ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

Human Action Recognition from Images and Videos

Ramanpreet Kaur
Assistant Professor, Department of Computer Science,
Khalsa College for Women, Sidhwan Khurd, Ludhiana, India

Abstract: Automatically recognizing human actions in real world environment is receiving utmost attention due to its large number of applications in a variety of domains. Such as Surveillance Systems for Security, Healthcare Systems, Personal Biometric Signature, Marketing, Content-based Search, Sports, and Human – Computer Interaction. Owing to its great potential, the problem of how to automatically discover and recognize human activities from video and image data has become a popular topic and attracted many researchers in the computer vision and pattern recognition area. An automatic action recognition process is basically the imitation of natural action recognition process in human brain. It is possible for computers to imitate the action recognition process of the human mind with the use of action recognition methods that can deal with variations of actions in images. The systems can act as intelligent as human brains with the use of effective action recognition methods.

Keywords: Human Action Recognition, Activity, Images, Surveillance

I. INTRODUCTION

Over some past decades, the expeditious advancements are performed in the field of technology. Along with these advancements a large number of devices such as computers, smartphones, digital cameras etc. are available for everyone. By performing repetitive and data-intensive computational tasks and extending possibilities to communicate, these devices have become prodigious part of our daily life. A large set of data captured by these devices is available in the form of images and videos. The images and videos are indispensible part of everyday life due to their easily accessible property. Videos and images are meaningful ways to capture and convey the information easily. Due to the faster internet access and growing storage capacities, it is easy to directly publish and share videos with others. According to the statistics of YouTube company over 400 hours of videos are uploaded to the YouTube every minute [1]. Despite the large availability of videos and image data, the techniques to analyze these in an automatic way are rather limited. The computer vision systems have fewer capabilities than the human vision systems. So there is an immense need to develop efficient systems to analyze and process images and videos in an automatic way.

Numerous works and techniques have been proposed in past few decades within the field of human action and activity recognition. Since action and activity recognition from images and videos is a challenging task, a lot of approaches have been considered for achieving higher accuracy in challenging environments. But it is still required to develop some efficient techniques for human action recognition from images and videos.

The remainder of this paper is organized as follows; second section describes human action recognition, third section contains a detailed discussion about the structure of human action recognition system. In the fourth section various methods used for action recognition process are explained.

II. HUMAN ACTION RECOGNITION

Human Action or Activity Recognition is used to identify the objectives of one or more actions from a series of examined actions of objects and their environmental conditions. Accurate activity recognition involves a set of challenges because human activity is complex and highly diverse. The terms action and activity are often used interchangeably. An Action is a basic human motion, e.g. jump, and walk. On the other hand, an activity is a complex human motion that can be defined as a combination of several basic actions, e.g. Drinking water can be defined as a sequence of actions, opening a bottle, pouring water into a glass and then drinking [2]. Human actions can be grouped into four different parts on the basis of their complexity. These parts are Gestures, Actions, Interactions and Group activities. Gestures are basic movements of human

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

body parts. They represent a meaningful motion of a person e.g. stretching an arm, raising a hand. An action is a single human's activity that can be defined as a combination of several gestures, e.g. walking, jogging, jumping etc. interaction can involve two persons or a person and object. For instance, a person is picking a suitcase, a person is snatching some object from another person, two persons are fighting etc. Group activities can be performed by a group of persons, a seminar, fight between two groups, crowd etc. The action recognition from group activities is a really difficult task.

III. HUMAN ACTION RECOGNITION SYSTEM

The Fig. 1 represents the general structure of an action recognition system.

The Human action recognition system is divided into two parts; Training stage and testing stage. In training stage, the first step is image acquisition. Image acquisition means capturing an image and converting it into digital form. The captured image may contain irrelevant information. After the image acquisition, the next stage is pre-processing stage. At this stage object segmentation is performed on each extracted frame to extract the target object from input data.

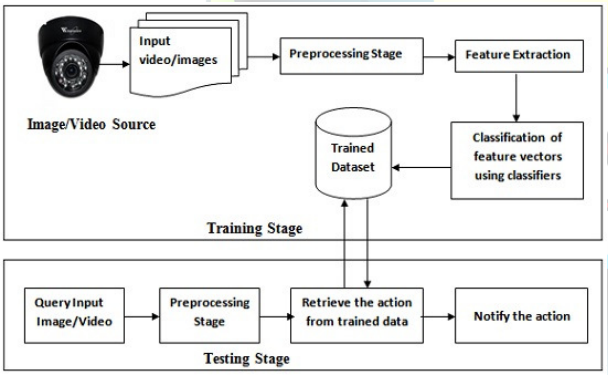

Fig. 1 A general structure of human action recognition system

The object segmentation can be classified into two categories; static camera object segmentation and moving camera object segmentation. In the case of static camera object segmentation the camera is fixed at a location and captures the images from a fixed angle, the background is fixed. The technique for static camera object segmentation is background subtraction [6, 7] , it is defined as an efficient method.

This method establishes a background image without any foreground object is first established. Then the current image in analysed video sequences is subtracted from the background image to extract the foreground objects. This is also known as a low-level stage, background subtraction is performed on extracted frames either by pixel-based, block-based or the combination of both. Gaussian Modeling, Mixture of Gaussians [8, 9], Kalman filter, and Hidden Markov Models are commonly used pixel-based background models. Block-based approach usually falls under Normalized Vector Distance, Histogram Similarity, Incremental PCA and Local Binary Pattern Histogram [10]. For the moving camera segmentation, the camera is moving with a person, it is more complex than static camera object segmentation. It considers the camera motion and background change. The simple background extraction method is not applicable here. The temporal difference [11, 12] method is used for moving camera object segmentation. It is the difference between the consecutive image frames.

The third step is feature extraction. It deals with the features of the tracked objects. The features of the objects such as shape, pose, motion etc. can be classified into four groups such as Space-time information, frequency transform, local descriptors, body modeling [3]. The Feature extraction can be performed by using feature extraction methods (local or global features) or model-based approach. After extracting the features, it is passed to next level, where classification is performed**.** At this level, activities can be recognized by using various activity detection and classification algorithms. In the testing stage, a video or image is entered in the system, after performing the pre-processing, it is passed to the next stage. The system tries to retrieve that image from a trained dataset. After that, the notification related to the action in the image is presented.

IV. HUMAN ACTION RECOGNITION TECHNIQUES

Human Action Recognition methods can be divided into two categories on the basis of techniques used for action recognition. This section presents various techniques used for action recognition.

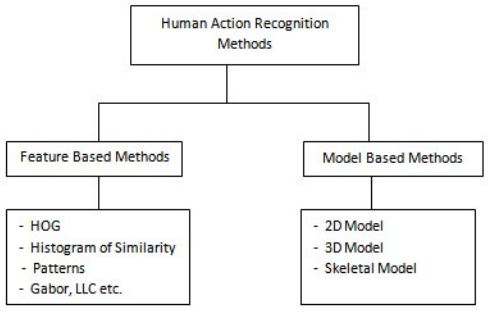

Fig. 2 Human action recognition methods

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

A. *Feature based Approach*

In this approach, after the pre-processing of input images, feature extraction methods are applied to them to extract the features. At this stage, the important characteristics of the image frames are extracted and represented in a systematic way as features. Feature extraction and representation have a crucial influence on the performance of recognition, it is essential to select or represent features of image frames in a proper manner. There are two types of feature extractors; Global feature extractors and Local feature extractors.

Global feature extractors are sensitive to noise, occlusion, and variation of a view point. The Space-Time Volume (STV) [13-15] concatenates the consecutive silhouette of objects along the time axis. The continuity of human action can be captured from the 3D XYT volume. The STV is limited to non-periodic activities. They consider space and time information of video sequences. They deal with spatial-temporal domain. Discrete Fourier Transform (DFT) [16] deals with frequency domain information. They belong to the global features, so they consider the whole image, not the parts of the image. The DFT of image frames spatially captures the image intensity variation along with the spatial and temporal information.

Local feature extractors are more robust to noise and occlusion, to rotation, and scale as compared to the global methods [3]. Histogram of Oriented Gradients (HOG) [17], Scale-Invariant Feature Transform (SIFT) [18], are local feature descriptors.

B. *Model based Approach*

In this approach, human model is constructed for action recognition [10].These methods directly or indirectly model human body, to which the pose estimation and body part tracking techniques, can be applied. Feature based methods do not fully capture whole body actions. Therefore, some human modeling methods [3] have been proposed to model the human body, including simple blobs, 2D body modeling, and 3D body modeling. In modeling based approaches, 2D/3D pose estimation is performed, after that 2D/3D coordinates of a human body can be converted into feature representations such as Boolean features [19], geometric relational features [20] etc.

After applying any of the best-suited methods for feature extraction and representation, the next step in human action recognition is classification. Different types of classification techniques are available for this purpose. Different types of classifiers need different types of feature representations. Some of the classifiers are Dynamic Time Wrapping (DTW), Hidden Markov Model (HMM), Support Vector Machine (SVM), Relevance Vector Machine (RVM), Artificial Neural Networks (ANN) etc.

V. CHALLENGES IN HUMAN ACTION RECOGNITION FROM VIDEOS AND IMAGES

Human action recognition has a large number of applications for real-world environments. It involves a large set of challenges that affect the performance of the system. Some of the research challenges related to the human action recognition are presented below:

A. *Multiple Viewpoint*

In most of the cases, it can be assumed that actions are performed from a fixed viewpoint. But it is not possible in real world. The motion pattern, location, posture can vary for the same action. Viewpoint variation is a major challenge in Human action recognition.

B. *Occlusion*

The existing systems require that the action being performed should be clearly visible in the video sequences. In a real-world surveillance video, this is not possible because of the number of people in the field of view of the camera. Occlusions can be either self-occlusions or those created by other objects in the field of view of the camera during the video capture. This poses a big challenge to the research community because not all the body parts performing the action are visible in the video sequence [4].

C. *Speed of Action*

The execution rate of an action depends on every individual. It is also not guaranteed that the same person will perform the same action at the same speed every time.

D. *Anthropometric Variations*

Each person has different body size, proportion, and comfort zone while performing an action. For example, a waving gesture of a person might involve moving the hand above the head and then wave the hand, but another person might not move his hand above his head and would just wave from a shoulder height. Thus, researchers developed a generalized approach to capture and handle these variations [4].

E. *Cluttered Background*

Dynamic or cluttered background is a form of distraction in the video sequence from the original action of interest as it introduces ambiguous information [21]. Flow-based methods that calculate motion are affected as they detect unwanted

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

background motion along with the actually required motion. In addition, color-based and region-based segmentation approaches require uniform non-varying background for reliable segmentation and tracking of the foreground object. To avoid the anomaly introduced, most applications assume a static background or a method to handle background segmentation from the videos prior to processing [22, 23].

F. *Camera Motions*

In most action recognition cases, researchers assume static cameras, which might not be the case in unconstrained systems. Camera motion severely affects motion features since erroneous and misleading motion patterns are induced in the videos. Shape features generally require a good tracking mechanism, background model, and stationary cameras [15, 24]. Background subtraction required by these features is affected by moving cameras [4].

G. *Semantic Gap*

Many detection results depend on the semantics of the activities. It is often very difficult to detect high-level semantics of an activity from low-level features.

H. *Low Resolution*

In the case of human action recognition from videos, the resolution is also a major factor that affects the technique applied. If a video is having low resolution, the extracted frames are also of very low quality and it becomes very challenging task to perform action recognition.

VI. CONCLUSION

This paper has presented some of the basic details about human action recognition from videos and images. Various approaches have been developed for action recognition, but action recognition in real life environments is still a challenging task. Most of the work done is giving good results for static and less challenging datasets. There is an immense need to develop systems for action recognition that can work in real life environments.

REFERENCES

[1] https://youtube.googleblog.com
[2] M. Selmi, M. A. El-Yacoubi, and B. Dorizzi, "Two-layer Discriminative Model for Human Activity Recognition," IET Computer Vision,vol. 10, no. 4, pp. 273-278, 2016.
[3] S. R. Ke, H. 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, pp. 88-131, 2013.
[4] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human Action Recognition With Video Data: Research and Evaluation Challenges," in the proceedings of IEEE Transaction on Human-Machine Systems, vol. 44, no. 5, pp. 650-663, 2014.
[5] R. Bodor, B. Jackson, and N. Papanikolopoulos, "Vision-Based Human Tracking and Activity Recognition," in the proceedings of the 11th Mediterranean Conference on Control and Automation, vol.1, pp. 18-20, 2003.
[6] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," in the proceedings of IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, 1997.
[7] M. Seki, H. Fujiwara, and K. Sumi, "A Robust Background Subtraction Method for Changing Background," in the proceedings of IEEE Workshop on Applications of Computer Vision, vol. 2, pp. 207-213, 2000.
[8] H. Permuter, J. Francos, and I. Jermyn, "A Study of Gaussian Mixture Models of Color and Texture Features for Image Classification and Segmentation," Pattern Recognition, vol. 39, no. 4, pp. 695-706, 2006.
[9] S. Yoon, C. S. Won, K. Pyun, and R. M. Gray, "Image Classification using GMM with Context Information and with a Solution of Singular Covariance Problem," in the proceedings of Conference on Data Compression, Snowbird, UT, 2003.
[10] T. Subetha and S. Chitrakala, "A Survey on Human Activity Recognition from Videos," in the proceedings of International Conference on Information Communication and Embedded Systems, vol. 15, no. 3, pp. 1-7, 2016.
[11] D. Murray and A. Basu, "Motion Tracking with an Active Camera," in the proceedings of IEEE Transaction on Pattern Recognition and Analytical Machine Intelligence, vol. 16, no. 5, pp. 449-459, 1994.
[12] J. Y. L. Kye Kyung Kim, S. H. Cho, and H. J. Kim, "Detecting and Tracking Moving Object using an Active Camera," in the proceedings of 7th International Conference on Advance Communication and Technology, vol. 2, pp. 2-5, 2005.
[13] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," in the proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65-72, 2005.
[14] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes", in the proceedings of 10th IEEE International Conference on Computer Vision,vol. 2, pp. 1395-1402, 2005.
[15] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal Shape and Flow Correlation for Action Recognition," in the proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recoginition, pp. 1-8, 2007.
[16] S. Kumari and S. K. Mitra, "Human Action Recognition Using DFT," in the proceedings of the 3rd National Conference on Computer Vision, Pattren Recoginition, Image Processing and Graphics, vol. 2, pp. 239-242, 2011.
[17] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in the proceedings of IEEE Computer Society Conference on Computer Vision and Patteren Recoginition, vol. 1, pp. 886-893, 2005.
[18] S. Lowe and D. G., "Distinctive Image Features from Scale-Invarient Keypoints," International Journal of Computer Vision,vol. 60, no. 2, pp. 91-110, 2004.
[19] M. Muller, T. Roder, and M. Clausen, "Efficient Content-based Retrieval of Motion Capture Data," ACM Transactions on Graphics, vol. 1, no. 212, pp. 677-685, 2005.
[20] H. L. U. Thuc, P. Van Tuan, and J. N. Hwang, "An Effective 3D Geometric Relational Feature Descriptor for Human Action Recognition," in the proceedings of IEEE Conference on Computing

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

and Communication Technologies, Research Innovation and Vision for the Future, pp. 1-6, 2012.

[21] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories using Spatial-Temporal Words," International Journal of Computer Vision, vol. 79, no. 3, pp. 299-318, 2008.

[22] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "Human Action Recognition using Robust Power Spectrum Features," in the proceedings of IEEE Transaction on Image Processing, United Kingdom, pp. 753-756, 2008.

[23] A. Iosifidis, A. Tefas, and I. Pitas, "Neural Representation and Learning for Multi-view Human Action Recognition," in the proceedings of IEEE World Congress on Computational Intelligence, Brisbane, Australia, pp. 10-15, 2012.

[24] Z. Jiang, Z. Lin, and L. Davis, "Recognizing Human Actions by Learning and Matching Shape-Motion Prototype Trees," in the proceedings of IEEE Transaction on Patteren Analysis and Machine Intelligence, vol. 34, no. 3, pp. 533-547, 2012.