Human Activity Recognition using Computer Vision based Approach – Various Techniques

Prarthana T V[1], Dr. B G Prasad[2]

[1]*Assistant Professor, Dept of Computer Science and Engineering, B. N. M Institute of Technology, Bengaluru*
[2]*Professor, Dept of Computer Science and Engineering, B. M. S. College of Engineering, Bengaluru*

Abstract - *Computer vision is an important area in the field of computer science which aids the machines in becoming smart and intelligent by perceiving and analyzing digital images and videos. A significant application of this is Activity recognition – which automatically classifies the actions being performed by an agent. Human activity recognition aims to apprehend the actions of an individual using a series of observations considering various challenging environmental factors. This work is a study of various techniques involved, challenges identified and applications of this research area.*

Key Words: Human Activity Recognition, HAR, Deep learning, Computer vision

1. INTRODUCTION

Activity Recognition, a sub domain of vision related applications, is the ability to identify and recognize the actions or goals of the agent, the agent can be any object or entity that performs action, which has end goals. The agent 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 is called as Human Activity Recognition (HAR). The goal of human activity recognition is to automatically analyze ongoing activities from an unknown video (i.e. a sequence of image frames). In a simple case, where a video is segmented to contain only one execution of a human activity, the objective of the system is to correctly classify the video into its activity category[1]

HAR is one of the active research areas in computer vision as well as human computer interaction. The wide variety and unexpected pattern of activities performed by humans, poses the challenge to the methodology adopted for recognition. There are various types of human activities. They are conceptually categorized as – gestures, actions, interactions, and group activities. Gestures are elementary movements of a person's body parts, and are the atomic components describing the meaningful motion of a person. Actions are single-person activities that may be composed of multiple gestures organized temporally. Interactions are human activities that involve two or more persons and/or objects. Finally, group activities are the activities performed by conceptual groups composed of multiple persons and/or objects. Approaches implemented should cater to all the requirements. The above listed category of activities is known as Usual Activities. Any activity which is different from the defined set of activities is called as Unusual Activity. These unusual activities occur because of mental and physical discomfort. Unusual activity or anomaly detection is the process of identifying and detecting the activities which are different from actual or well-defined set of activities and attract human attention.

Applications of human activity recognition are many. A few are enumerated and explained below:

- Behavioral Bio-metrics - Bio-metrics involves uniquely identifying a person through various features related to the person itself. 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.
- 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, etc
- Security and surveillance - Systems which implement HAR and unusual activity detection are very helpful to monitor the environment without the interference human operators. The efficiency and accuracy of such system increase drastically.
- Interactive applications and environments - This involves understanding the activities of the humans to respond to the human activity, which is one of the main goals of Human Computer Interaction Systems. This is one of the main modes 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.
- Healthcare Systems - In the healthcare systems, recognition of the activities can help in better analyzing and understanding the patient's 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.
- Elderly care – Systems involving HAR would be helpful in improving the quality of life of elders. Such systems can be

used to monitor the activities of elders, process the videos captured and in case of disastrous events, alerts can be sent to the concerned.

2. LITERATURE SURVEY

Classification of Activity recognition can be based on multiple parameters. This taxonomy has been well explained in [2] and stated diagrammatically below.

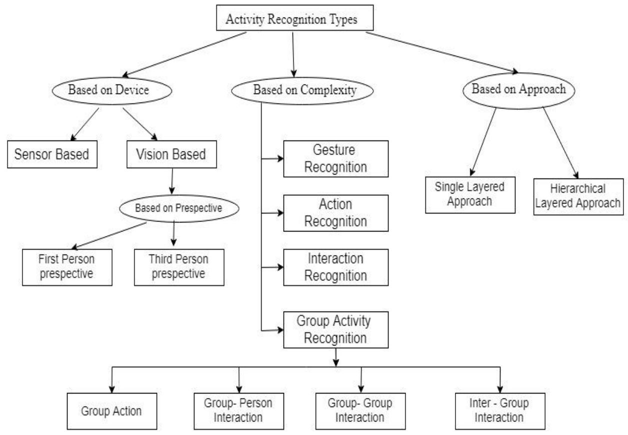

Fig -1 Classification of Activity recognition

A. Types of Activity Recognition based on devices used:

Based on the devices used in the system, Activity Recognition is classified as sensor-based activity recognition and vision-based activity recognition.

1. Sensor based activity recognition uses network of sensors to monitor the behavior of an actor, and some monitor the surroundings. Such data collected from various sensors may be aggregated and processed to derive some essential information from them. They are further used for training the model using different data analytics, machine learning and deep learning techniques.

2. Vision based activity recognition is a camera- based system that captures the video that can be processed and used to identify the activities in the given environment. These systems normally use digital image processing to extract meaningful information from the video, which is considered as sequence of images.

B. Types of Activity Recognition based on Complexity [3]

1. Gesture Recognition - Gestures are elementary movements of a person's body parts, and are the atomic components describing the meaningful motion of a person. "Stretching an arm" and "raising a leg" are good examples of gestures.

2. Action Recognition – Actions are single-person activities that may be composed of multiple gestures organized temporally, such as "walking," "waving," and "punching."

3. Interaction Recognition - Interactions are human activities that involve two or more persons and/or objects. For example, "two persons fighting" is an interaction between two humans and "a person handing over a suitcase to another" is a human-object interaction involving two humans and one object.

4. Group Activity Recognition – This involves activities performed by a group of actors, a person interacting with the group, two groups interacting with each other and people within a group interacting with each other.

C. Types of Activity Recognition based on perspective

1. First person perspective – This means one of the persons engaged in performing the activity will also be engaged in capturing the video. An instance that can be thought of is of a situation where two people are interacting with each other, one of which has worn a head mounted camera which records the video for further processing.

2. Third person perspective – Here, the observing entity and the performing entity are different. The camera mounted at a static point, which is recording the interaction between two people forms a third person perspective.

D. Types of Activity Recognition based on Approaches [1]

1. Single Layered Approach involves identifying the primary activities or independent activities that are directly obtained from the video. A simple example would be a person extending hand.

2. Hierarchical Approach involves identifying set of sequential but atomic activities which when aggregated forms a different activity. For instance, the simple handshaking activity involves multiple atomic activities like two people extending hands, where in the hands must meet followed by withdrawal of the hands.

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

Intelligent visual identification of abnormal human activity (AbHAR) has raised the standards of surveillance. [4] summarizes the existing approaches to recognize abnormal human activity depending on the dimension of the information available. Based on the features and available information AbHAR is further classified as the below Taxonomy. Each of these approaches are in detail presented in the paper.

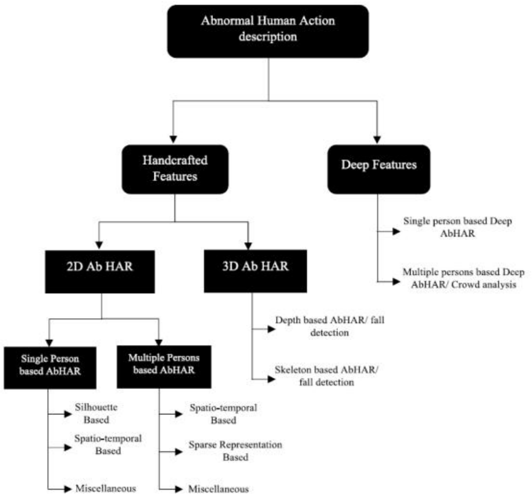
Abnormal Human Action description
Handcrafted Features
Deep Features
2D Ab HAR
3D Ab HAR
Single person based Deep AbHAR
Multiple persons based Deep AbHAR/ Crowd analysis
Depth based AbHAR/ fall detection
Skeleton based AbHAR/ fall detection
Single Person based AbHAR
Multiple Persons based AbHAR
Silhouette Based
Spatio-temporal Based
Miscellaneous
Spatio-temporal Based
Sparse Representation Based
Miscellaneous

Fig -2 Classification of AbHAR

Authors have also mentioned about the available datasets, accuracy of results across these datasets for different approaches and the various challenges presented in these data sets. Below mentioned are a few future work mentioned in the paper –

- Meaningful datasets must be developed to represent abnormal actions in different scenarios-office, home, the coffee shop.
- Availability of deep architectures based efficient and deeper abnormal action recognition system with required computational resources to the common man.
- Challenge lies in transferring the knowledge of three-dimensional data as depth or skeleton based feature descriptor to a realtime AbHAR systems with improved true detection rate and less computational complexity.

[5] compiles and categorizes recent existing methods involved in various stages of HAR - image or data input, feature extraction, descriptor formation, and classification. Summary of different existing data sets based of input types like RGB, RGB+depth+skeleton is tabulated well. This review categorizes HAR techniques in to two hierarchical layers. The top layer is categorizes based on the ordinary process flow components which are; input type, feature extraction, descriptor formation, and classification. In the second layer, it further categorizes each component into independent sub-categories based on the current proposed methods. Such categorization is presented below –

- Input type - colour-based, colour-based + depth, and skeleton based.
- Feature extraction - holistic and local features,
- Descriptor formation - global-based, local-based, and hybrid.
- Classification - supervised, unsupervised, semi supervised and deep learning classifiers.

An introduction to all these approaches and the scenario in which it is used is discussed. A detailed classification of features used, comparison in performances, results achieved is well presented. This paper also mentions the different datasets available and the diversity in their input types.

Thien Huynh-The, et al. [6], 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 codewords 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 pose lets, 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 dataset on temporal distance angle features.

Slim Abdelhedi, et al. [7], 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. [8] , 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. [9], proposed a method to perform activity recognition on Depth-Silhouette. Initially, 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. 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.

Jinhui Hu, et al. [10], 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 cancelling 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 Euclidian's 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.

3. METHODOLOGY

Survey of various papers suggested the below steps in sequence as methodology commonly used across for solving the problem of vision-based HAR.

1. Capture the video – Depending on the complexity of activity recognition and the resource availability, the camera with appropriate resolution is used to capture the video for further processing.

2. Segmentation of the video, where the region of interest or presence of humans is detected.

3. Feature Extraction, where the required features are extracted based on the motion or the pose of the humans.

4. Feature Representation, here the extracted features are represented using the feature vectors or feature descriptors.

5. Finally, training and testing is done using classification model.

The model is depicted in the below figure [2] –

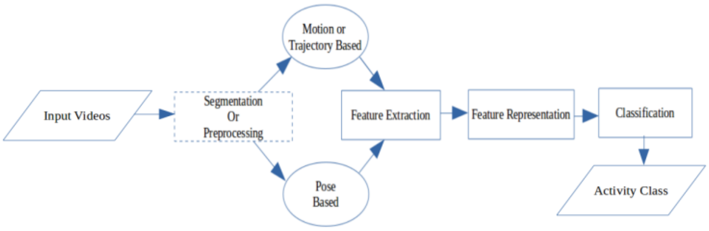

Fig -3 Generic flow of activity recognition

It depicts generic flow of activity recognition based on the existing systems. Each of the steps in the sequence are explained briefly.

Segmentation in human based activity recognition acts like a pre-processing 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 defined as dividing the entire image into group of subsets, typically one of the subsets 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 [11]. The other pre-processing techniques may involve marking of key start and end frame manually.

Feature Extraction is the step that 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 approach has been used in [7],[8] and [12] of the study. 2. Pose based approach – where the position of the human is the input for recognizing the activity. This approach is used in [6], [9] in the study.

The extracted featured have to the represented so that the classification algorithms can be applied. The features extracted can be represented as a single descriptor or a topic model- where the set of words map to a topic, spatial distribution of edge gradient (SDEG), Translation and Scale Invariant probabilistic Latent Semantic Analysis model (TSI-pLSA), Hu moments and Zernike moments feature vector. The representation depends on extracted features.

Classification algorithms are used to create the classification model based on the training data. The model created is then used to test the video for activity recognition and classification. Few of the classification algorithms used for activity recognition are Naïve Bayes, K-Nearest Neighbours, Expectation Maximum (EM) Algorithm, multi-class Support Vector Machine (SVM) classifier, Bayesian decision Hidden Markov Models (HMMs), Feed-forward neural networks etc.

Neural Networks and Deep Learning are also used in recent approaches. Networks such as Convolutional Neural Networks(CNN) which is used to find the hidden patterns in the given data-set, Recurrent Neural Networks(RNN) which uses time series data to retrieve the temporal information and LSTMs are used. The amount of pre-processing required is considerably lessened when these neural networks are used.

4. CHALLENGES IDENTIFIED

Human activity recognition and Anomaly detection has great implications in various fields with some useful applications like Security healthcare, Visual Surveillance and many others. Although it has such wide spread presence, there are a few challenges that are obstructing the development of the technology to its complete potential. These challenges encourage and bring in a lot of attention for good research.

Challenges in Human Activity Recognition –

1. Recognizing Concurrent Activities – Concurrent activities are activities performed by one or multiple individuals at the same time. For Example, one user might drink coffee while interacting with somebody. Processing videos to recognize both the activities is a challenge.

2. Recognizing Interleaved Activities - In real life, activities are conducted by people in an unpredictable manner. People might choose to switch between the actions of two or more activities. Such activities are termed to be interleaved activities. An instance for this would be the actor might answer a phone call, watch T V while cooking food.

3. Ambiguity in interpretation – Since multiple actors have their own style of performing an action, the technique for recognizing the activity should be versatile to take care of all these. Different people have different walking styles, a system should be able to recognize as all the styles as "Walking" activity without ambiguity.

4. Multiple resident complexity – When multiple actors are present in the same scene, it becomes complex to recognize all the activities performed by each of them. There might be scenarios where some actors are not clearly or partially visible.

5. Amidst of ample applications to be thought of, scope for accurate and simple systems.

Challenges in Anomaly detection –

1. Due to different motion patterns of different subjects at different time, the accuracy of activity recognition decreases.

2. It is difficult for any classification algorithm to recognize the motion during the transition period between two activities.

3. Resource constraints – Complex activities may need better resolution images for recognizing the activity without any ambiguity, which in turn needs better resolution cameras.

4. Datasets – Cleaning the dataset to match to the specific application of the research area.

5. CONCLUSION

This work is a survey of various techniques involved in activity recognition in general and human activity recognition in specific. The different methodologies used are also introduced. Applications reported in brief and challenges identified gives an insight into the area of computer vision based human activity recognition.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," ACM Comput. Surv., vol. 43, no. 3, 2011, doi: 10.1145/1922649.1922653.

[2] A. G. D'Sa and B. G. Prasad, "A survey on vision based activity recognition, its applications and challenges," 2019 2nd Int. Conf. Adv. Comput. Commun. Paradig. ICACCP 2019, pp. 1–8, 2019, doi: 10.1109/ICACCP.2019.8882896.

[3] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in Human Action Recognition: A Survey," no. February, 2015, [Online]. Available: http://arxiv.org/abs/1501.05964.

[4] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," Eng. Appl. Artif. Intell., vol. 77, no. August 2018, pp. 21–45, 2019, doi: 10.1016/j.engappai.2018.08.014.

[5] S. A. R. Abu-Bakar, "Advances in human action recognition: An updated survey," IET Image Process., vol. 13, no. 13, pp. 2381–2394, 2019, doi: 10.1049/iet-ipr.2019.0350.

[6] 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. (Ny)., vol. 369, pp. 317–333, 2016, doi: 10.1016/j.ins.2016.06.016.

[7] S. Abdelhedi, A. Wali, and A. M. Alimi, "Fuzzy logic based human activity recognition in video surveillance applications," Adv. Intell. Syst. Comput., vol. 427, pp. 227–235, 2016, doi: 10.1007/978-3-319-29504-6_23.

[8] P. Guo, Z. Miao, Y. Shen, W. Xu, and D. Zhang, "Continuous human action recognition in real time," Multimed. Tools Appl., vol. 68, no. 3, pp. 827–844, 2014, doi: 10.1007/s11042-012-1084-2.

[9] A. Jalal, M. 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 Trans. Consum. Electron., vol. 58, no. 3, pp. 863–871, 2012, doi: 10.1109/TCE.2012.6311329.

[10] J. Hu and N. V. Boulgouris, "Fast human activity recognition based on structure and motion," Pattern Recognit. Lett., vol. 32, no. 14, pp. 1814–1821, 2011, doi: 10.1016/j.patrec.2011.07.013.

[11] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," Vis. Comput., vol. 29, no. 10, pp. 983–1009, 2013, doi: 10.1007/s00371-012-0752-6.

[12] B. Boufama, P. Habashi, and I. S. Ahmad, "Trajectory-based human activity recognition from videos," Proc. - 3rd Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2017, pp. 1–5, 2017, doi: 10.1109/ATSIP.2017.8075536.