(viii) Crossing arms
(ix) Scratching head
(x) Getting up
(xi) Turning around

Resolution of these videos is 390x291. Ground truths and bounding boxes have also been provided along with the videos. It consists of both un-occluded and partially occluded single human actions [9]. It aids in the development of more robust and realistic systems where the human actions are expected to be partially occluded in most general cases. Volumetric voxel representations and silhouette of human actions obtained by eliminating backgrounds have been included in the dataset. [10]. A few examples of occluded actions and some of un-occluded actions have been displayed in Figure 4A and Figure 4B respectively.

(i) Turn Around **(ii) Get Up** **(iii) Punch**

Figure 4A: IXMAS (without occlusion)

(i) Punch **(ii) Turn Around** **(iii) Get Up**

Figure 4B: IXMAS (with occlusion)

UIUC Dataset

UIUC dataset includes video sequences of 14 general human actions and 4 badminton sequences [17]-[18]. Eight actors have cooperated in the generation of this dataset. It comprises of 532 videos in total.

The action classes included in this dataset are very diverse which are listed below:

(i) Crawling
(ii) Turning
(iii) Clapping
(iv) Walking
(v) Running
(vi) Jumping
(vii) Waving
(viii) Jumping jack
(ix) Jump from sit-up
(x) Raise one hand
(xi) Stretch out
(xii) Sit to stand
(xiii) Pushing up
(xiv) Stand to sit

Apart from these action classes, badminton sequences have also been included in this dataset. These videos have been recorded from a single camera view. All the video recordings have a uniform and clean background [18]. Foreground masks are also provided. Figure 5 shows selected collection of actions from this dataset.

(i) Clapping **(ii) Waving** **(iii) Pushing Up**

(iv) Turning **(v) Jumping** **(vi) Crawling**

Figure 5: UIUC dataset

HMDB51 Dataset

HMDB dataset has been collected from various public sources such as YouTube and Google videos. A total of 51 actions have been added in this dataset out of which nineteen are single human actions and remaining include interactions of human beings with objects or with other human beings. Each category has 101 video clips on an average [5]. It consists of images taken from a static camera and in some cases, taken from a moving camera. This dataset contains 6766 video sequences. Thirty percent of the data is used for testing while 70 percent for training against each action class [18]. The dataset has both the originally generated videos and the stabilized version of these videos i.e. standard image stitching techniques were used to align frames of the video [5], [18]. A few sample instances of the dataset have been revealed in Figure 6.

(i) Standing **(ii) Kicking** **(iii) Clapping**

Figure 6: HMDB51 Dataset

Meta information for each video is also provided in the dataset along with video sequences. The validation of the dataset was manually checked. This vast dataset has five subcategories [5].

1) General facial actions
2) Facial actions with object interaction
3) Single human actions
4) Human to object interaction
5) Multiple human actions

WVU Dataset

This dataset consists of 2 main divisions i.e. single human unit actions and single human prolonged composite actions. There are 12 videos for unit actions and 8 videos for composite actions available in this dataset. Jogging, clapping and nodding head are a few of these many action classes. To help understand the dataset better, WVU dataset also contains meta information and description of the videos. The camera resolution is 640x480 pixels for unit actions and 960x720 pixels for composite actions. Twenty frames per second are present in each video sequence. The data has been recorded from 8 cameras that are synchronized in time [19]. All eight camera views of nodding head have been exhibited in Figure 7.

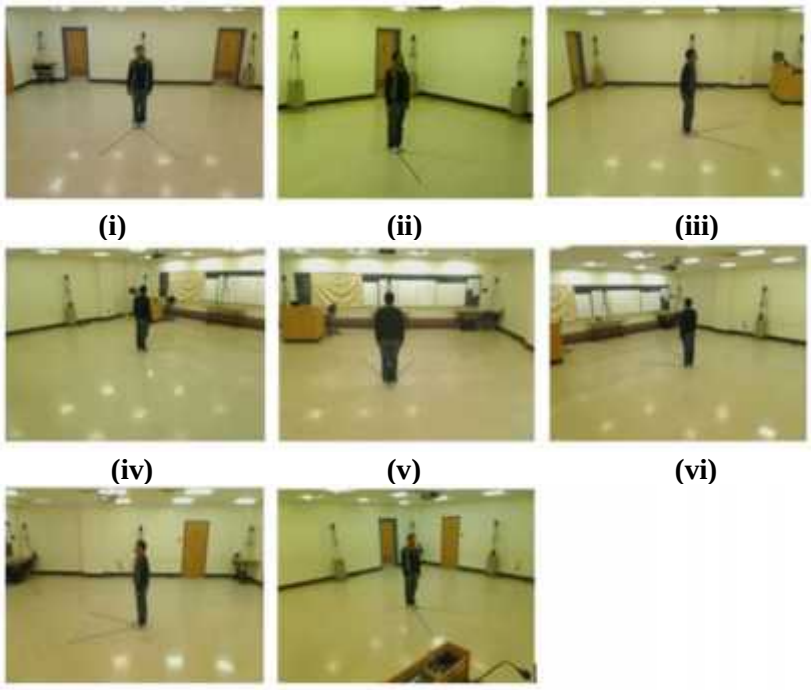


Figure 7: WVU dataset (8 camera views of nodding head)

MPII Dataset

This dataset comprises of 21 main categories (e.g. running, home activities, sports and others) which are further subdivided into 823 individual human action classes [20] such as running, jogging, sitting, exercising and so forth. These image frames and short video are taken mainly from YouTube. The dataset contains 24,920 videos in total. The dataset consists of a huge range of videos among which, some may consist of occluded human actions, and the camera viewpoint of these videos may also vary from one video to another. It is also possible that most videos may not show the complete silhouette of humans present [22]-[23], rather some portion of human body may be truncated in some or all of the frames in any video [21].

Weizmann Dataset

The dataset is composed of ninety videos with low resolution of 180x144[22]-[23]. Ten natural action classes are focused in this dataset which include running, walking, skipping, bending and others. Each of the ten actions was performed once by all nine actors [22]-[23]. This dataset includes videos without occlusions for most actions, however, walking action has been covered with object occlusions too, such as walking occluded by a bag, a log or a pole. In addition to this, moon-walk, limp-walk and normal walk captured from 10 different angles is also part of this dataset. Frames of 3 actions, randomly selected for illustration, have been shown in Figure 8. However, these varying walking styles have been included in the test data where training will be carried out by video sequences of normal walk. Extracted masks and background sequences are also part of this dataset.

(i) Jumping **(ii) Idling** **(iii) Skipping**

Figure 8: WEIZMANN dataset

CASIA Dataset

CASIA action database [24]-[26] for recognition includes 8 single person actions and 7 two-people interactions. The single human actions consist of:

(i) Bending
(ii) Crouching
(iii) Walking
(iv) Running
(v) Jumping
(vi) Fainting
(vii) Wandering
(viii) Punching a car

Two persons interaction instances consist of:

(i) Rob
(ii) Follow
(iii) Follow and gather
(iv) Meet and part
(v) Meet and gather
(vi) Fight and overtake

The resolution of video sequences is 320x240. All actions have been recorded with three camera views i.e. top-down view, angle view and horizontal view. The recordings have been made in a parking lot, an outdoor cluttered background, where frame to human ratio is large. All 3 views of five actions from CASIA dataset have been displayed in Figure 9.

(i) Walking

(ii) Running

(iii) Bending

(iv) Jumping

(v) Crouching

Figure 9: Three views of a few instances of CASIA Action dataset

Berkeley MHAD Dataset

The dataset consists of 11 actions performed 5 time by each of the 12 actors [27]-[29]. The actions include:

(i) Jumping
(ii) Jumping jacks
(iii) Throwing
(iv) Waving
(v) Clapping hands
(vi) Sitting
(vii) Standing
(viii) Bending

As an example, a sample frame of each of the eleven actions have been depicted in Figure 10. The dataset consists of diversity since no specifications were provided on how to perform a particular action [27]. The actions have been recorded in indoor environment. In addition to these actions, a T-pose of every subject is also provided in the dataset to estimate the structure of the subject. Moreover, background information is also available.

Figure 10: Berkeley MHAD dataset
(Some actions along with their point clouds obtained using a Kinect camera)

DISCUSSION AND SUMMARY

The collection of single human actions obtained from the datasets mentioned in this review can be categorized into 3 subdivisions i.e. sports and fights, body gestures and some common actions including both unit and composite actions. Sports and fights category include a summary of all those actions in these datasets that are related to either sports or fights and disputes e.g. climbing, punching, running etc. Gestures category is meant to familiarize the reader of all the instances in this collection which represents bodily gestures in one way or the other. This category includes crossing arms, nodding head, bending the body etc. The third category i.e. common actions include all those examples which cannot be linked to any particular class of actions, rather include commonly performed actions in daily routine e.g. walking, sitting, standing etc.

Following fact and figures are determined by comparing all datasets in Table 4.

From 2004 to 2010; KTH, MuHA Vi-MAS, IXMAS Action, WEIZM ANN, CASIA, UIUC and UT Tower were compiled.

- KTH dataset's background consists of indoor as well as outdoor; IXMAS Action & UIUC was captured indoor and in uniform environment and outdoor only for WEIZM ANN and CASIA. However MuHA Vi-MAS and UT tower were recorded outdoor.
- Illumination in KTH is grey-scale, MuHA Vi-MAS is Night Street Lights, IXMAS Action is Dim Light, UIUC is Moderate Light, WEIZM ANN, CASIA and UT Towers is Daylight.
- All of them have no occlusions except IXMAS Action, it contains occluded and non-occluded both videos.
- Resolution of KTH and UT Tower is very low; IXMAS Action and UIUC have normal resolution. However, MuHA Vi-MAS has high quality resolution with 720 pixels. Camera views are different for all of these datsets.
- KTH and WEIZM ANN have single camera View only; UT Tower has aerial view, MuHA Vi-MAS consists of 8 camera views, IXMAS Action is recorded with 5 different angles, CASIA has 3 and UIUC has only frontal view.
- KTH is sorted into 6 Classes with 25 subjects, MuHA Vi-MAS is in 17 classes with 4 subjects, UT tower is in 9 classes with 6 subjects, IXMAS action is in 11 classes with 10 subjects, WEIZM ANN is in 10 classes with 9 subjects, CASIA is with 8+7 with 24 classes and UIUC is in 14 classes with 8 subjects.

From 2011 to 2013; HMDB 51, WVU, MPII and Berkeley MHAD were collected.

- HMDB 51 and MPII were recorded in varying lightning condition while XVU and Berkeley MHAD in moderate light.
- Background conditions are varying for HMDB 51 and MPII whereas indoor for WVU and Berkeley MHAD.
- WVU and Berkeley MHAD has no occlusions but other 2 has different occlusions.
- Camera view for MPII and HMDB 51 is single. WVU has 8, and Berkeley MHAD has 12 views of camera.
- Class division of HMDB 51 is 51, WVU is of 12+8 classes with 1 subject, MPII of 823, and Berkeley MHAD of 11 classes with 12 subjects.

Moreover an additional row of "others" has been added in the tables to compensate all those actions that occur in a single dataset or are found to have rare occurrences in the discussed datasets. HMDB51 and MPII action datasets are very versatile and cover a large number of actions. The details of action events in datasets other than these two have been summarized in Tables 1-3. Moreover, a detailed tabular summary and analysis of these datasets have been presented in Tables 4 and 5 respectively.

Table 1: Presence of action instances related to sports category in datasets

Sports and Disputes										
	KTH	***MuHAVi -Mas***	***UT Tower***	***IXMAS***	***UIUC***	***Berkeley MHAD***	***CASIA***	***WVU***	***MPII***	***WEIZ MANN***
Climbing		Yes								
Punching				Yes			Yes	Yes		
Kicking		Yes		Yes				Yes		
Running	Yes		Yes	Yes	Yes	Yes	Yes		Yes	Yes
Jumping		Yes	Yes		Yes		Yes			Yes
Jogging	Yes							Yes	Yes	
Others (dive, somersault etc.)									Yes	

Table 2: Body gesture instances in datasets

Gestures									
	KTH	***MuHAVi -Mas***	***UT Tower***	***IXMAS***	***UIUC***	***Berkeley MHAD***	***CASIA***	***WVU***	***WEIZ MANN***
Cross arms				Yes					
Scratch head				Yes					
Nod head								Yes	
Bend						Yes	Yes		Yes
Waving	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes
Pointing				Yes					
Raising hand					Yes				

Table 3: Some common action instances in datasets

Common Actions										
	KTH	***MuHAVi -Mas***	***UT Tower***	***IXMAS***	***UIUC***	***Berkeley MHAD***	***CASIA***	***WVU***	***MPII***	***WEIZ MANN***
Walking	Yes		Yes	Yes	Yes		Yes		Yes	Yes
Crawling		Yes								
Clapping	Yes							Yes		
Sitting				Yes	Yes	Yes			Yes	
Standing			Yes	Yes	Yes	Yes		Yes		
Turning										
Pick up/ carry		Yes	Yes	Yes				Yes		
Drop/throw		Yes						Yes		
Two people Interactions							Yes			
Others (crouching, fainting etc.)				Yes		Yes			Yes	Yes

Table 4: A summary of datasets and their specifications

Datasets	Background	Illumination	Occlusions	Resolution	Camera views	Scale variations	No. of Videos	Size	Classes	Subjects	Format	Year of Publication	Ref
KTH	Indoors, Outdoors, Uniform	Grey-scale images	No	160x120	1	Y	2391	1.14 GB	6	25	AVI	2004	[1-4], [12], [13]
MuHAVi MAS	Outdoor, complex	Night street lights	No	720x256	8	N	952	171.8 GB	17	4	AVI	2010	[6], [7]
UT Towers	Outdoor, uniform	Daylight	No	Person height no more than 40 pixels	Ariel view	N	108	670 MB	9	6	BMP	2010	[15], [16]
IXMAS Action	indoors, Uniform	Dim light	Both	390x291	5 different angles	N	1148	7.4 GB	11	10	AVI	2006	[8], [9], [10]
UIUC	Indoor, uniform	Moderate light	No	400 pixels	Front view	N	532	57.7 GB	14	8	TIP	2008	[17], [18]
HMDB51	Varying	Varying light conditions	Rarely	Varying	1	Y	6766	5.5 GB	51	-	AVI	2011	[5], [18]
WVU	Indoor, Complex	Moderate light	No	640x480	8	N	6240	23.1 GB	12+ 8	1	JPEG	2011	[19]
MPII	Varying	Varying	Both	Varying	1	Y	24920	438 GB	823	-	-	2014	[20]. [21]
WEIZ MANN	Outdoor, uniform	Daylight	only in walking actions	180x144	Single with exception of 'walking'	N	90	340 MB	10	9	AVI	2005	[22] [23]
CASIA	Outdoor, complex	Daylight	No	320x240	3	N	-	-	8+7	24	-	2007	[24]-[26]
Berkeley MHAD	Indoor	Moderate light	No	640x480	12	N	56	822.8 GB	11	12	GRB G	2013	[27]-[29]

Table 5: A brief analysis of datasets

Datasets	Strengths	Weaknesses
KTH	Varying background	Poor resolution
MuHAVi MAS	Multi-camera views, silhouettes of human body are provided in MAS dataset	Frame to human height ratio is higher
UT Tower	Recorded from an aerial view	Large frame to human ratio results in poor resolution of the area of interest
IXMAS action	Occluded instances of videos have also been made part of the dataset	Poor resolution and blurred videos
UIUC	High resolution	Slight edges appear in background which might cause trouble in training
HMDB51	Versatility in actions	Inconsistent recording conditions
WVU	8 different camera views for each action, neat background	Edges in background might become problematic in training
MPII	Versatility in actions	Inconsistent recording conditions
WEIZMANN	Uniform background	Low resolution
CASIA	Unique classes have also been included	Frame to person ratio is large
Berkeley MHAD	Versatility in style of performing actions is a plus in testing	Versatility in style of performing actions becomes challenging in training

CONCLUSION

Most of the 11 datasets discussed in this review have managed to provide uniform backgrounds in their recordings, which is vital for training the system. On the contrast, some other datasets provide complex and real backgrounds, aiding in testing purposes. Relatively lesser number of instances of occluded actions is provided in this collection. Majority of these datasets provide annotations or bounding boxes around humans.

The mentioned datasets have a dominance of common actions performed in daily routine. However, it is observed that there is a lack of instances in context of surveillance for security systems in these online repositories. UT Tower and WEIZMANN datasets have low frame to person height ratio and KTH dataset comes with low resolution. Due to this reason, it may become tedious to utilize these datasets in body parts segmentation. HMDB51 and MPII comprise of videos based on real life human actions with multiple backgrounds, different resolutions, varying camera views and scale variation. These traits are significant when it comes to testing the accuracy of a well-trained system; however, they provide little benefit in training stage.

Analysis of these datasets reveals that there are negligible instances of locomotion-based actions that have been captured from a frontal or backside view i.e. approaching or moving away from a static camera. Moreover, no datasets of Asian people are available who are likely to have varying clothes and styles of performing chores.

REFERENCES

[1] G. Khan *et al.*, "Egocentric visual scene description based on human-object interaction and deep spatial relations among objects," *Multimed. Tools Appl.*, pp. 1–22, 2018.

[2] H. Naveed *et al.*, "Human activity recognition using mixture of heterogeneous features and sequential minimal optimization," *Int. J. Mach. Learn. Cybern.*, pp. 1–12, 2018.

[3] A. Klaser *et al.*, "A spatio-temporal descriptor based on 3d-gradients," in *Proc. BMVC 19th British Machine Vision Conference*, 2008, pp. 271–275.

[4] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[5] H. Kuehne *et al.*, "HMDB: A large video database for human motion recognition," *IEEE Int. Conf. Computer Vision (ICCV)*, 2011, pp. 2556–2563.

[6] S. Singh *et al.*, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *7th IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 48–55.

[7] M. A. Khan *et al.*, "An implementation of optimized framework for action classification using multilayers neural network on selected fused features," *Pattern Anal. Appl.*, pp. 1–21, 2018.

[8] D. Weinland *et al.*, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Underst.*, vol. 104, no. 2–3, pp. 249–257, 2006

[9] D. Weinland *et al.*, "Making action recognition robust to occlusions and viewpoint changes," in *European Conf. Computer Vision*, 2010, pp. 635–648.

[10] R. Pope, "A survey on vision-based human action recognition," in *Image and Vision Computing*, 2010, vol. 28, no. 6, pp. 976–990.

[11] E. C. Alp and H. Y. Keles, "A Comparative Study of HMMs and LSTMs on Action Classification with Limited Training Data," in *Proc. SAI Intelligent Systems Conf.*, 2018, pp. 1102–1115.

[12] J. M. Chaquet *et al.*, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Underst.*, vol. 117, no. 6, pp. 633–659, 2013.

[13] C. Schüldt *et al.*, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. Pattern Recognition*, 2004, vol. 3, no. September 2004, pp. 32–36..

[14] Y. Tian *et al.*, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Trans. Syst. Man Cybern. C-Applications Rev.*, vol. 42, no. 3, p. 313, 2012.

[15] C. Chen and J. K. Aggarwal, "Recognizing human action from a far field of view," in *Workshop Motion and Video Computing (WMVC)*, 2009, pp. 1–7.

[16] M. Z. Khan *et al.*, "Deep CNN Based Data-Driven Recognition of Cricket Batting Shots," in *Int. Conf. Applied Engineering Mathematics (ICAEM)*, 2018, pp. 67–71

[17] J. Liu *et al.*, "Recognizing human actions by attributes," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3337–3344.

[18] F. Naiar *et al.*, "A fixed-point estimation algorithm for learning the multivariate GGMM: application to human action recognition," in *Proc. IEEE Canadian Conf. Electrical & Computer Engineering (CCECE)*, 2018, pp. 1–4.

[19] V. Kulthumani *et al.*, "WVU multi-view action recognition dataset", 2012. Available: http//csee.wvu.edu/`vkkulathumani/wvu-action.html

[20] M. Andriluka *et al.*, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[21] L. Gorelick *et al.*, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[22] M. Blank *et al.*, "Actions as space-time shapes," in *10th IEEE Int. Conf. Computer Vision (ICCV'05)*, 2005, vol. 2, p. 1395–1402.

[23] Y. Wang *et al.*, "Human activity recognition based on r transform," in *IEEE Conf. Computer Vision and Pattern Recognition, (CVPR'07)*, 2007, pp. 1–8.

[24] Z. Zhang *et al.*, "Multi-thread parsing for recognizing complex events in videos," in *European Conf. Computer vision*, 2008, pp. 738–751.

[25] Z. Zhang *et al.*, "Trajectory series analysis based event rule induction for visual surveillance," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR'07)*, 2007, pp. 1–8.

[26] F. Ofli *et al.*, "Berkeley MHAD: A comprehensive multimodal human action database," in *IEEE Workshop Applications Computer Vision (WACV)*, 2013, pp. 53–60.

[27] F. Ofli *et al.*, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition" *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 24–38, 2014.

[28] R. Chaudhry *et al.*, "Bio-inspired dynamic 3d discriminative skeletal features for human action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2013, pp. 471–478.

[29] L. Pishchulin *et al.*, "Fine-grained activity recognition with holistic and pose based features," in *German Conf. Pattern Recognition*, 2014, pp. 678–689.