A Survey on Vision Based Activity Recognition, its Applications and Challenges

Ashwin Geet D'Sa
Department of Computer Science
BMS College of Engineering
Bengaluru, India
ashwin.dsa@gmail.com

Dr. B G Prasad
Department of Computer Science
BMS College of Engineering
Bengaluru, India
drbgprasad@gmail.com

***Abstract*—Tracking objects is gaining momentum in the field of research due to its widespread applications. Such systems can be enhanced or made more meaningful by allowing the activities of those objects to be recognized. Activity recognition is the ability to identify and recognize the action or goals of the agent. The agent can be any object or entity that performs action, which has end goals. The agents can be a single agent performing the action or group of agents performing the actions or having some interaction. Human activity recognition has gained popularity due to its demands in many practical applications such as entertainment, healthcare, simulations and surveillance systems. Vision based activity recognition is gaining advantage as it does not require any human intervention or physical contact with humans. Moreover, there are networked set of cameras, which makes it both easier to track and recognize the activities. Through this work, approaches involved in activity recognition, advantages and challenges involved in activity recognition are discussed.**

***Index Terms*—activity recognition, computer vision, video processing, classification**

I. Introduction

Activity recognition is the ability to identify and recognize the action or goals of the agent, the agent here can be any object or entity that performs action, which has end goals. The agents can be a single agent performing the action or group of agents performing the actions or having some interaction. One such example of the agent is human itself, and recognizing the activity of the humans can be called as Human Activity Recognition (HAR) [1]. In the last few years, automatic recognition of human activities has gained much attention in the field of vision based technologies due to its increasing demands in practical applications, such as surveillance environments, healthcare systems and entertainment environments. In a surveillance system, the automatic identification and classification of unusual and abnormal activities can be made. This would aid in alerting the concerned authority or person monitoring the given environment, for example group attacks or fights can be recognized and the concerned authority can be informed about it. In systems belonging to entertainment environment, activity recognition can be used for Human Computer Interaction(HCI) systems, which would involve identifying the activity of the person and responding to the activity of the actor. For example, we can use this in simulation of the game, where the human remains one of the players and the computer responds to the action of human player by simulating the computer based animations. In a healthcare system, the recognition of activities can help in activities such as rehabilitation of patients, where the activities of the patients can be automatically monitored such that the patient is undergoing the rehabilitation process. Human activity recognition is not just limited to a few of these applications, but also being used in various other applications.

A. Types of Activity Recognition based on the Device Used

The types of activity recognition based on the devices used can be classified as- Sensor based activity recognition and Vision based activity recognition.

1) Sensor based activity recognition: It uses a network of sensors to monitor the behavior of the actor, where some systems also monitor the actors surroundings or the environment. Here the stream of data is collected from various sensors and the information may be aggregated to derive some essential information pertaining to the activity performed by the actor. It usually involves training the model using techniques of data analytics and machine learning, to train the system to classify various activities. Thus, human activities are monitored and meaningful activities are recognized. Example of such human activities recognition includes physical movements like walking, running, cycling, etc [2]. In sensor-based approach, object-attached sensors or wearable sensors can be used. However one of the key challenges in wearable sensors includes the size and the battery life of the device used.

2) Vision based activity recognition: In vision based activity recognition, the sensors used are the ones that can sense the visuals such as sequence of images or the video. In general, it can be camera-based system that captures the video which can be further processed to identify the ongoing activities in a given environment. Vision based activity recognition involves human detection, human tracking, behavior tracking and human action recognition. Different approaches can be used in vision based activity recognition, such as use of a single camera or multiple set of stereo cameras or infrared cameras. Here the generated data is the video, which are sequence of frames or images [3]. Vision based approach derives the features of the limbs and other human pose, which can be used for activity recognition and, it generally involves

978-1-5386-7989-0/19/$31.00 ©2019 IEEE

digital image processing to extract meaningful information from the images or the video.

B. Human Activity Recognition Based on Complexity

The human activity recognition based on the complexity can be classified as below [4].

1) Gesture Recognition: Gesture recognition involves identification of the simple motions of the human body parts such as hands, arms, face, etc. Being one of the most useful recognition system widely used in Human Computer Interaction system, it can be used to understand and interpret the sign language [5]. It can be used in entertainment systems involving virtual or augmented reality.

2) Action Recognition: Here, the action of single actor or single person is recognized, which involves recognizing the actions such as walking, running, jumping, bending, etc.

3) Interaction Recognition: This involves recognizing the activities between two actors, wherein the activities may be two people talking to each other, shaking hands, hugging each other, etc. Set of action recognition between the two actors sometimes leads to interaction recognition.

4) Group Activity recognition: This involves recognizing the activities performed by -

a) group of actors (group action): [6] such as marching, choir singing, etc.

b) person interaction with group (group-person interaction): such as group of-people performing the activities based on the signal given by the other person, such as choir master giving signal to choir members.

c) Two groups interacting with each other (group-group interaction):

d) people within group interacting (inter-group interaction): such as two people within a group fighting or one member of the choir singing while other person playing instrument.

C. Types of Vision Based Activity Recognition- Based on Perspective

The types of activity recognition based on the perspective of view can be classified as follows:

1) First-person Perspective: Here the activity or interaction between the two entities that are interacting will be sensed, or recorded by one of the entity performing the action. For example, use of head mounted camera while performing the activity or while interacting with another person [7]. This also means that the person engaged in the activity is the observing entity of the action or interaction.

2) Third Person Perspective : Here the activity performed by a person or the interaction between the people or entities is captured from the perspective where the observing entity is different from that of the entity performing the activity. For example, one person watching interaction between two people will become third person perspective. Likewise, recognizing action or activity from a still camera mounted at some point, provides third person perspective.

Following are the types of interaction that can be observed from first person perspective as well as third-person perspective:

a) Human-Human Interactions: Involves two human interacting, such as shaking the hands or waving the hands at each other.

b) Human-Object-Human Interactions: Involves two humans interacting with a particular object or two humans interacting with each other using an object, example involves passing the object or throwing the object.

c) Human-Object Interactions: Here the single human entity interacts with a particular object. Example for this interaction can be, writing, typing, opening the door, eating, etc.

D. Types of Activity Recognition Based on Approaches

Based on approaches, recognition of activities can be classified as Single Layered or Hierarchical approach:

1) Single Layered Approach: This involves identifying the primitive activities directly from the acquired video, that does not involve identifying sub activity or set of simpler actions that constitute an activity [6]. Example of such single layered activity can be extending the hand.

2) Hierarchical Approach: This involves identifying set of primitive activities or action or set of sub activities that constitute an activity. It involves identifying set of primitive activities and aggregating the acquired information of these activities to define one activity. Example for activity that involves hierarchical approach is 'handshake', which requires two people to 'extend their hands', wherein the hands must meet each other and then recognizing withdrawing of the hands.

Figure 1 describes the overview of types of activity recognition based on the information gathered from various sources [1]–[8]. Here the types of activity recognition basically focuses more on vision based activity recognition.

II. LITERATURE REVIEW

Survey on various methods of activity recognition and action recognition were made. Where, some works for action recognition were based on trajectory of the motion and few of them were based on pose estimation. The works that are used for activity recognition follow the steps in the sequence: 1. Segmentation of the video, where the region of interest or presence of humans is detected, 2. Feature Extraction, where the required features are extracted based on the motion or the pose of the humans. 3. Feature Representation, here the extracted features are represented using the feature vectors or feature descriptors. In case of topic modeling, code-book is used to represent these features. Finally, training and testing is done using classification model. Figure 2 shows the generic flow of activity recognition based on the existing systems.

Segmentation in human based activity recognition acts like a preprocessing step and this may or may not be performed based on the steps used in feature extraction and feature representation. It is observed that some algorithms perform feature extraction without the use of segmentation. Segmentation is

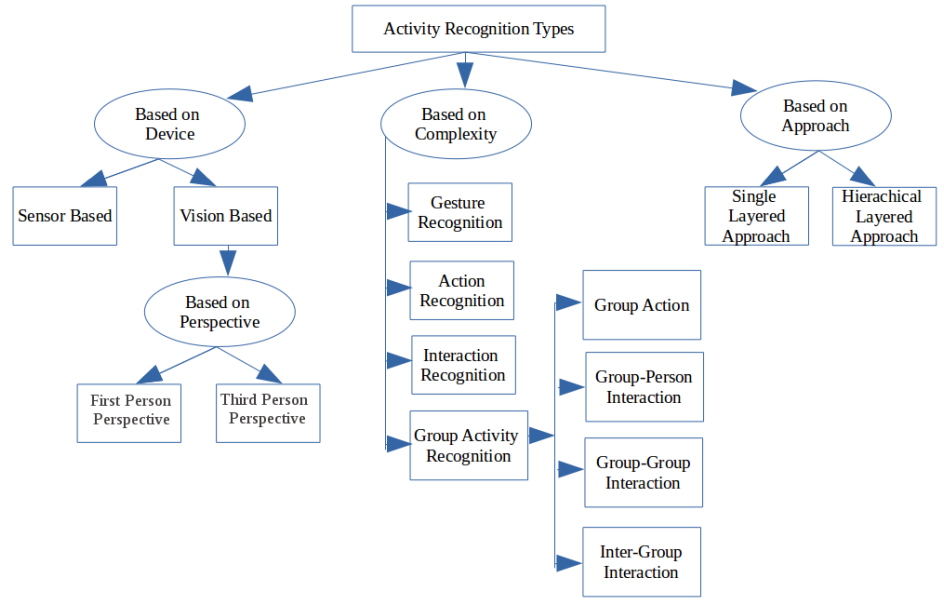

Fig. 1. Types of Activity Recognition

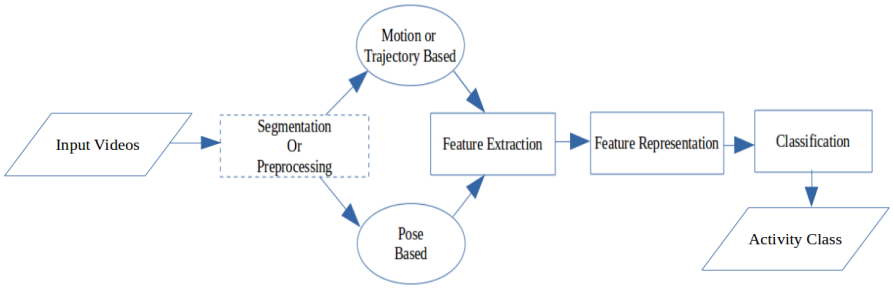

Fig. 2. Generic Flow of Activity Recognition

defined as dividing the entire image into group of subsets, typically one of the subset must contain the region of our study that has to be processed further. Pre-processing techniques such as background subtraction or foreground object extraction has been used for this purpose [9]. The other preprocessing techniques may involve marking of key start and end frame manually.

Feature Extraction, this step involves extraction of features such as shape, silhouette, motion information, etc, that can be represented so that the classification algorithm may be applied over it. Feature extraction varies based on the type of approach that is used for activity recognition. Activity recognition can be achieved using two approaches, 1. Motion or Trajectory Based Approach- where the features represent the motion information of the humans or objects. This type of approach is used in few of the works [10]–[12] and 2. Pose Based Approach- where the pose of the human is considered and acts as feature for the action or activity recognition [13]–[15].

The features used for motion based approach in few of the notable works are- Interest point (IP), Lucas-Kanade (LK) and Farnback (FB) trajectories [10], Optical flow [12], etc. The features used for pose based approach used in few of the notable works are- Human Joint-Coordinates along with distance and angle features representing joints, where the work [15] has performed 14-part and 26-part joint coordinates, silhouette extraction [16], fuzzy model for the selection of key pose [16], depth silhouettes [14], etc.

The extracted features are then represented, so that the classification algorithms may be applied over them. Here the feature representation depends on the approach used, since the feature representation depends on the extracted features. The features can be represented after applying dimension reduction algorithms like principal component analysis (PCA), local linear embedding (LLE) or Linear Discriminant Analysis (LDA) [14], [16].

The features extracted can be represented as a single descriptor or a topic model- where the set of words map to a particular topic. In the work [15], Pachinco allocation model, a topic model is used for the feature representation, where the features are angle and distance parameters corresponding to the human pose. Other topic models used are bag of words algorithm and Latent Dirichlet Allocation, which requires the generation of

code-book of words. Here the words are essentially derived per video frame, where the set of words map to particular poselet, which in turn may map to an action, and which finally maps to an activity. The other feature representation use Radon-Factor or R-descriptor obtained after applying radon-transform, here the Radon filters are invariant to scaling of the shapes of the human pose, which are effective when the size of appearance of the person changes [16]. Other ways of feature representation includes spatial distribution of edge gradient (SDEG) [16], Translation and Scale Invariant probabilistic Latent Semantic Analysis model (TSI-pLSA) [12]. Hu moments and Zernike moments feature vector are used in the case of work [11], where the optical flow was the feature extracted.

The classification algorithm is used to create the classification model based on the training data, where this created model is used to test the video for recognizing and classifying the activity. Few of the classification algorithms used for activity recognition are multi-class Support Vector Machine(SVM) classifier [10], [16], Expectation Maximum (EM) Algorithm, Bayesian decision [12], Hidden Markov Models (HMMs) [14], Feed-forward neural networks [11], etc.

It can also be observed that the concepts of Neural Networks and Deep Learning [17]–[20], are used in recent approaches and networks such as Convolutional Neural Networks, Recurrent Neural Networks and LSTMs are used. These types of neural networks have reduced amount of preprocessing, since CNNs are used to find the hidden patterns in the given data-set and also RNN takes time series data, which is very useful in gaining the temporal information.

The methodology used by various authors for the activity recognition along with the scope for future work as mentioned by the authors are discussed:

Thien Huynh-The, et al. [15], studied the complex person-person activities based on the knowledge coming from pose. The authors used articulate-body estimation to obtain the human joint coordinates with 2 patterns: 14-part and 26-part. The authors have calculated the distance and angle features between the 2 joints and their angle with respect to the horizontal. Eight feature categories were obtained: Intra spatio joint distance, Intra spatio joint angle, Inter spatio joint distance, Inter spatio joint angle, Intra temporal joint distance, Intra temporal joint angle, Inter temporal joint distance and Inter temporal joint angle. K-Means clustering was then used to quantize the features and construct codebook. The code-words are: d-word corresponding to Spatio-Temporal distance and a-word corresponding to Spatio-Temporal angle. Here, the single frame of the video was represented using collection of d-words and a-words from the codebook. 4-level PAM (Pachinko Allocation Model) based on Hierarchical model of topic modeling was proposed for topic-modeling. The levels are Root-topic, Super topics at the second level corresponding to interactive activities, followed by subtopics corresponding to interactive poselets, then the N unique codewords as the last level. Binary Tree of SVMs are used for N-class classification by using the topic model. The authors concluded that 26-part pattern produced better results than 14-part and obtained the accuracy of 91.2% on BIT data-set using temporal distance angle features and accuracy of 91.7% on UT interaction data-set on temporal distance angle features.

Slim Abdelhedi, et al. [11], proposed the method of human activity detection using optical flow by using Hu and Zernike Moments together for feature representation. The authors extracted the information of motion in the video using Optic Flow Vector modeling to obtain the motion descriptor. Here Lucas-Kanade algorithm is used to obtain Optic Flow and the result of this step is curvature of orientation. Using the Curvature of Orientation, the action features are obtained using i) Hu moments method (it is invariant to scale, rotation or translating); and ii) Zernike moments (rotation invariant and robust to noise), The combined features of both are used as for the feature representation (This forms the mid-level video representation method). Here, Feed Forward neural network (FFNN) was used for training and classification. The work obtained an accuracy of 97.3% and 63.7% on KTH and Weizmann data-sets respectively. Inference is that usage of mid level curvature increased the accuracy of detection. The authors proposed future work of using shape mid level representation instead of the applied curvature representation.

Ping Guo, et al. [12], solved the problem of identifying the key start and end frame of the action in the video clip which had multiple actions combined. The spatial-temporal interest points(STIPs) are first extracted. For each STIP, HOG (Histogram of Gradient) & HOF (Histogram of Flow) are extracted to form Feature Descriptors. K-means algorithm is used to group the feature vectors into clusters. Each cluster forms the visual word. Probabilistic Latent Semantic Analysis (pLSA) was originally used for document analysis, in Translation and Scale Invariant probabilistic Latent Semantic Analysis model(TSI- pLSA), the number of model parameters increase with increase in training data, hence generative TSI-pLSA is proposed (a new mathematical model proposed by the authors). EM is used for Model Fitting, for classification of activities. The authors have used Bayesian decision for deciding the boundary of the action, which is used to decide key start and end frames. A threshold is used to decide the end of round. If the action categories of two rounds are different, then the decision boundary of the action is made. Here, the proposed method works better when each action has independent words, hence the suggested scope for the future work is to work on those different actions that have similar words. An accuracy of 90.8% and 97.8% are achieved on KTH and Weizmann data-sets respectively.

A. Jalal, et al. [14], proposed a method to perform activity recognition on Depth-Silhouette. At first Depth-Silhouette is obtained from the depth camera, which produces depth maps and RGB images. The region of interest in obtained image is then resized. In order to compare the depth silhouette approach with that of binary silhouette, the binary silhouette is obtained from depth silhouette. Each image is converted to a single row vector, which is then mean normalized. Radon Transform is applied on the depth silhouette and R transform

is used to get 1-D R transform profile, which provides a scale invariant shape representation. Sequence of R transform was obtained for 10 consecutive frames of every video, which outputs the 3-D data. Principal Component Analysis (PCA) is used to reduce the dimensions of R-Transform profile. Linear Discriminant Analysis is used to map the R transformed profiles to different activity classes. LDA further reduced the dimension. Clustering algorithm based on Vector quantization is used to generate codebook of vectors. Then HMM algorithm is used for classification. The authors used their own data-set on which an accuracy of 91.6% for depth Silhouette and 67.08% for binary silhouette was obtained.

Jenhui Hu, et al. [21], proposed a method for activity recognition, which involved constructing set of templates for each activity. The templates are designed to capture the structural and motion information. Method is used to solve the temporal variability of the activities in the same class. Binary Silhouettes are used as the input and are scaled and centered in an image. The centering is done in two ways: i) frame-to-frame basis and ii) horizontal displacement canceling using the same displacement across all the frames. Temporal segmentation is done on the sequence of images using clustering algorithm to have four temporal segments. The clustering is done based on the Euclidians distance between the frames. Motion Energy Images on centered sequence of silhouettes are constructed. The four stages of templates are obtained, where each template shows the information that changes. The movement of the foreground objects information is used to obtain motion profile. Activity recognition is performed based on the distance calculation with the above templates and motion profile. The weight map is designed to discriminate different activities having same templates. Hence, weighted template distance is used to obtain the activity class. An accuracy of 85.83% on IXMAS data-set was achieved.

Weiyao Lin, et al. [22], proposed a method of using the features called heat-map to capture the temporal detail for Group Activity Recognition. The entire video scene is divided into small squares. If the motion trajectory goes through this square, it will be defined as one heat source. Hence, series of heat maps is obtained. The decay information of heat are used to obtain the temporal information (The first frame will be described using low heat, while the last frame will be described using more heat). A key-point-based method is used to handle the alignment in the Heat maps. Alignment works as feature scaling, since same activities show different scales and rotations. Surface fitting method is used for recognition of activities based on HM feature. A standard surface is first obtained for every class, then the surface of test video is compared against the standard surfaces of every class and the best match is selected as the predicted class. Total Error Rate (TER) was used as metric for evaluation, TER of BEHAVE Data-set and Traffic Data-set are 8.8% and 4.5% respectively, and UNM data-set- ROC curve was used to prove that HMB+optic flow algorithm proposed works better than normal optic flow methods, Social force model and velocity-field models.

Dinesh K. Vishwakarma, et al. [16], suggested use of i) fuzzy model for the selection of key pose. Histogram distance is used to identify the single optimal key pose. Spatial distribution of edge gradient (SDEG) of human pose was found and ii) human silhouette were extracted on which Radon transform is used to obtain R factor. The work reported that classifier works better when SDEG and R-transform is used in combination. Here the effect of Radon filter or R-transform is that this filter is scale invariant to the human pose. This added an advantage for activities to be recognized irrespective of the position of the actors in the video frame, and the size of appearance of the actor in the video frame did not impact the performance of the classifier. The inference of this work is- i) As the number of levels used for computing the SDEG features increased, the dimensions of SDEG feature vector also increased but the accuracy of recognition did not increase significantly, ii) As the number of key poses considered for computation of temporal information increased, the accuracy of recognition improved slightly, but it results in high dimension of feature vector, and iii) It is also observed that R-transform was more effective for the activities which have significant orientation like bending, crouching, etc. The work reported the future work to increase the number of key poses for estimation of orientation. Accuracy of 100%, 95.5%, 93.25%, 92.92% and 85.5% respectively was achieved on Weizmann, KTH, Ballet, i3dpost and IXMAS data-sets.

S. U. Park, et al. [18], proposed used of Recurrent Neural Network (RNN) for HAR. The joint angles are computed, and input feature matrix is created for the obtained joint angles, Recurrent Neural Networks are used for training the data. RNN consisted of 50 Long Short-Term Memory (LSTM)s with 90 hidden units, which prevented the vanishing gradient problem. Accuracy of 95.55% was achieved using MSRC-12 data-set.

Haiam A., et al. [23], worked on improving the trajectory-based human action recognition approaches. STIP (Space-Time Interest Points) in each video is extracted using the technique of matching SIFT(Scale invariant feature transform) descriptors. A windowing approach is used to reduce the incorrect matches. Motion trajectories are generated by matching consecutive descriptor of the frames. Trajectories which are formed in less than 10 frames are tried to be matched (this has happened due to self-occlusion and noise), which are then tried to be linked, by matching SIFT descriptors. The short trajectories that still exist are pruned by considering it as noise. Four trajectory shape descriptors were used: Trajectory shape descriptor, HOG (Histogram of Oriented Gradients), HOF (Histograms of optic flow) and MBH (Motion Boundary histogram). Then the visual words are built using k-means clustering by using training data for bag of words algorithm. SVM is used for classification. This algorithm achieved an accuracy of 95.36% on KTH data-set with HOF descriptor, 97.77% on Weizmann data-set with HOG feature, and 89.97% on UCF-Sports with combined features.

Kuan-Ting Lai, et al. [24], proposed a new algorithm based on SIFT key points for action recognition. The key-points are detected based on SIFT to detect the local features. SIFT

keypoints are matched in the consecutive frames. The large erroneous keys that are matched were removed as it affected the later stages. The maximum matching key-length of 10 for running actions, and a key-length of 6 for all the other actions in the KTH data-set was considered. SIFT displacement vectors were further quantized into bins, weighted by vector lengths, hence obtaining the histogram. SVM was used to learn and recognize the action features. An accuracy of 78.67% on KTH data-set was reported.

M.M. Youssef, et al., [13], proposed a new feature extraction technique using hull convexity defects for recognition of human activity. Background subtraction is first done to obtain the silhouette. Then, the contour detection and removal is done using the threshold, wherein the largest contour (represented using polygons in this work) is assumed to be that of human. The silhouettes are then normalized. The feature is extracted using convex hull defects, i.e, use of negative space created by the convex hulls. The feature vector is derived using the silhouette hull defects, which are the space (triangle) between the contour lines and the actual object. A threshold is used to remove the triangle whose area is less than threshold value, which removed the defects caused by small triangles. K-means clustering algorithm is used to cluster the training set into its own subspace and minimum distance metric is used for the classification of the test data. Accuracy of 97.5% is achieved using this technique on Weizmann data-set.

Chiranjoy Chattopadhyay, et al. [25], proposed a method to automatically classify the interaction videos and, retrieve the videos having similar interactions from the database. The local structures in space-time with significant variations in the pixel intensities are considered from the input video. The detected points are ensured to concentrate on the part of video where there is significant amount of motion. The points are represented using descriptors such as HOG, HOF and LBP-TOP(Local binary patterns on three orthogonal planes). Vectors of Locally Aggregated Descriptors(VLAD) are constructed by generating codebook by clustering the obtained feature descriptors. The authors have followed the method of Bag of Visual Words algorithm. SVM algorithm (1-against-all SVMs) is used for classification. In order to retrieve the similar videos from the database, VLAD of all the videos are stored in the feature database. Given a query video, VLAD of the given video is first obtained, and then matched against feature descriptor stored in the database, similarity is measured, and the videos are retrieved based on the ranking of the similarity of the video descriptors in the database. Combining all the 3 descriptors together gave better results than use of single or combination of 2 descriptors. Method achieved an overall accuracy of 88.52% on UT-Interaction data-set and 87.12% on BIT interaction data-set respectively.

Tushar Dobhal, et al. [19], proposed a method to classify the human actions by converting the 3D videos to 2D binary motion images. For the input video, background from each image is subtracted using Gaussian Mixture model. All the action sequence images are then combined to obtain a single image known as Binary Motion Image (BMI). Then, Convolutional Neural Networks(CNN) is used for learning, which does both extraction of features as well as classification. CNN requires less pre-processing compared to ANN. Accuracy of 100% on Weizmann data-set and 98.5% on MSR Action 3D data-set was achieved. The authors used MATLAB for extracting the BMI, and ConventJS to implement a 3 layer CNN.

Sheng Yu, et al. [20], proposed use of a two stream CNN in order to avoid the problem of overfitting in CNN and perform the action recognition. The input data is passed into two separate streams (Spatial and Temporal). The RGB video frames became the input to the spatial stream. Stacking of optical flow obtained using TLV1 method is used as input to the temporal stream. Learning rate of 0.00001 is used for 1st 10k iterations followed by 0.000001. Stochastic gradient descent is used for training the model. The streams of CNNs are treated as feature extractors and the last max-pooling layer is used as vector of features. Two fusion techniques are used to fuse the features: i) Linear weight fusion method is used to add the pixels of spatial and temporal feature maps where its weights signifies the importance; ii) Concatenation fusion, reshapes the combination of both the features into single vector. A Vector of locally aggregated descriptor (VLAD) and temporal pyramid pooling (TPP) are used together to obtain video level features. The classification is done based on SVM. Caffee toolbox is used to implement the CNN. Accuracy of 90.5% on UCF101 data-set using linear weighted fusion technique and Accuracy of 63.4% on HMDB51 data-set using linear weighted fusion technique has been reported.

Lei Wang, et al. [26], proposed a method to recognize the action using only RGB data. The video frames are used as input to convolutional neural network to extract the features, the output of fully connected(FC) layer and convolutional(Conv.) layer of CNNs are different. The features obtained from FC layer are further processed by FC-LSTM, and Conv. features are processed by Conv.-LSTM. LSTMs are used to obtain the temporal information. The output of LSTMs are passed through attention model based on Rectified linear units activation (relu). The attention model here focuses on the frames which are important for action classification as well as the important region of the frames. A joint optimization module is built to combine the feature vectors obtained from the two LSTMs. The joint optimization module contains a FC layer followed by dropout operation, to avoid over-fitting problem. The softmax layer is used for the final prediction. Accuracy of 98.76%, 91.89% and 84.10% on UCF-11, UCF-Sports and UCF-101 data-sets are achieved respectively.

Thus, we can infer that there is no single straight forward method that can be employed for activity recognition. However, we have a choice of variety algorithms at every step that can be used for recognizing the activity, where few of the important steps include, feature extraction followed by feature representation, and then the classification over the represented feature, used to classify the activities.

III. APPLICATIONS OF ACTIVITY RECOGNITION

Some of the applications of human activity recognition are enumerated and explained below:

- *Behavioral Bio-metrics:* Bio-metrics involves uniquely identifying a person through various features related to the person itself. Traditional bio-metrics features are physiological bio-metrics based on physical attributes of person, such as finger-print, eye-print, etc [9]. Such traditional algorithms require intervention and cooperation of the person intended to be identified. Behavior bio-metrics is based on the long term observation of humans, which do not require any intervention or which requires the least amount of intervention. It involves using methods such as motion history methods for authentication [27].
- *Content based video analysis:* Applications of HAR in content based video analysis encompass platforms such as video sharing and other application, where such systems can be made more effective, if the activities in the video are recognized. HAR in such systems can improve user experience, storage, indexing, summarization of contents.
- *Security and surveillance:* This traditionally involved the human authority or a person monitoring network of cameras. With advances in the camera and imaging technology, the efficiency and accuracy of human operators has increased. Hence, it is further easier for the security persons to monitor the security environment if the vision based activity recognition is incorporated in the system. This would also ease the task of manually analyzing multiple videos to identify unusual activities. A related application to this involves searching the huge database typically containing long videos in order to identify the set of similar activities, wherein such application can extend to content based video retrieval.
- *Interactive applications and environments:* This involves understanding the activities of the humans to respond to the human activity, which is one of the main goal of Human Computer Interaction systems. This is one of the main mode of nonverbal communication. Effective implementation of such systems that uses actions or gestures as input can aid in development of better robots or computers that will be able to respond and interact with humans efficiently.
- *Animation and synthesis:* The gaming and animation industries rely on efficient synthesis of natural human motions, which involves production of large amount of human motions based on the natural human motions. A related application is learning in simulated environments, where the activities are first tracked, and then simulated in the simulation environment. Such simulations can be used in training firefighter men or army personal where natural human motions can be simulated.
- *Healthcare Systems:* In the healthcare systems, recognition of the activities can help in better analyzing and understanding the patients activities, which can be used by the health workers to diagnose, treat and care for patients. Recognizing the activities could improve the reliability on the diagnosis, also decreases the work load for the medical staff, reduces the duration of stay of the patients at hospitals, and improves patients quality of life [28].
- *Rehabilitation Applications:* Traditional rehabilitation systems work in a manner where the patients are required to undergo several visits to the clinic for the treatment, exercises and the periodical evaluation until his/her complete recovery of normal daily behavior occurs. These visits to the clinic can be avoided by using creative rehabilitation systems, and with the application of video-based activity recognition system, self-health care and home-centered rehabilitation systems can be created. This would also aid in continuously monitoring the daily activities and movements, which would further help in detection of early symptoms of few diseases, so that the diagnosis can be made at the early stages.

IV. CHALLENGES IN VISION BASED HAR

- *Human behavior:* Some humans can perform multiple tasks at the same time, which makes the process of recognition more difficult.
- *Intra-class variability:* A given activity, can be performed differently by different individuals.
- *Inter-class similarity:* Classes that are fundamentally different, may show similar characteristics. Example of such activity may be, skipping and running has the similar pose, which becomes challenging when pose based algorithms are used for activity recognition [13].
- *Illumination changes:* There may be dynamic variations in factors such as brightness, contrast and other parameters in a video, or these parameters may be affected due to the factors such as change in environmental condition like change in weather, etc [29], [30].
- *Shadow Effect:* The shadow of the object or human may create silhouettes, which may lead to false detection and tracking of the activity.
- *Partial or full occlusions:* Subject of interest whose action is to be recognized may be occluded by another object in the video. This makes it difficult to identify the activity.
- *Self-occlusions:* One body part of the person may block the view of other part of the body (which may be significant in recognizing the activity), which typically occurs due to the viewpoint.
- *Scaling:* The people performing the activity, may be near or far from the device used to sense the activity.
- *Bootstrapping:* The background in training environment may not be the same as the one which is available where the cameras has to be actually deployed [30].
- *Camera jitter:* The quality of video may be degraded due to low resolution, or poor quality of the recording device.
- *Camera automatic adjustments:* The modern cameras have the feature of auto-focus, white balancing and brightness adjustments, which may lead to problems since the same image may appear different in different frames.

- *Noisy frames in video:* The pixels related to people involved in the activity may be similar to the pixels in the background, leading to difficulty in segmenting out the humans or object of interest from the background.
- *Camouflage:* The pixels related to people involved in the activity may be similar to the pixels in the background, leading to difficulty in segmenting out the humans or object of interest from the background.
- *Moving background objects or humans:* This may lead to the wrong activity being identified, as they may be considered as the actual objects or humans involved in activity recognition.

V. Conclusion

Through this work, various types of activity recognition systems and techniques are reported. It is observed that interaction and action recognition are solved in two different ways. Further there is no single solution that fits all class of problems. There are many algorithms available to solve activity recognition. However, the same sequence of algorithms perform with different efficiency on different data-sets. The activity recognition has widespread applications in different domains. However, vision based activity recognition has its own set of challenges.

References

[1] J. Aggarwal and M. Ryoo, "Human Activity Analysis: A Review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1922649.1922653

[2] J. T. Sunny and S. M. George, "Applications and Challenges of Human Activity Recognition using Sensors in a Smart Environment," vol. 2, no. 04, p. 8.

[3] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-Based Activity Recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, Nov. 2012.

[4] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in Human Action Recognition: A Survey," *arXiv:1501.05964 [cs]*, Jan. 2015, arXiv: 1501.05964.

[5] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, May 2007.

[6] M. S. Ryoo and J. K. Aggarwal, "Recognition of High-level Group Activities Based on Activities of Individual Members," in *Proceedings of the 2008 IEEE Workshop on Motion and Video Computing*, ser. WMVC '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 1–8. [Online]. Available: https://doi.org/10.1109/WMVC.2008.4544065

[7] S. Narayan, M. S. Kankanhalli, and K. R. Ramakrishnan, "Action and Interaction Recognition in First-Person Videos," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2014, pp. 526–532.

[8] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From Handcrafted to Learned Representations for Human Action Recognition," *Image Vision Comput.*, vol. 55, no. P2, pp. 42–52, Nov. 2016. [Online]. Available: https://doi.org/10.1016/j.imavis.2016.06.007

[9] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[10] B. Boufama, P. Habashi, and I. S. Ahmad, "Trajectory-based human activity recognition from videos," in *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, May 2017, pp. 1–5.

[11] S. Abdelhedi, A. Wali, and A. M. Alimi, "Human activity recognition based on mid-level representations in video surveillance applications," in *2016 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2016, pp. 3984–3989.

[12] P. Guo, Z. Miao, Y. Shen, W. Xu, and D. Zhang, "Continuous human action recognition in real time," *Multimedia Tools and Applications*, vol. 68, no. 3, pp. 827–844, Feb. 2014. [Online]. Available: https://link.springer.com/article/10.1007/s11042-012-1084-2

[13] M. Youssef and V. Asari, "Human action recognition using hull convexity defect features with multi-modality setups," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1971–1979, 2013.

[14] A. Jalal, M. Z. Uddin, and T. S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 863–871, Aug. 2012.

[15] T. Huynh-The, B.-V. Le, S. Lee, and Y. Yoon, "Interactive Activity Recognition Using Pose-based Spatio-temporal Relation Features and Four-level Pachinko Allocation Model," *Inf. Sci.*, vol. 369, no. C, pp. 317–333, Nov. 2016. [Online]. Available: https://doi.org/10.1016/j.ins.2016.06.016

[16] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "A Proposed Unified Framework for the Recognition of Human Activity by Exploiting the Characteristics of Action Dynamics," *Robot. Auton. Syst.*, vol. 77, no. C, pp. 25–38, Mar. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2015.11.013

[17] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential Deep Learning for Human Action Recognition," in *Human Behavior Understanding*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Nov. 2011, pp. 29–39. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-25446-8_4

[18] S. Park, J. Park, M. Al-masni, M. Al-antari, M. Z. Uddin, and T.-S. Kim, "A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services," *Procedia Computer Science*, vol. 100, pp. 78–84, 2016.

[19] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human activity recognition using binary motion image and deep learning," *Procedia computer science*, vol. 58, pp. 178–185, 2015.

[20] S. Yu, Y. Cheng, L. Xie, and S.-Z. Li, "Fully convolutional networks for action recognition," *IET Computer Vision*, vol. 11, no. 8, pp. 744–749, 2017.

[21] J. Hu and N. V. Boulgouris, "Fast Human Activity Recognition Based on Structure and Motion," *Pattern Recogn. Lett.*, vol. 32, no. 14, pp. 1814–1821, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2011.07.013

[22] W. Lin, H. Chu, J. Wu, B. Sheng, and Z. Chen, "A Heat-Map-Based Algorithm for Recognizing Group Activities in Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1980–1992, Nov. 2013.

[23] H. A. Abdul-Azim and E. E. Hemayed, "Human action recognition using trajectory-based representation," *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 187–198, 2015.

[24] K.-T. Lai, C.-H. Hsieh, M.-F. Lai, and M.-S. Chen, "Human action recognition using key points displacement," in *International Conference on Image and Signal Processing*. Springer, 2010, pp. 439–447.

[25] C. Chattopadhyay and S. Das, "Supervised framework for automatic recognition and retrieval of interaction: a framework for classification and retrieving videos with similar human interactions," *IET Computer Vision*, vol. 10, no. 3, pp. 220–227, 2016.

[26] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks," *IEEE Access*, vol. 6, pp. 17 913–17 922, 2018.

[27] A. Drosou, D. Ioannidis, K. Moustakas, and D. Tzovaras, "Spatiotemporal analysis of human activities for biometric authentication," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 411–421, 2012.

[28] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A Review on Video-Based Human Activity Recognition," *Computers*, vol. 2, no. 2, pp. 88–131, Jun. 2013. [Online]. Available: http://www.mdpi.com/2073-431X/2/2/88

[29] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: a review," *Artificial Intelligence Review*, pp. 1–57, 2017.

[30] A. Bux, P. Angelov, and Z. Habib, "Vision based human activity recognition: a review," in *Advances in Computational Intelligence Systems*. Springer, 2017, pp. 341–371.