A Fine Grainedresearch Over Human Action Recognition

S. Sandhya Rani, G. Appa Rao Naidu, V. Usha Shree

Abstract: ***Human Action Recognition from videos has been an active research is in the computer vision due to its significant applicability in various real-time applications like video retrieval, human-robot interactions, and visual surveillance, etc. Though there are so many surveys over Human action Recognition, they are limited to various constraints like only focusing on the methods in few orientations only. Unlike the earlier ones, this paper provides a detailed survey according to the basic working methodology of Human action recognition system. Initially, a detailed illustration is given about various standard benchmark datasets. Further, following the methodology, the survey is accomplished in two phases, i.e., the survey over feature extraction approaches and the survey over action classification approaches. Further, a fine-grained survey is also accomplished under every phase based on the individual strategies***

Keywords: Human Action Recognition, Feature Extraction, Classification, Spatio-temporal interest points, Trajectories, Support Vector Machine, Deep Learning, action datasets.

I. INTRODUCTION

In recent years, with a widespread applicability in various applications,like visual surveillance, sports analyses, human-computer interactions (HCI) [1], robotics [2], elderly care [3], and intelligent space [4] and video retrieval, Human Action Recognition (HAR) has gained a significant research interest in the field of computer vision. The main aim of HAR is to determine, and then recognize what humans do in the unknown videos. Compared to the individual still images, video sequences provide more detailed information about actions. For example, consider a still image in which a man moving. We can't say what he is doing. Whether he is running? Or he is just jogging? In such a case, video sequences can give more detailed information about the human action. Video gives both spatial and temporal information but still,the image gives only spatial information.

(a)

(b)

Figure.1 (a) Still image and (b) frames of a video sequence

The main objective of a HAR system is to identify actions in a video sequence under different situations like occlusion, cluttering, and different lighting conditions. The main center of this system is the computational algorithms that understand the human actions. Similar to the human vision system, these computational algorithms ought to produce a label after the analysis of partial or entire action in the video sequence [5], [6]. Developing such algorithms is typically addressed in computer vision research, which studies how to make the computers to gain high-level understanding regarding human actions from digital images and videos?

A. Applications of HAR

HAR algorithms empower many real-world applications and these approaches remarkably reduce the manual effort to analyze a large-sized video data and provide sufficient understanding about the present end future states of an ongoing video data. The major real-world applications which utilize the action recognition are formulated as;

Visual Surveillance: In the Visual Surveillance system, a camera equipped with HAR algorithms may increase the chances of capturing a criminal on video, and decrease the dangerinstigated by illegal actions [7]. The cameras also make some people feel more secure, knowing the criminals are being watched.

Video Retrieval: Due to the vast number of videos on the internet, Video Retrieval is becoming a tremendous challenge as most search engines use the associated text data to manage video data [8]. The text data, such as tags, titles, descriptions, and keywords, can be incorrect, obscure, and irrelevant, making video retrieval unsuccessful. An alternative method is to analyze human actions in videos, as the majority of these videos contain such a cue.

Entertainment: In the entertainment field, the HAR has a significant importance due to its deployment in the sensors which are used to detect human actions [9]. These sensors provide an in-depth channel data which has encoded rich information about the entire scene of video.

Revised Manuscript Received on November 05, 2019.

S. Sandhya Rani, Assoc.Prof at MREC(A), Research Scholer JNTUH, Department of CSE. sarlanasandhya@gmail.com.

Dr. G. Appa Rao Naidu, Professor at JBIET, Department of CSE. apparaonaidug@yahoo.com

Dr.V. Usha Shree, Principal at JBREC, Department of ECE. valasani_usha1@yahoo.com

Human-Robot Interaction: Assume that a patient is undertaking a recovery exercise at house, and his/her robot assistant is capable of recognizing the patient's actions, analyzing the correctness of the exercise, and preventing the patient from further injuries. Such an intelligent machine would be significantlyhelpful as it hoards the rounds to visit the therapist, reduces the medical cost, and makes remote exercise into reality [10].

Autonomous Driving Vehicle: Action Recognition algorithms can predict a person's intention in a short period of time. In an emergency situation, a vehicle equipped with an action recognition algorithm can predict a pedestrian's future action or motion trajectory in the next few seconds, and this could be very helpful to avoid a collision [11].

This paper outlines a detailed literature survey over various methods proposed to recognize the human action in multiple environments with different strategies. This paper conducted the survey based on the basic methodology of HAR system, i.e., feature extraction and classification. In the feature exaction phase, the HAR system extracts a sufficient set of features and in the classification; the obtained features are processed for classifiers to recognize the action. Following the same strategy, initially, the survey is carried out over various feature extraction techniques and then over various classification techniques.

The remaining paper is organized as follows; Section II explores the details of various datasets used for simulation experiments. Section III Outlines the complete survey and section IV provides the concluding remarks.

II. DATASETS

Under this section, this paper describes the details of the datasets used for simulation experiments. In earlier, various datasets are developed and based on the development;the environment considered during the development, the appearance of objects, camera view, background variations and total number of subjects, they are categorized into two classes such as constrained and unconstrained.

2.1 Constrained Datasets

Under this category, the action videos are captured under constrained environments which have fixed settings. INRIA XMAS multi-view dataset, KTH dataset, Weizmann dataset, and UT-interactions dataset are some examples of this category. Further, the first three are individual action dataset and the remaining one isthe group action dataset.

A. KTH dataset [12]

This dataset consiststotal of six different actions such as *Handclapping, running, jogging, walking, hand waving and boxing.* The entire actions are accomplished various times under four different environments such as Outdoors, Indoors, Outdoors with various clothes and outdoors with scale variations. Totally 25 subjects are used to create this dataset and this hasa totally $25 \times 6 \times 4 = 600$ vides in.AVI format and 2391 sequences. All the sequences are captured with a static camera with 25fps frame rate and the background is homogeneous in nature. Some sample frames of this dataset are shown in fig.2.

B. Weizmann Dataset [13]

This dataset totally consists of ten different actions like *walking, running, skipping, jumping jack, jump forward on two legs, bend, single hand waving, doble hands waving, gallopsizeways, and jump in the same place with two legs,* and totally consist of 90 videos. all the actions are captured with the help of a static camera. The frame rate of every video of 50 fps ane the resolution of each frame is 180×144.Some sample frames of this dataset are shown in fig.3.

C. INRIA XMAS multi-view dataset [14]

This dataset consists of 12 action classes such as*point out, pick up, wave, punch, turn around, walk, sit down, get up, cross arms, scratch head and check watch*. Each action is performed three times and 12 different subjects are recorded with five cameras, four are fixed at four sides and one is fixe on the top. These five cameras capture five views such as left, right front back and top. The frame rate is 23 frames per second and the size of the frame is 390×291 pixels.Fig.4 shows some samples of different actions under multiple views.

D. UT-interactions dataset [15]

This dataset consists of a total of six human interactions such as *Push, Punch, Point, Kick, Hug, and Handshake*. There are totally 20 video sequences which have a length of around one minute. Several subjects with more than 15 different clothing conditions are captured in this dataset. The frame rate is 30fps and the size of the frame is 720×480 pixels. Further, the height of a person in every video is approximately 200 pixels. Fig.5 shows some samples of different actions.

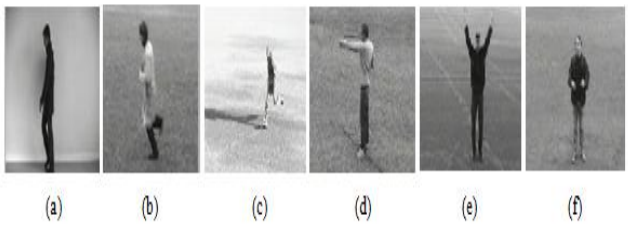


Fig.2 action samples of KTH dataset (a) Walking, (b) Jogging, (c) Running, (d) Boxing, (e) Handwaving, (f) Handclapping

Fig.3Action samples of Weizmann dataset, (a) Bend, (b) Jump, (c) Run, (d) Boxing, (e) Wave two hands, (f) Walk

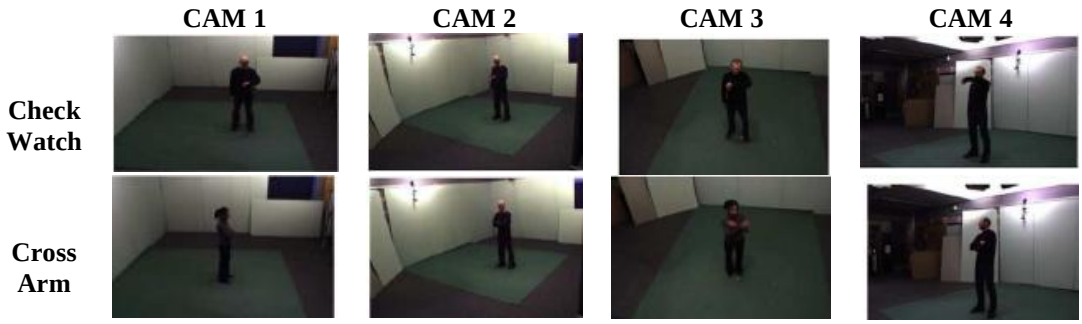


Figure. 4 Samples of INRIA XMAS Multi-View Dataset

Fig.5 Action samples pf UT-interaction dataset, (a) Push, (b) Punch, (c) Point, (d) Kick, (e) Hug and (f) Handshake,

2.2Unconstrained Datasets

Under this category, the action videos are captured under unconstrained environments which have no fixed settings, simply real-time environments. The datasets under this class are created by capturing the videos from the internet. Moments-in-time (MIT) dataset, UCF101 dataset, HMDB51 dataset are some of the examples of this unconstrained datasets.

A. Moments-in-time dataset [16]

This dataset containsa totally of 339 types of video clips, each one is of a 3-second duration and totally this dataset has 1,00,000 videos clips. The visual objects appear in the video clips are natural scenes, animals, people, and objects. This is a very large human-annotated dataset comprised of so many events and generally used for the recognition tacks sin complex environments. Fig.6 shows some samples of this dataset.

B. UCF101 dataset [17]

This dataset consists of a totally 101 action classes and the total number of videos is 13320. All the videos of this dataset are acquired from YouTube. With the presence of 13320 videos, the UCF101 dataset provides diversified actions with the presence of different variations such as illuminations, cluttered background, viewpoint, scale, pose and appearance, and variations in the motion of camera, etc. The complete 101 videos are categorized into 25 groups, where every group consists of 4-7 video for an action. Further, the actions are categorized into five classes and they are Sports, playing musical instruments, Human-Human Interactions, Human-Object interactions, and body motions. Fig.7 shows some samples of this dataset.

C. HMDB51 Dataset [18]

This dataset consists ofa totally 51 action classes and the total number of video clips is approximately 7000. Every class of action carries a minimum of 101 video clips. The total 51 action categories are grouped into five classes and they are body movements for human interactions, body movements for object interactions, general body movements, facial actions with the manipulations in an object, and the general facial actions. In this dataset, most of the videos are collected from different sources like Google videos, YouTube and several movies.Fig.8 shows some samples of this dataset.

Fig.6 Action samples of MIT dataset, (a) Bouncing, (b) Swimming, (c) Falling, (d) Opening,

Fig.7Action sample of UCF101 dataset, (a) Sky Diving, (b) Shaving Beard, (c) Apple Eye makeup, (d) Rafting, (e) Playing dhol

Fig.8 Action samples of HMDB51 dataset, (a) Climb Satirs, (b) fall floor, (c) draw sword, (d) flic flac, (e) had stand

Table.1 Comparison Details of various datasets

Dataset Name	Number of clips	Actions	Year	Environment
KTH	10	6	2004	Constarined
Weizmann	9	9	2005	Constarined
INRIA XMAS multi-view	390	13	2006	Constarined
UT-Interaction	60	6	2010	Constarined
HMDB51	101	51	2011	Unconstarined
UCF101	13,320	101	2012	Unconstarined
Moments In time	1,00,000	339	2017	Unconstarined
Hollywood[19]	30-140	8	2008	Unconstarined
MSR [20]	14-25	3	2009	Constarined
Sports1-M [21]	11,33,158	487	2014	Unconstarined

III. LITERATURE SURVEY

According to the methodology followed for Human Action Recognition, the entire literature survey is accomplished here under two phases, one is the approaches focused over the feature extraction and the other is the approaches focused over the classifier. In the case of feature extraction oriented approaches, the main focus is made on the feature extraction, i.e, the representation of a video or frame with an efficient set of features. An effective set of features provides a perfect discrimination from action to action such that the classifier will get clarity over the action sequences. Further, in the case of classification, the main focus is done on the reduction of complexity and the provision of robustness. A classifier must be like that it has to produce effective recognition results for all types of action sequences. In such a case, the classifier is said to be robust. Further, the recognition framework also concentrates on the computational complexity and mainly tries to choose a classifier which has a less computational complexity. The details of different earlier developed approaches are described in the below sub-sections.

3.1 Feature Extraction

Feature extraction is the foremost important problem in the human action recognition system. The human actions appearing in the video have so many variations due to the variations in the pose, camera movement, appearance, speed of motion, etc., makes the feature extraction really a most complex task. The main intention of a feature extraction technique is to provide perfect discrimination between action sequences, computationally efficient, and can effectively characterize the human actions, such that the recognition system should have less false positives or classification errors.One more major challenge in the human action recognition system is the large variations in the pose and appearance even in one action class, which results in more confusion for the recognition system. In this scenario, the main goal of the feature extraction is to nullify the variations and to convert the video into one feature vector such that it can provide sufficient discrimination between different forms of same human action, and minimize the variations such that the recognition performance will be improvised.

3.1.1 Local Features

These are extracted from local regions which have more salient information regarding the human action. Since the information present in the local regions is more salient and also more informative, most of the researchers focused on local feature only. Furthermore, the local features are more robust to variations in the appearance and translation, etc. Motion trajectory [22, 23, 55] and "space-time interest points (STIPs)" [24-26] are the two most popular local feature extraction techniques which had shown their superior performance in the HAR.

Wang et al. [23] proposed a new motion trajectory method by integrating the "Motion boundary Histogram (MBH)" [29] and "Histogram of Gradients (HoG)" [32] to extract a more effective and information-rich feature. In this approach, the trajectories are calculated based on the optical flow vectors [27]. Further, Jiang Y. G et al. [28] focused to integrate the global and local motion reference pints such that the obtained feature vector is compact to camera movement. To further improve the performance, Wang et al. [30] considered to predict the camera motion and performed the feature point matching between the frames using SURF descriptors. After generating the trajectories of both human action and camera motion, the trajectories which have inconsistent matches are removed. This estimation can also be used to remove the camera motion in optical flow vector-based approaches.

A similar action recognition technique is proposed by H. Wang et al. [33] based on the discovery of feature point matches between frames using SURF and optical flow vectors. In this approach, the homograph is measured and the trajectories which are inconsistent with homography are assumed to occur due to the camera motion and they are removed. Further, Hamim A. Abdul Azim and E. E Hemayed [31] proposed a new version of trajectory-based human action recognition system which captures the discriminative temporal relationships. In this approach, the trajectories are extracted based on STIPs named cuboid features and they are derived by matching it's SIFT descriptors over the successive frames. Then the obtained trajectory points are described in a visual "Bag-f Words (BoW)" model and then fed to "support vector machine (SVM)" classifier. However, the main drawback with cuboid featuresis that the obtained interest points may or may not locate at the same place in some frames which have cuboid temporal bounds.

Next, STIPs based approaches have gained the most popular due to its ease of implementation [34, 35].Bergonzi et al. [35]proposed to extract two different set of interest points. The first set is concerned about the shape and speed of foreground moving and to obtain these features, a 2-D Gabor filter is accomplished to every frame. Next, the second set of features is oriented to different scale and they are extracted from interest point clouds. Further, the method proposed by Willems et al. [36] used the Hessian matrix to derive STIPs which are scale-invariant both temporally and spatially.

Furthermore, some more 2-D detectors are also considered which are extended into the Spatiotemporal domain. In such regard, the HoG is extended to 3-D and formulated as HoG3D and applied to extract the STIPs in the method proposed by Kaiser et al. [37] to perform human action recognition. Next, a new approach is developed by N Li et al. [38] for HARfrom unconstrained videos by combining the HoG3D with Self-organizing map (SOM) which has a significant impact on the training parameters.Generally, the STIPs are applied over a grayscale image which makes the system sensitive to disturbing the photometric phenomena like shadows and highlights. Furthermore, the important information is also neglected by discarding the chromaticity. To solve these issues, Ivo Everts et al. [39] extended the STIPs to multiples channels, called color STIPs. These points improve the quality of subsequent STIP detection and description.

Further several 2-D descriptors are combined to develop new descriptors such that they can find the Spatio-temporal features more effectively. Towards this process, optical flow vectors are combined with histogram features and formulating into a new descriptor called “Histograms of Optical Flows (HOF)” [19]. Next, the Gradients are computed over optical flow vectors and formulated into a new descriptor called MBH for describing the motion trajectories [40]. Similarly, Nazir et al. [54] also proposed to integrate the 3D SIFT with 3D-Harris Spatio-temporal features to extract the key regions of a video. This approach used the conventional bag of visual word histograms for representing a human action.

3.1.2 Depth and Skelton Features

With the application of depth evaluation sensors (RGBD sensors), the human action detection approaches have gained improved recognition results due to the depth analysis of data. These approaches are further classified as depth sequence-based approaches [41, 42, 47-50], and Skelton based approaches [43-46, 51]. These approaches use the basic global and local features s well to extract a composite feature vector from every action sequence.

In the depth-based feature extraction approaches, initially, the motion changes are analyzed through the depth map of the human body. Under this class, a video captured through an RGBD sensor is seen as a space-time structure having the depth information. For a given action sequence, the feature is extracted as a Spatio-temporal feature with a motion or an appearance having changes in the depth.Xiaodong Yang et al. [47] proposed a new action recognition technique based on the additional depth information, i.e., body shape and motion information. In this approach, the depth maps are projected into three orthogonal planes and then accumulate global activities to generate “Depth Motion Maps (DMM)”. Further HoGs are computed form DMM as a feature. The sampled actions of DMMS for different actions are exposed in figure.9.

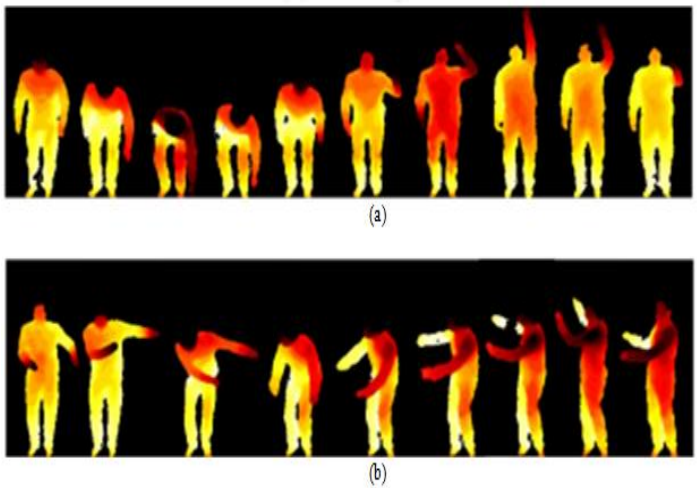
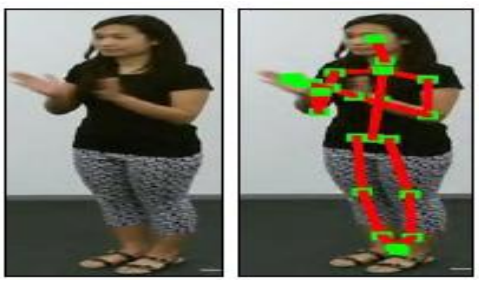
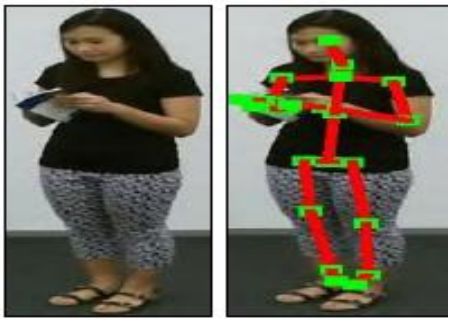


Figure.9 Sampleactions of depth maps of (a) Pick up & Throw (b) Golf swing [47]

Next, Omar Oreifej and Zicheng Liu [42] proposed a novel descriptor, called as “histogram of oriented 4D surface normals (HON4D)” for HAR from depth maps. This descriptor is equaledto the HoGs in color action sequences and spreads the histogram of normal in static images. H. Rahman et al. [48] focused to integrate the 3D joint positions and discriminative information from depth images. For every joint, a composite feature vector, called “3D space time-motion volume” is calculated to provide a perfect discrimination between actions. Further, the method proposed by C. Chen at el. [49] developed a multi-fusion based action recognition technique based on the DMM and “Local Binary Patterns (LBPs)”. This approach employs DMMs for three projection views (top, side, and front) to capture the motion cues and then applies LBPs to extract a composite feature for every action. This approach accomplished two fusion phases; one is feature level fusion and the other is decision level fusion.Jie Miao et al. [50] proposed to consider the compressed depth maps for action recognition. In this approach, every depth map is encoded with a scalable encoder which has multi-scale breakpoints and ad Adaptive “Discrete Wavelet Transform (DWT)”. Here the sharp edges are obtained through breakpoints and the smooth variations are obtained through DWT and are extracted from the bit-stream and are used to construct a set of features that are fed to classifier for recognizing the action.

From depth information of a human body, the Skelton can be estimated more easily. The samples of skeletons for some actions are shown in figure.10.





(a) (b)

Figure.10 original and skeletons of different actions (a) clapping (b) reading [52]

J. Shotton et al.[51] focused to estimate the body position of 3D joints effectively and quickly even from only one depth frame without seeking any help of temporal data. The joint consideration has been regarded as a simple super-pixel problem.Papadopoulos, G.T et al. [43] proposed a real-time tracking approach for HAR. In this approach,first, the skeleton is tracked and then applied a new representation mechanism for action based on the calculation of spherical angles between the joints and corresponding angular velocities. Finally, the recognition is done with the help of "Hidden Markov Models (HMMs)".

Unlike the above approaches which consider the joint angles and joint locations to represent the human Skelton, V. Raviteja et al. [46]developed a new method to represent the human skeletal based on the evaluation geometric relationships between the parts of body based on translations and rotations in 3D space. This approach modeled the human action as a curved manifold. Further, the classification is with an integration of "dynamic Time warping (DTW), Fourier temporal pyramid representation" and linear SVM. Pazhoumanddar et al. [45] proposed to consider a skeleton-based number of action descriptors for action recognition. This approach used the longest common subsequence (LCSS) algorithm to select high-discriminative power features fromthe relative motion trajectories of the skeleton to determine the associated action.Hany ElGhaith et al. [52] proposed to combine three different modalities such as body part images, 3D skeletons, and "Motion History Image (MHI)" into a deep learning mechanism for HAR. Based on the three different modalities, the entire information like pose of the body, motion of body and part shape can be extracted. Since the 3D skeleton can't acquire the shape body and also the shape of manipulated objects, MHI [53] and body parts are included in the feature extraction process.

Based on the above discussion, the depth and skeleton-based action recognition approaches are efficient in the representation of joint features which are more important in describing an action. Hence, more discrimination cane be acquired by the system form these features. But, the performance of these approaches completely depends on the prediction of a human pose. Furthermore, this approach has a severe effect in the presence of occlusions in the scene which results in heavy classification errors.

3.2 Action Classification

After extracting features from video sequences, they are processed for learning through action classifiers such that classifier will get sufficient knowledge about the actions present in the video. Further, this knowledge helps to determine the class label of various action classes.

3.2.1 Traditional classifiers

Among various available traditional classifiers, Support Vector Machine [54, 56, 58, 59, 60, 62, 63]and random forest [65, 66] has gained an excellent performance in the classification of different actions for all types of action videos.

K. G. C. Manocha, R. Rodrigo [56] accomplished the SVM algorithm for action recognition. In this approach, the Optical flow values of a silhouette are extracted as features and based on these features, a new motion descriptor id described for every action. Further, the SVM is accomplished for classification. To reduce the high dimensional feature space, this approach used Principal component analysis.Simulation is done over two datasets namely, Weizmann and UIUC1 Dataset [57].

Jalal A. Nassiri et al. [58] introduced a version of SVM, called "Energy-based Least squares Twin Support Vector Machine (ELS-TSVM)". This is an extension to the conventional "Least Square Twin Support Vector Machine (LS-TSVM)" [59] which performs the classification based on two non-parallel hyperplanes instead of a single hyperplane. This approach effectively handles the unbalanced dataset problem. The performance of this classifier is tested over KTH, Weizmann, Hollywood and UCI datasets.

L. Gonzalez et al. [60] considered the "Multi-class Support Vector Machine (MC-SVM)" to classify the human actions, using a multi-camera dataset called MUHAVI dataset [61]. Silhouettes are extracted as features for every action. Since the videos of the MUHAVI dataset noisy and also comprises shadows, this approach considered MC-SVM to achieve effective classification results.

M D. Praveen and C K Niranjan [62] used the SVM classifier for action recognition after extracting the features for a given video sequence. Simulation experiments are conducted over KTH dataset with hand-clapping and running actions.

Y Wang e al. [63] combined the 3D skeleton joints, gesture potential energy, and kinetic energy and others to extract one feature matrix. Further, K-means clustering is accomplished for the extraction of semantic features by the BoW. This combination of features not only reveals the information about the kinematics but also explores the biology of the human body and the natural visual saliency. Finally, the SVM kernels are used to accomplish the HAR.

SamehMerghi et al. [64] accomplished the SVM classifier for action recognition after detecting the moving human sin moving field of views based on optical flow and dense SURF. After SURF extraction, the video is represented through a set of fused features that combines visual descriptors, trajectory, and motion and finally exploited the BoW approach. Experiments are conducted on the standard datasets such as KTH and UCF101 and HMDB51 datasets. H. Zhang et al. [65] developed an action recognition method which combines the sparse coding with gradient information. In this approach, initially, the depth of gradient information and distance between the joints of the 3D skeleton are extracted to find the coarse depth-skeleton features. Next, the sparse coding and max pooling are

applied to finalize a coarse DS feature and fed to random decision forest classier to perform action recognition.

Vikas Tripathi et al. [66] proposed to use the random forest algorithm for human action recognition after extracting a novel feature descriptor based on two algorithms They are distance mean histogram of gradient and segmented block of means image with normalization. The performance evaluation is accomplished over two standard benchmark datasets and they are ATM and HMDB.

3.2.2Deep Learning classifiers

In recent years, the application of deep learning to computer vision applications has gained more importance due to its deep studying strategy. In the human action recognition also, various deep learning strategies are developed by researchers. Based on the deep learning network structure applied for action recognition, the earlier developed approaches are formulated into three different types and they are 2D convolutional networks (2D-CNNs) [69, 72] based, and 3D convolutional networks (3D-CNNs) [68] based approaches. In the 2D CNNs based HAR approaches, both the image and its optical flow information are trained to the network in the training phase and at the output, the fusion process is applied at the output layer. The 2D CNN follows 2D image convolution while 3D CNN follows 3D convolution which is the major difference between these two methods.

3.2.2.1 2D CNNs

C. Feichtenhofer et al. [70] proposed a new ConvNet architecture for spatiotemporal fusion of video action recognition. This approach has the following modifications compared to the tradition ConvNet. 1. Fusion is accomplished at spatial and temporal network instead of a softmax layer. 2. Spatial fusion is done at the last layer. 3. Pooling is done over Spatio-temporal neighborhoods. Simulation experiments are conducted over two standard datasets such as UCF101 and HMDB51. Next, Limin Wang et al. [71]proposed a new ConvNet architecture based on the long-range temporal structure modeling, called “temporal segment network (TSN)”. This approach combined the video-level supervision and sparse temporal sampling strategy to provide efficient learning to the network with an entire action video. This approach is effective even with few learning samples. Simulation experiments are conducted over two standard datasets such as UCF101 and HMDB51.

B Zhou et al. [73] developed a new “Temporal Relation Network (TRN)” into the CNNs for HAR. TRN is an effective and interpretable network module which was intended to learn and reason about the temporal dependencies between frames of a video. Further, the TRN is simulated over three recent video datasets; they are something-Something [74], Jester [75] and Charades [76]. Further, C. B. Jin et al. [77] proposed a hierarchical CNN model for real-time HAR based on the temporal images. Under the hierarchical model, this model has three CNN layers, namely, posture layer, motion layer, and action layer. Simulation experiments are conducted over the “Imperial Computer Vision and Learning Lab (ICVL)” action dataset [78] over the standing, walking, nothing, and texting action sequences.

3.2.2.2 3D CNNs

Shuiwang Ji et al. [79] developed 3D CNN model for HAR. In this approach, the features are extracted from both temporal and spatial dimensions by applying 3D convolutions such that more encoded information can be acquired from adjacent frames of a video sequence. Initially, the multi-channel information is acquired from all the frames and then the final feature is obtained by combining the information from all the channels. Simulation experiments are conducted over TREC Video Retrieval Evaluation(TRECVID) 2008 dataset [80]. Later Du Tran et al. [81] developeda novel Spatio-temporal feature-based learning approach using 3D CNNs. In this approach, the learned features, namely C3D (convolutional 3D) with a simple linear classifier is far better than the conventional 2D CNNs. Further to check the performance, totally three datasets are used and they are UCF101, ASLAN dataset [84], YUPENN [82] and Maryland [83]. Further, J.Arun Nehru et al. [85] considered the 3D motion cuboid features for action recognition through 3D convolution neural networks. Simulation experiments are conducted over a standard Weizmann and KTH datasets.

3.3 Results and Discussion

From the above discussion, we can understand that the approaches proposed based on deep learning are more effective compared to the handcrafted feature-based approaches. Though they offer a superior performance, some issues exist, particularly the fusion of multi-modal data in the deep learning architecture.Most of the deep learning approaches focused on the fusion of multi-modal data and thereby realizing the concept of deep learning. For any system, the deep knowledge will be acquired when it has all possible information. For example, the skeleton data of an action explores only the skeletal information. Next, the statistical measures explore only the rotational and translational variations in the action data, and the RGB data explores only the multi-channel information with respect to three channels in the RGB action video. Compared to individual features, a combined feature gives more detailed information and helps very much in the accurate recognition of an action. The main problem assisted with the feature fusion in the computational complexity. Furthermore, an effective combination of multimodal data such as skeleton data, depth data, optical flow, and RGB data still remains an open issue for human action recognition. This issue gives better direction of research in the field of HAR.

A graph is provided in Fig.1, wi20th two parameters such as methodologies accommodated and the accuracy obtained with the datasets.

$$\boldsymbol{F-score}= ((2 * P * R))/(P + R)$$

$$\boldsymbol{Precision} = \frac{No.of\ instances\ of\ correct\ positive\ recognition}{Total\ No.of\ positive\ recognition}$$

$$\boldsymbol{Recall} = \frac{No.of\ instances\ of\ positive\ recognition\ found}{Total\ No.of\ relevant\ input\ instances}$$

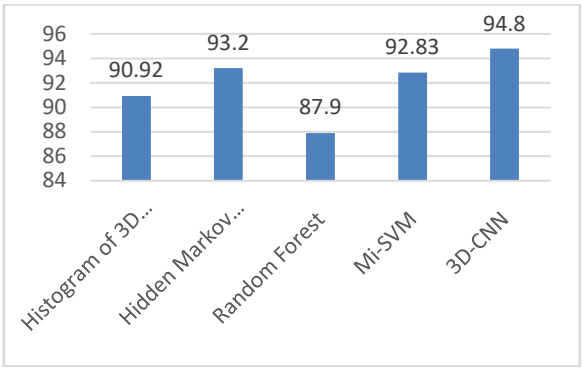


Fig.1 Accuracy of various Algorithms

IV. CONCLUSION

In this paper, a detailed and comprehensive survey is carried out over various human action recognition techniques. Initially, an introduction to HAR is explored and in that the need for human action recognition is illustrated. Further detailed information is provided about various datasets. Further, a detailed survey is described and the complete description is done under two phases, one is the survey over various feature extraction techniques and the other is on various action classification techniques. Under the feature extraction techniques, this paper reviewed the local features, depth, and skeleton features. Next, under the action classification techniques, this paper reviewed traditional classifications techniques and also the deep learning strategies. Further, a fine-grained analysis is accomplished over the deep learning approaches with respect to the convolution dimension. Based on the review explored, this paper makes the following conclusions.

1. To perform effective human action recognition, first, the feature extraction must be more effective. Since a system with more detailed information can only recognizes the action even under occlusions, noisy and complex backgrounds. For this purpose, the feature fusion will get priority and need an effective combination.
2. Though the feature fusion gives more prominent results in action recognition, there will be an excessive computational complexity at the classifier. Definitely, the complexity is more for an HAR system which analyzes the data in multiple views than the HAR system which analyzes the data in only one point of view. This can be compensated by an effective classifier design which is also more important in the HAR system.

REFERENCES

1. YangY,AloimonosY., 2015, "Robot learning manipulation action plans by watchingunconstrained videos from the world wide web", pp.3686–3693.
2. Kru¨gerV, KragicD, 2007,"The meaning of action: a review on action recognition andmapping", *Advanced Robotics* 21 (13),pp.1473–1501.
3. NiitsumagM., Hashimoto H, 2007, "Spatial memory as an aid system for human activity inintelligent space",pp.1122–1131.
4. Khan Z. A.Sohn W, 2011, "Abnormal human activity recognition system based on r-transform and kerneldiscriminant technique for elderly home care",pp.1843–1850.
5. Bobick A. and Davis J, 2001,"The recognition of human movement using temporal templates", pp. 257–267.
6. RyooM. S., 2011, "Human activity prediction: Early recognition of ongoing activities from streaming videos", pp.1-5.
7. FeichtenhoferC, PinzA, 2017, "Spatiotemporal multiplier networks for video action recognition", 7445–7454.
8. RamezaniM. and YaghmaeeF, 2016, "A review on human action analysis in videos for retrieval applications", pp. 485–514.
9. Xia L. andAggarwal J,2013,"Spatio-temporal depth cuboid similarity features for activity recognition using depth camera".
10. Koppula H. S. and Saxena A,2016, "Anticipating human activities using object affordances for reactive robotic response",pp. 14–29.
11. Li K. and Fu Y, 2014, "Prediction of human activity by discovering temporal sequence patterns",pp.1644–1657.
12. C. Sch¨uldt, I. Laptev, and B. Caputo Recognizing human actions: Alocal SVM approach,*in IEEE ICPR*, 2004.
13. Blank M., Gorelick L., ShechtmanE, 2005, "Actionsas space-time shapes",*in Proc. ICCV*, 2005.
14. Weinland D, RonfardR, andBoyer E, 2006,"Free viewpoint action recognitionusing motion history volumes",page no. 2-3.
15. RyooM. S. and. K J. Aggarwal, 2010, "UT-Interaction Dataset, ICPRcontest on Semantic Description of Human Activities (SDHA)",http://cvrc.ece.utexas.edu/SDHA2010/HumanInteraction.
16. MonfortM., Zhou B, S. A. Bargal, T. Yan, "Momentsin time dataset: one million videos for event understanding".
17. Khurram Soomro A. R. Z. and Shah M, 2012, "Ucf101: A dataset of 101human action classes from videos in the wild".
18. Kuehne H., Zhuang H., and Serre T, 2011,"HMDB: Alarge video database for human motion recognition".
19. Laptev I., MarszalekM, and RozenfeldB,2008,"Learningrealistic human actions from movies".
20. W. Li, Z. Zhang, and Z. Liu, Action recognition based on a bag of 3dpoints,*in CVPR workshop*, 2010.
21. KarpathyA, TodericiG, Shetty S., and L. Fei-Fei, 2014, "Large-scale video classification with convolutional neuralnetworks",*in CVPR*, 2014.
22. H.Wang, A. Klaaser, C. Schmid, and C. L. Liu, "Dense trajectories andmotion boundary descriptors for action recognition", *IJCV*, vol. 103, no.60-79, 2013.
23. H. Wang, A. Kl¨aser, C. Schmid, and C.-L. Liu, "Action Recognitionby Dense Trajectories",*in IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs*, United States, Jun. 2011, pp.3169–3176.
24. I. Laptev and T. Lindeberg, "Space-time interest points",*in ICCV*, 2003,pp. 432–439.
25. M. Bregonzio, S. Gong, and T. Xiang "Recognizing action as clouds ofspace-time interest points",*in CVPR*, 2009.
26. A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-temporal descriptorbased on 3d-gradients",*in BMVC*, 2008.
27. Cutler, R.; Turk, M. "View-based interpretation of real-time optical flow for gesture recognition". *In Proceedings*
28. *of the Third IEEE International Conference on IEEE, "Automatic Face and Gesture Recognition"*, Nara, Japan,14–16 April1998; pp. 416–421.
29. [28] Jiang, Y.G.; Dai, Q.; Xue, X.; Liu, W.; Ngo, C.-W. "Trajectory-based modeling of human actions with motionreference points". *In European Conference on Computer Vision*; Springer: Berlin, Germany, 2012; pp. 425–438.
30. Dalal, N., Triggs, B.: "Human detection using oriented histograms of flow and appearance". *In: ECCV*. (2006)
31. H. Wang and C. Schmid, "Action recognition with improvedtrajectories",*in IEEE International Conference on Computer Vision*,Sydney, Australia, 2013.
32. Hamim A. Abdul Azim and E. E Hemayed, "Human action recognition using trajectory-based representation", *Egyptian informatics journal*, Volume 16, Issue 2, July 2015, Pages 187-198.
33. Dalal, N.; Triggs, B. "Histograms of oriented gradients for human detection". In Proceedings of the CVPR2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA,20–26 June 2005; Volume 1, pp. 886–893.
34. H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficientvideo representation for action recognition",*IJCV*, 2015.
35. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognitionvia sparse Spatio-temporal features",*in ICCV VS-PETS*, 2005.
36. M. Bregonzio, S. Gong, and T. Xiang "Recognizing action as clouds ofspace-time interest points",*in CVPR*, 2009.
37. G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariantSpatio-temporal interest point detector",*in ECCV*, 2008.
38. A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-temporal descriptorbased on 3d-gradients",*in BMVC*, 2008.

39. N Li, Xu Cheng, S. Zhang, Zhejiang Wu, "Realistic human action recognition by Fast HOG3D and self-organization feature map", 2014, pp.1793-1812.
40. Ivo Everts, Jan C. van Gemert and Theo Gevers, "Evaluation of Color STIPs for Human Action Recognition", 2013.
41. H. Wang, M. M. Ullah, A. Kl aser, I. Laptev, and C. Schmid, "Evaluationof local Spatio-temporal features for action recognition",*in BMVC*,2008.
42. Ye, M.; Zhang, Q.; Wang, L.; Zhu, J.; Yang, R.; Gall, J. "A survey on human motion analysis from depthdata", October 2012; pp. 149–187.
43. Oreifej, O.; Liu, Z. HON4D: "Histogram of Oriented 4D Normals for Activity Recognition from DepthSequences",June 2013, pp. 716–723.
44. Papadopoulos, G.T.; Axenopoulos, A.; Daras, P. "Real-time skeleton-tracking-based human action recognition using Kinect data", January 2014; pp. 473–483.
45. Presti, L.L.; Cascia, M.L. 3D "Skeleton-based Human Action Classification: A Survey". *Pattern Recognit*. 2016,53, 130–147.
46. Pazhoumanddar, H.; Lam, C.P.; Masek, M. Joint "movement similarities for robust 3D action recognition using skeletal data". *J. Vis. Commun. Image Represent*. 2015, 30, 10–21.
47. Vemulapalli, R.; Arrate, F.; Chellappa, R. "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group", June 2014; pp. 588–595.
48. Miao, J.; Jia, X.; Mathew, R.; Xu, X.; Taubman, D.; Qing, C. "Efficient action recognition from compresseddepth maps", September 2016; pp. 16–20.
49. Chen, C.; Jafari, R.; Kehtarnavaz, N. "Action Recognition from Depth Sequences Using Depth MotionMaps-Based Local Binary Patterns", January 2015; pp. 1092–1099.
50. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A. "Real-time action recognition using histograms of depth gradients and random decision forests", March 2014; pp. 626–633.
51. Yang, X.; Zhang, C.; Tian, Y.L. "Recognizing actions using depth motion maps-based histograms oforiented gradients", November 2012; pp. 1057–1060.
52. Shotton, J.; Sharp, T.; Fitzgibbon, A.; Blake, A.; Cook, M.; Kipman, A.; Finocchio, M.; Moore, R. "Real-Timehuman pose recognition in parts from single depth images". *Commun.* ACM 2013, 56, 116–124.
53. Hany ElGhaish, Mohamed E. Hussien, Amin Shoukry, andRikioOnai, "Human Action Recognition Based on IntegratingBody Pose", Part Shape, and Motion, *IEEE Access*, Volume 6, 2018, pp. 49040-49055.
54. S. Jetley and F. Cuzzolin,"3D activity recognition using motion historyand binary shape templates", Springer, 2014, pp. 129-144.
55. Nazir, S.; Yousaf, M.H.; Velastin, S.A. "Evaluating a bag-of-visual features approach using Spatio-temporalfeatures for action recognition",2018.
56. Gaidon, A.; Harchaoui, Z.; Schmid, C. "Activity representation with motion hierarchies". *Int. J. Comput. Vis*.2014, 107, 219–238.
57. K. G. ManoshaChathuramali, Ranga Rodrigo, "Faster Human Activity Recognition with SVM", *The International Conference on Advances in ICT for Emerging Regions – ICTer 2012*: 197-203.
58. D. Tran, A. Sorokin, and D. Forsyth, "Human activity recognitionwith metric learning", 2008, pp. 549–562.
59. Jalal A.Nasiri, NasrollahMoghadamCharkari, KouroshMozafari, "Energy-based model of least squares twin Support Vector Machines for human action recognition", November 2014, Pages 248-257.
60. M Arun Kumar, M Gopal, "Least squares twin support vector machines for pattern classification", May 2009, Pages 7535-7543.
61. L Gonzalez, S.A. Velastin, G. Acuna, "silhouette based human action recognition with a multi-class support vector machine", 2018 page (5 pp.)
62. A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to Human Behavior Analysis for Ambient-Assisted Living", *Expert Systems with Applications*, vol. 39, no. 12, pp. 10873–10888, 2012.
63. M D. Praveen and C K Niranjan, "Human Action Detection and Recognition Using SIFT and SVM", *International Conference on Cognitive Computing and Information Processing*, CCIP 2017, pp.475-491.
64. YongxiongWang, YuboShiGuoliangWei, "A novel local feature descriptor based on energy information for human activity recognition", *Neurocomputing*, Volume 228, 8 March 2017, Pages 19-28.
65. SamehMegrhi, MarwaJmal, WidedSouidene, AzeddineBeghdadi, "Spatio-temporal action localization and detection for human action recognition in big dataset", November 2016, Pages 375-390.
66. HanlingZhang, PingZhong, JialeHe, ChenxingXia, "Combining depth-skeleton feature with sparse coding for action recognition", *Neurocomputing*, Volume 230, 22 March 2017, Pages 417-426.
67. Vikas Tripathi, DurgaprasadGangodkar, Ankush Mittal, Vishnu Kanth, "Robust Action Recognition framework using Segmented Block and Distance Mean Histogram of Gradients Approach",August 2017, Cochin, India.
68. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T,"Long-Term Recurrent Convolutional Networks for Visual Recognition and Description",*IEEE Trans. Pattern Anal*. 2017, 39, 677–691.
69. Tran, D.; Bourdev, L.D.; Fergus, R.; Torresani, L.; Paluri, M,"Learning Spatiotemporal Features with 3DConvolutional Networks". *In Proceedings of the IEEE International Conference on Computer Vision* 2015, Las Condes, Chile, 11–18 December 2015; pp. 4489–4497.
70. Simonyan, K.; Zisserman, A. "Two-Stream Convolutional Networks for Action Recognition in Videos", December 2014; pp. 568–576.
71. Feichtenhofer, C.; Pinz, A.; Zisserman, A. "Convolutional Two-Stream Network Fusion for Video ActionRecognition". *In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*(CVPR), Las Vegas Valley, NV, USA, 27–30 June 2016; pp. 1933–1941.
72. Wang, L.; Xiong, Y.;Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. "Temporal Segment Networks: TowardsGood Practices for Deep Action Recognition". *In Proceedings of ECCV*; Springer: Cham, Switzerland, 2016;pp. 20–36.
73. Lan, Z.; Zhu, Y.; Hauptmann, A.G.; Newsam, S. "Deep Local Video Feature for Action Recognition".*In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW*), Honolulu, HI, USA, 21–26 July 2017; pp. 1219–1225.
74. Zhou, B.; Andonian, A.; Torralba, A. "Temporal Relational Reasoning in Videos". *arXiv2017*, arXiv:1711.08496.
75. Goyal, R., Kahou, S., Michalski, V., Materzy´nska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: "The something-something video database for learning and evaluating visual common sense". *Proc. ICCV* (2017)
76. "Twentybn jester dataset: a hand gesture dataset".https://www.twentybn.com/datasets/jester (2017)
77. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: "Crowdsourcing data collection for activity understanding". *In: European Conference on Computer Vision*, Springer (2016) 510–526.
78. Cheng Bin Jin, Shengzhe Li,Trung Dung Do, and Hakil Kim, "Real-Time Human Action RecognitionUsing CNN Over Temporal Images for StaticVideo Surveillance Cameras", *Springer International Publishing* Switzerland, pp. 330–339, 2015.
79. T.H. Yu, T K. Kim, and R. Cipolla, 2013, "Unconstrained Monocular 3D Human Pose Estimation by Action Detection and Cross-modality Regression Forest".
80. Ji Shuiwang, XuWei, Yang Ming, Member, IEEE, and Kai Yu, "3D Convolutional Neural Networks for Human Action Recognition".
81. Huang Xin, WeiGang, and AValery. "PetrushinShot Boundary Detection and High-level Features Extraction for the TREC Video Evaluation", 2003, *Accenture Technology Laboratories*.
82. Tran D., BourdevL., Fergus R., L. Torresani, and M. Paluri. "Learning spatiotemporal features with 3d convolutional networks", December 2015.
83. K. Derpanis, M. Lecce, K. Daniilidis, and R. Wildes. "Dynamic scene understanding: The role of orientation features in space and time in scene classification". *In CVPR*, 2012.
84. N. Shroff, P. K. Turaga, and R. Chellappa. "Moving vistas: Exploiting motion for describing scenes". *In CVPR*, 2010.
85. Orit Kliper-Gross, Tal Hassner, Lior Wolf, "The Action Similarity Labeling Challenge", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. X, No. X, X 201X.
86. J.Arunnehru, G.Chamundeeswari, S. PrasannaBharathi, "Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos", *Proceida computer science*, Volume 133, 2018, Pages 471-477.