iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** V **Month of publication:** May 2019

DOI: https://doi.org/10.22214/ijraset.2019.5308

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

A Survey on Human Action Recognition

Neema V. Volvoikar[1], Prof. Vineet Pukhraj Jain[2]
[1, 2]*Dept. of Computer Science Engineering, Goa college of Engineering, Goa, India*

***Abstract:** In Computer Vision and image processing field, understanding the actions of human from videos is demanding task. Humans performing the actions in the video is automatically cognized and designating their actions is the prime purpose of astute video systems. Its motive is to acknowledge the actions and objectives of one or more multiple objects from a series of examination on the action of objects and their environmental condition. Human action recognition has gained popularity because of its wide applicability varying from Content- based Video Analytics, Visual Surveillance and Video Indexing etc. This paper depicts survey on person to person interaction, single person and multiple people action recognition and deciphers the methodologies and limitations of diverse methods for human action recognition are explored.*
***Keywords:** Human action recognition, challenges, applications, single person action recognition, human-human interaction recognition methods, multiple people action recognition*

I. INTRODUCTION

Human Action Recognition is a resourceful research issue in the area of computer vision. Due to the increase in number of surveillance cameras, Human Action Recognition system has gained its popularity. The objective of Human Action Recognition (HAR) is to identify actions of one or more humans from a series of examination on the actions of humans and their environmental condition. The system's major applications are not limited to sports, surveillance security, autonomous driving and video retrieval. Actions are elementary with respect to one single activity. Human Action Recognition systems conventionally follows a hierarchical manner. First step is to perform Background subtraction, feature extraction, tracking and detection which comes under the low level. Midlevel module is used to recognize followed by the reasoning engines on the high level based on the units of lower level. In the lower level, background subtraction is implemented on the extracted frames either by block-based, pixel-based or the combination of both. Mixture of Guassians, Guassian Modelling, Kalman Filter and Hidden Markov Model are commonly used pixel – background models. Block- based approach typically falls under Histogram Similarity, Normalized Vector Distance, Incremental PCA and Local Binary Pattern Histogram. After the detection of foreground, the feature extraction is performed using a technique where a human model is built for recognition of an action or else the global or local features or both features are extracted that aid in the numerical computation for action detection. During mid-level after the detection and tracking process, it is given to the classifier for the action recognition. After the action identification it is given to a high level that includes a reasoning engine that interprets the actions of performers from the multifarious works of various authors. This paper is structured as follows. Section 2 depicts about the various challenges in Human Action Recognition. Section 3 depicts about various applications. Section 4 depicts about single person action recognition methods. Section 5 describes person to person interaction recognition methods. Section 6 depicts multiple people action recognition methods and in section 7 inferences are made and future research direction followed by conclusion.

II. CHALLENGES IN HUMAN ACTION RECOGNITION

Despite of having a significant progress in Human Action Recognition, state-of-the-art algorithms still misclassify actions due to several major challenges in these tasks.

A. Intra and Inter-Class Variations

People behave differently for the same actions. For a given instance, for example, "running", a person can run slow, fast, or even jump and run. Likewise one action category may contain multiple different styles of human movements. In addition, same action in videos can be captured from various endpoints. They can be taken in front of the human subject, on top of the subject, or even on the side of the subject, showing appearance variations in different views. Furthermore, different poses shown by different people in executing the same action. All these factors will result in large intra-class and pose variations, which leads to confusion of existing action recognition algorithms. These variations will be even larger on real- world action dataset [1]. This encourages investigation of more advanced action recognition algorithms that can be deployed in real- world scenarios. Surveillance exist in different action categories. For instance, "running" and "walking" involve similar human action patterns. These similarities would also be challenging to differentiate for intelligent machines and consequently contribute to misclassifications.

B. Cluttered Background and Camera Motion

A number of Human Action Recognition algorithms work very well in indoor environments but not in outdoor uncontrolled environment due to background noise in the outdoor environment. In fact, most of existing action features such as interest points and histograms of oriented gradient [2] also encode background noise and thus degrade the recognition performance. The motion of the camera is another factor that should be observed in real-world applications. Due to significant camera motion, action features cannot be precisely extracted. To better extract action features, camera motion should be modelled and compensated. Other environment- related issues such as viewpoint changes, illumination conditions and dynamic background are also the challenges that results in prohibiting action recognition algorithms from being used in real time scenarios.

C. Uneven Predictability

Not all frames are equally discriminative. As shown in [3] a small set of key frames can be used to effectively represent a video indicating that lots of frames are redundant. However, action recognition methods require the beginning portions of the video to be discriminative in order to maximize predictability. Context information is transferred to the beginning portions of the videos [4] to solve this problem, but the performance is still limited due to insufficient discriminative information. In addition, actions differ in their predictabilities [4], [5]. As shown in [4], some actions need more frames to be observed while the remaining ones are instantly predictable. However, in practical scenarios, it is necessary to predict any actions as early as possible, especially in real time. This demand to create action prediction algorithms that can make accurate and early predictions for most of or all actions.

III. APPLICATIONS

Human Action Recognition algorithms enfranchise many genuine world applications. State-of-the-art-algorithms[6] have remarkably minimize the human labour in scrutinizing a tremendous scale of video data and provide better understanding on the current state and future state of an ongoing video data.

A. Video Retrieval

In recent times, due to thriving growth of technology, people can with ease upload and share videos on the Internet. However, administrating and retrieving videos according to the video content from a query is becoming a great challenge as most of the search engines use the associated text data to manage video data [7]. The text data, such as titles, tags, descriptions and keywords, can be obscure, incorrect and irrelevant making the video retrieval unsuccessful. An alternative method is to analyse human actio0ns in videos, as the majority of these videos contain such an important cue. For example, in [8], researchers in a bid to help children with autism created a video retrieval framework by computing the similarity between action representations and implemented framework in a classroom setting to retrieve videos of children suffering from autism. Compared to conventional Human Action Recognition task, the video retrieval task relies more on the retrieval ranking instead of classification [7].

B. Human –Robot Interaction

Human- Robot interaction is popularly applied in industry and home environment. Let's imagine that a person is interacting with a robot and asking it to perform tasks, such as "performing an assembling task" or "passing a cup of water" an interaction requires communications between robots and humans and visual communication is one of the most efficient ways [8].

C. Autonomous Driving Vehicle

Action prediction algorithms could be one of the potential and may be most important building components in an autonomous driving vehicle. Action prediction algorithms can predict a person's intention [8] in a short period of time. In a situation, a vehicle equipped with an action prediction algorithm can predict a pedestrian's future action in the next few seconds and this could be critical to avoid a collision. By analysing human motion at an early stage of an action using interest points or convolutional neural network [4], action prediction algorithms [4] can understand the possible actions of human by analysing the action progression without the need to observe the entire action recognition.

IV. SINGLE PERSON ACTION RECOGNITION METHODS

Single person action recognition is also used to analyse the motion of a player in a game. Single person action includes actions like walking, running, falling and loitering. R. Bodor, B. Jackson and N. Papanikolopoulos [9] have worked on video- based tracking of human and activity recognition. The image of the region inside the tracked blob of pedestrian is detected and tracked by a smart video system. The acquired images are arranges sequentially to make video. To estimate the velocity and the shape of motion of pedestrian; kalman filter is used. These pedestrian motions are classified as running, walking, loitering and falling. Warning sign is

signalled if a person enters a prohibited area, loiters for a long time, fall down and if a pedestrian exceeds walking speed. The limitation is that performance varies with the change in illumination. It also does not differentiate between objects moving with the same speed but with different means such as bicyclist and runner.

Temporal key poses for Human Action Recognition was done by A. Eweiwi, S. Cheema, C.Thurau and C. Bauckhage [10]. To identify human actions from videos, Motion History Images (MHI) and Motion Energy Images (MEI) temporal templates are used. Nearest Neighbour classifier is used to classify query videos. The experimental result for Weizmann dataset and MuHAVi is shown using leave-one-out cross validation. Accuracy obtained for MuHAVi dataset consisting of 8 actions is 98% and on MuHAVi dataset with 14 actions having similar setup as previous one gives 92% accuracy, 100% accuracy on Weizmann dataset consisting of 10 actions is obtained. In this approach the recognition rate decreases with severe change in camera view.

W. Lu and J. Little [11] have proposed a method to track the person of interst and to recognize action of that person. Previous time template is used to track interest region to estimate the current state of player.the tracking region in frames can be estimated by computing Principle Component Analysis- Histogram of Oriented Gradient (PCA- HOG) descriptor and Maximum Likelihood Estimation of previous observations. The experiments are performed for two games soccer and hockey. For hockey, images of 6 actions like skate right, skate left, skate in, skate out, skate left 45 , skate right 45 are collected. For soccer, the categories of actions used are run right, run left, run left 45, run right 45, run in/out, walk left, walk right, and walk in/out. For both the experiments, 10 possible templates for action are used.

V. TWO PERSON OR PERSON-OBJECT INTERACTION RECOGNITION METHODS

Two person interactions consist of actions like punching, pushing, kicking, handshake, hug and kiss. Human- object interactions like dial phone, answer phone, drink and eating activities need to be recognized. Many methods had been carried forward for this but still there is a necessity to improvise the results and overcome many challenges.

K. Slimani, Y. Benezeth and F. Souami [12] have presented work on human interaction recognition based on the co-occurrence of visual words. All people need to be detected and it is done manually. From two interacting performers, 3D-XYT volume is extracted.

The volume is represented by a set of tuple. Tuple is a collection of 4 things- time, spatial, position and word index. For two person interaction detection, two dimensional co-occurrence matrix is constructed. Co-occurrence matrix is changed to make invariance to the relative position of a person. UT-interaction dataset is used for experiment. For set1, Euclidean distance and k-nearest neighbour classifier (KNN) are used as distance function and for set2, SVM classifier with polynomial kernel is used. By the proposed average 40.63% accuracy on set1 and on set2 66.67%.

The experiment is carried out for only two performers with constant background.

The research work on recognition of interactions between human to human by ‘Dominating Pose Doublet’ was presented by S. Mukherjee, S. Biswas and D. Mukherjee [13].

The pose descriptors of detected humans are obtained by optical flow of video frames sequence. People appearances are detected by Histogram of Gradient (HOG).

The codebook for both the humans are created separately that consisted of pose descriptors. Taking into consideration dominating poses of two humans, Bipartile graph is created. The least no. of poses are the set dominating poses that are required to cover all the variation of poses.

Different codebooks having nodes representing poses and depending on the frequency of occurrence of the poses in the videos, edges have weights. On UT- interaction dataset 86.67% accuracy is obtained. The limitation is that the approach is implemented only for tow performer’s interactions. It can be extended to multiple human interaction.

A system to explore TV videos, a STIP- based model for the recogntio of human interactions is presented by M. Jimenez, E. Yeguas and N.Blanca [14]. From the videos spatio-temporal interest points (STIP) are acquired and can be selected based on dense sampling of STIP or Harris3D. All feature points are selected of Harris 3D and sense sampling of STIP is done only for the regions that consist of persons.

Dense sampling of STIP in person region gives good results. The volume descriptor HOG and HOF are calculated for STIP. For encoding BOW model is used. Support Vector Machine is used for the purpose of classification. TV Human Interaction Dataset (TVHID), UT Interaction dataset (UTID) and Hollywood -2 dataset are used for experiments. Performance accuracy on TVHID is 0.3661, 0.88 and 0.86 for set2 and set1 respectively of UTID and Hollywood – 2 dataset obtained 0.6077 accuracy.

VI. MULTIPLE PEOPLE ACTION RECOGNITION METHODS

Multiple people action recognition has gained popularity due to increasing need of security in case of video surveillance. Many researchers have tackled with the problems like abnormal behaviour recognition, pedestrian counting and crowd motion analysis.

G. Santhiya, K. Sankaragomathi and S. Selvarani [15] have presented work on abnormal crowd tracking and motion analysis. Crowded environment is very difficult for human to recognize live or through video surveillance as it connate be surveyed 24.7 by a human. Adaptive background modelling is performed for a given input video. Threshold is decided based on the pixels of the input video. This threshold is used for separation of background and foreground pixels crowd detection and blob analysis is carried out. Crowd detection is important for crowd density and in literature it is performed by two methods indirect and direct. To detect abnormal activities a model is created of crowd activities. The experiment is carried out on available UMN dataset.

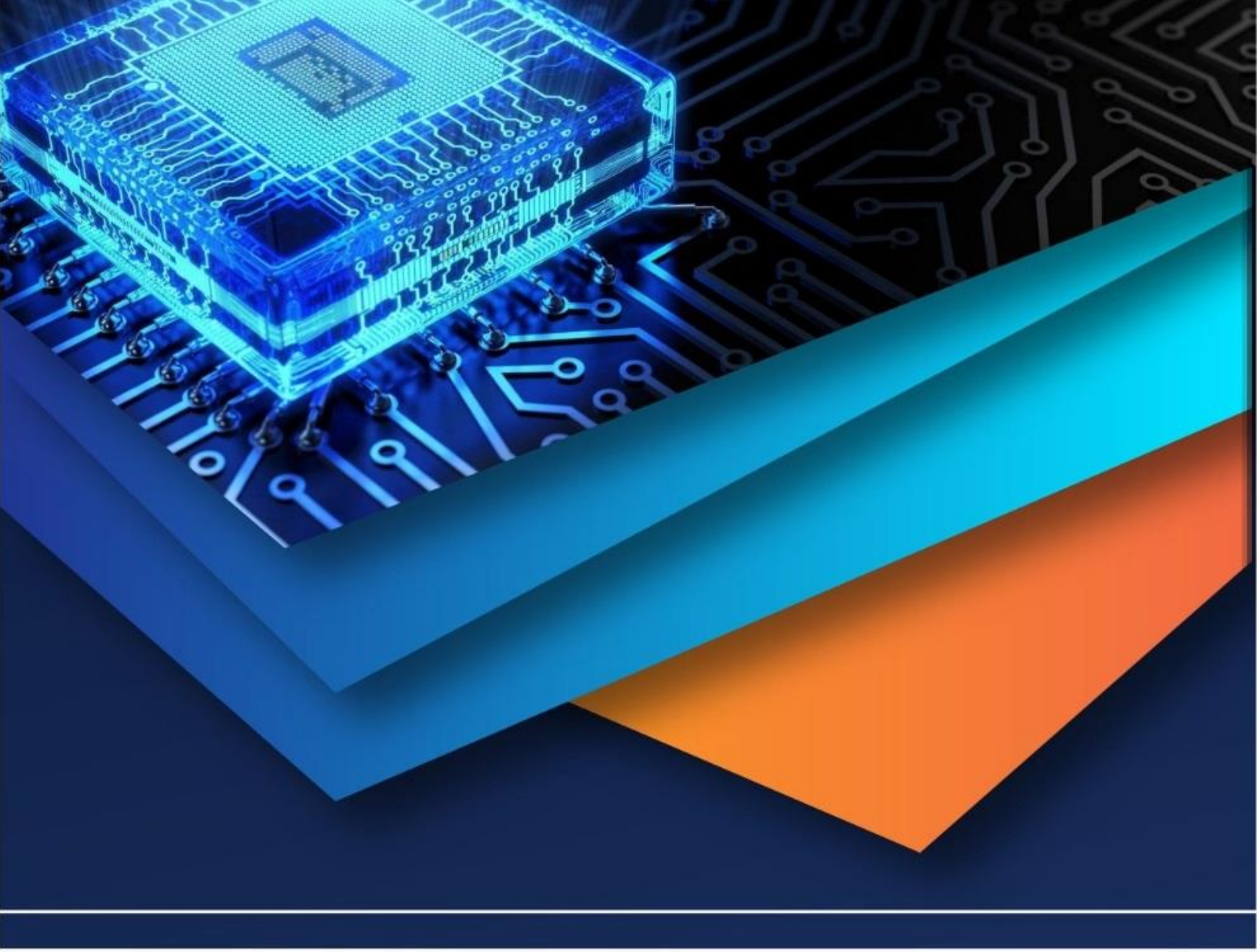
A real-time crowd motion analysis was proposed by Nacim and C. Djeraba [16]. In recent years, the need for autonomous video surveillance systems has increased a lot. A system in public areas is proposed to detect abnormal activities of crowd. Motion heat map is computed of the image which computes hot and cold regions. The high motion is represented by hot regions in the frame and cold region represent the low motion intensities. The experiment is carried out using real time videos of airport to monitor situations of escalator that consist of 20 videos of normal situation and 20 videos of collapsing situations. This approach is able to detect all collapsing situation.

VII. CONCLUSION AND FUTURE DIRECTION

This paper discusses the methods and limitations in the field of human action recognition. Hierarchical approach, Spatio-temporal interest point based, semantic descriptor based approaches are widely used for human action recognition. Thus the human action recognition methods conclude that the progress in the field of action recognition is developing and encouraging. In the future, there are some performance issues that need to be considered and solved for real time deployment. Many challenges like change in appearance, change in illumination, high computational cost, changing camera view point and low recognition rate need to be solved.

REFEERENCES

[1] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, “Activitynet: A large-scale video benchmark for human activity understanding,” in CVPR, 2015.
[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in CVPR, 2008.
[3] M. Raptis and L. Sigal, “Poselet key-framing: A model for human activity recognition,” in CVPR, 2013.
[4] Y. Kong, Z. Tao, and Y. Fu, “Deep sequential context networks for action prediction,” in CVPR, 2017.
[5] K. Li and Y. Fu, “Prediction of human activity by discovering temporal sequence patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 8, pp. 1644–1657, Aug 2014.
[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, “Temoral segment networks: Toward good practices for deep action recognition,” in ECCV, 2016.
[7] M. Ramezani and F. Yaghmaee, “A review on human action analysis in videos for retrieval applications,” Artificial Intelligence Review, vol. 46, no. 4, pp. 485–514, 2016.
[8] H. S. Koppula and A. Saxena, “Anticipating human activities using object affordances for reactive robotic response,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 14–29, 2016.
[9] B. Robert, B. Jackson, and N. Papanikolopoulos. "Vision-based human tracking and activity recognition", In Proc. of the 11th Mediterranean Conf. on Control and Automation, vol. 1, 2003.
[10] E. Abdalrahman, S. Cheema, C. Thurau and C. Bauckhage, "Temporal key poses for human action recognition", International Conference on Computer Vision Workshops, IEEE, pp. 1310-1317, 2011.
[11] W. Lu and J. Little, "Simultaneous tracking and action recognition using the pca-hog descriptor", 3rd Canadian Conference on Computer and Robot Vision, IEEE, 2006, pp. 6- 6. IEEE, 2006.
[12] K. Slimani, Y. Benezeth, and F. Souami, "Human interaction recognition based on the co-occurrence of visual words", Computer Vision and Pattern Recognition Workshops, IEEE, pp. 461-466, 2014.
[13] M. Jiménez, E. Yeguas and N. Blanca, "Exploring STIP-based models for recognizing human interactions in TV videos", Pattern Recognition Letters, Elsevier, vol. 34, pp. 1819-1828, 2013.
[14] S. Mukherjee, S. Biswas and D. Mukherjee, "Recognizing interactions between human performers by ‘Dominating Pose Doublet’ ", Machine Vision and Applications, Springer Berlin Heidelberg, vol. 25, pp. 1033-1052, 2014.
[15] G. Santhiya, K. Sankaragomathi and S. Selvarani, "Abnormal crowd tracking and motion analysis", International Conference Advanced Communication Control and Computing Technologies, IEEE, pp. 1300-1304, 2014.
[16] I. Nacim and C. Djeraba, "Real-time crowd motion analysis", 19th International Conference Pattern Recognition, IEEE, pp. 1- 4, 2008.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY