DOI: https://doi.org/10.53555/nnbma.v1i2.131 **Publication URL: https://nnpub.org/index.php/BMA/article/view/131**

A REVIEW ON CLASSIFICATION TECHNIQUES FOR HUMAN ACTIVITY RECOGNITON

Sonali[1]*, Ashok Kumar Bathla[2]
*[1]*Research Scholar, CE Deptt. YCOE, Punjabi University, Patiala, Indiakaushiksonali19@gmail.com*
[2]*Assistant Professor, CE Deptt. YCOE, Punjabi University, Patiala, Indiaashokashok81@gmail.com*

Corresponding Author: -
Email: Indiakaushiksonali19@gmail.com

Abstract: -

Recognizing human actions from video sequences has many important applications like video surveillance, patient monitoring, human computer interaction, dance choreography analysis, analysis of sports events and entertainment environments. It involves processing the video into frames firstly and finding out the interest points, then extracting the features and lastly specifying and labelling the videos following an appropriate classifying approach like Support Vector Machine, bag of words or nearest neighbour. This paper provides a detailed overview of various state-of-the-art research papers on human activity recognition using different types of classifiers. We surveyed various challenges exhibited by computer vision researchers like the problem of occlusion, 2D/3D pose estimation, variations in viewpoints, human body modelling especially of a person who is paralyzed or injured. From this survey, we can make conclusion of various advantageous and disadvantageous facts about different classifiers used in the detection and classification task.

Keywords: *action recognition, classification, support vector machine, nearest neighbour, bag of visual words*

1. INTRODUCTION

Human actions are not merely due to the movement or motion of body-parts of a human being, rather it is the depiction of one's intentions, behavior and thoughts. "Action Recognition" as the term itself is self-suggesting, it is the recognition of an activity or action by using a system that analyzes the video data to learn about the actions performed and uses that acquired knowledge to further identify the similar actions.

Recognizing human actions and activities is a key-component in various computer applications like video-surveillance, healthcare systems, recognition of gestures, analysis of sports events and entertainment events. In a Video Surveillance environment, the detection of various unusual or abnormal actions in a video sent to court for criminal investigation.

Likewise, in a healthcare environment, 'patient monitoring', the process of automatic recognition of a coma patient's actions can help the concerned doctors to check out the patient's recovery status.

Furthermore, in an entertainment environment, the efficient recognition of sports actions of a person is done to create an avatar for him. He can then play over the computer system imitating his real-world actions. However, in today's era, we are having many traditional approaches for human action recognition, be it be low-level features recognition on basis of trajectory, shape, color or bag-of-visual-words (BoVW), still it is really hard to semantically classify the actions in videos. Although a lot of work has been done to recognize the human actions but still it remains a challenging task to effectively uncover all the actions performed in the video due to different challenges opposed [15].

As far as we know, human action recognition poses various challenges like occlusion, execution rate, cluttered background, camera motion and variation in view-point. Due to varying datasets and testing strategies a clear distinction could not be made amongst the proposed techniques, all of the pre-existing techniques respond in a different way to distinct problems [16]. Samples taken up from 2008 TRECVID surveillance event detection dataset are shown in Figure1.

Figure 1: -Action recognition from 2008 TRECVID surveillance event detection dataset [16]

2. CLASSICATION IN HUMAN ACTIVITY RECOGNITION

Generally speaking, the task of human activity recognition can be divided into three levels comprising of pre-processing and object segmentation, feature extraction and representation and activity detection and classification. The pre-processing stage involves the extraction of frames from the video as most of the previously done work in the field of human activity employs a frame-by-frame processing. Segmentation is done to extract the target object from the frames depending upon the camera mobility from which the videos were captured. Once the region of interest (ROI) is obtained from a frame, feature extraction is done where features like color, silhouette, shape are extracted. The features could be space-time information, body modeling, local descriptors etc [6]. Then comes the classification which helps to recognize the human activities on basis of the features extracted. The classifiers use to recognize and classify the actions are SVM, KNN, DTW, HMM etc [6].

2.1 Support Vector Machine

Support vector machine is a supervised learning model used for classification. It analyzes data and recognizes patterns by classifying them to be belonging to a particular class. Given a data to be classified and the training set, SVM classifier constructs a model which assigns the data into one of the categories. Human activity recognition is modeled as a multidimensional, multiple actions classification problem where we have one class for each action and our task is to assign a class label to a given activity or action. So, in this way one can use SVM for action classification. However, there are various supervised learning algorithms prevalent from earlier times to train the action recognizer, to recognize the motion patterns over time but SVM has a higher generalization capability and provides high accuracy. SVM creates a

hyperplane for classifying the data into a high dimensional space for separating the data with different labels as shown in figure 2. On each side of the hyperplane created initially, two separate hyperplanes are created. SVM tries to find that hyperplane which maximizes the distance between the two parallel hyperplanes. A wisely done separation means largest distance between the hyperplane and the nearest training data point of any class [13].

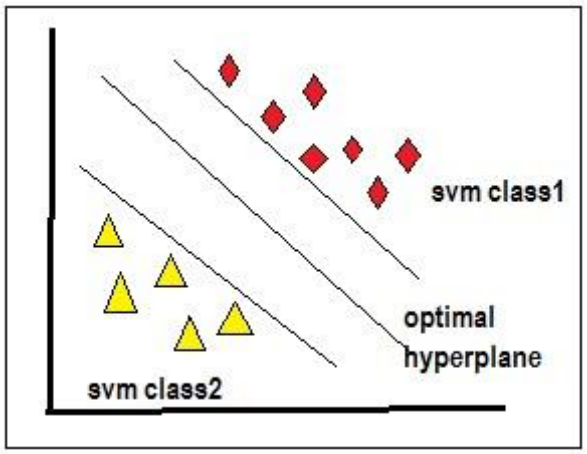

Figure 2: - Classification in Support Vector Machine

2.2 Nearest neighbor (KNN)

K- NN classifier measures the distance between the image or frame representation obtained from an observed sequence of video and the training set. The most common sequence from the training set is chosen for classification. NN classification is done either at frame level or for the whole video data sequence. K-NN is termed as instance-based learning, or lazy learning algorithm where the function is only approximated locally and all computations are deferred until classification of features. It is amongst the simplest of all machine learning algorithms wherein the features extracted are classified by a majority vote of its neighbors, with the feature or object being assigned to that class, which is the most commonly occurring in its k nearest neighbors. We assume k to be a positive integral value, which is not in general very large**.** Most of the methods use NN classification along with dimensionality reduction or by classifying global features using Euclidean distance. The different distance metric that can be used for NN classification are minimum mean frame-wise distance, frame order preserving variant, parametric and non-parametric density functions, etc. The drawback of K-NN is that for large training sets, computing distances and making comparisons can be very expensive [5].

2.3 Bag of words

(a)Features and visual words: This approach uses 3D space time local features to predict the human activities can be predicted. Firstly, the feature extractor converts the input video into 3D by concatenating the image frames along the time axis. Motion changes are then located. The local features extracted then form clusters based on their appearances i.e. feature vector values. The clusters of features thus formed is called as 'visual words. Once the clusters are formed, we can then use k-means clustering algorithm to form the visual words by extracting features from the given set of videos [11].

(b) Integral bag of words: Integral bag-of-words is a probabilistic activity prediction approach. It is used to construct integral histograms to represent the human activities. In order to predict the ongoing activity being carried out in a video, we compute histograms of visual words for those activities. Integral bag-of-words method is a histogram-based approach, in which the feature histogram representations are compared with histograms of video being tested in order to measure the similarities among them. It computes the likelihood P(O|Ap,d) where 'O' is the video observation, 'Ap' is activity and d is the progress level. The advantage of using the histogram representation is that even if the scale is varying, the noise can be handled. For all (Ap,d), the histograms are computed and compared with the histogram of testing video. Features can be represented in a histogram by taking a set of 'k' histogram bins, where k is the number of features extracted saved as visual words. In a video, each histogram bin represents the number of extracted features of the same type and the feature histograms are averaged to compute the histogram representation of activity model (Ap,d) which describes the expected number of corresponding visual words of occurrences. Integral histogram is a function of time which states how the histogram values changes as the duration of observation varies [11].

3. RELATED WORK

After a detailed extensive study on Human action recognition many approaches have been reported in the literature. One of the excellent demonstrations of human activity analysis and many of the related aspects of this task has been presented in Aggarwal et al. [1]. We have observed that most of the methods used previously recognizes the simple human activities like standing, bending, walking, running, sitting etc but did not focus on recognizing activities like sports events, dance choreography, patient monitoring etc. Various techniques are now used by using different types of classifiers or by combining one or more classifiers for identifying complex activities. Some of them seemed better than the others but quite complex or time-consuming at times. Ke et al. presented in detail the three aspects of human activity recognition

and addressed them as three critical processing stages mainly the human object segmentation which is the extraction of interest points detection from the data using methods like background subtraction, temporal difference consideration etc. The feature extraction and representation process were described. The working of each of the Activity detection and classification algorithms like the SVM, HMM, nearest neighbor, DTW etc was also discussed and stated in the paper [6]. Gorelick et al. represented actions as space-time shapes using a relatively simple classification scheme of nearest neighbors classification and Euclidian distance measure [4]. The various advantages provided were resolving the problems faced due to occlusions, significant changes in view-point and scale and various irregularities in low quality videos. Vemulapalli et al. inspired by the human action recognition based on real time skeleton estimation algorithms of the past, modeled the features representation task. The features or actions were modeled as curves and the classification of curves or the features extracted is done using a combination of dynamic time warping and linear SVM classifiers [14]. With an increasing demand of work in human action recognition from videos different researchers were expressing their own views. Brendel et al. gave another different kind of approach where which is exemplar based [2] also came meanwhile where human actions like cycling, running, swinging etc. were represented as short time series. The approach was almost common as used in existing algorithms of the times, but the difference was representing the video data as temporal sequences of learned codewords.

Sadanand et al. devised a method named as 'action bank' that leveraged on the fact that a large number of small action detectors, could yield high-leveled semantically rich features which are superior to low-level features in video data [12]. Every researcher has his own technique for extracting and classifying features for recognizing human actions. Wang et al. proposed the usage of high-level action units to represent human actions in videos and, based on such units, a novel sparse model is developed for human action recognition. Three interconnected steps are carried out, Firstly, a contextaware descriptor, named locally weighted word context is used. Secondly, from the statistical values of the contextaware descriptors, learning of the action units using the matrix factorization, which leads to a part-based representation and encodes the geometrical information was done. Lastly in order to suppress the noise, a sparse model was used in it [15]. Besides just recognizing simple human actions Yang et al. presented an approach to detecting basic human actions from surveillance videos. Actions like making cell phone calls, putting down objects, and hand-pointing were tested effectively on 2008 TRECVID surveillance event detection dataset. It was also generalized on KTH and Weizmann datasets [16]. Rani et al. suggested the method of tracking region-of-interest in a video sequence by estimating the similarity measure between the frames. Color feature extraction and geometric feature extraction are also used for the same [10]. The drawback of this approach was that, it could track only one object in a video sequence. Owing to the problems encountered while treating both the pose-estimation and detection tasks as separate tasks, Maji et al. presented a Poselet activation vector [7] for representing the different poses of people in various challenging datasets and also detecting the person at the same time. It is quite suited for 3D pose problem of persons which remained quite a challenging task at that time. A comparative study of different types of object segmentation methods like shape-based, texture based, appearance based etc. is stated in detail in Paul et al. which draws up an outline for us to choose the features to be classified based on their pros and cons. Chua et al. gave a fusion-based framework of the sum-rule and fuzzy-KNN, where motion and shape were the main features used for classification [8]. Ramanathan et al. provided an overview of the existing methods based on their ability to handle these challenges as well as how those methods could be generalized and their ability to detect abnormal actions. The authors demonstrated how the vision-based human action recognition is affected by several challenges due to view changes, occlusion, variation in execution rate, anthropometry, camera motion, and background clutter. The various challenges like occlusion and cluttered background are identified and efforts to resolve them are quite high but still they specify that a lot of improvement could be made in the real-world scenario [9].

4. COMPARATIVE ANALYSIS

This section shows up for the advantageous and disadvantageous features of the classifiers SVM, BoVW, KNN discussed in previous section. The areas where each of them is applicable are also described in Table 4.1.

Table 4.1 Comparative study of classifiers

Classifier	Description	Advantage	Disadvantage	Application areas
SVM	It is a Supervised learning model used to analyze data and recognize the patterns used for classification	Binary classifier basically, that performs linear and non linear Classification for different forms of data, be it be images, videos, texts etc.	The Size and Speed of training and testing database is quite insufficient that some features remain unclassified.	Human actions recognition, Text categorization, Image classification
BoVW	It is a Probabilistic activity prediction-based classifier used to construct visual word histograms	Quite Easy to implement and use as based on simple probabilistic measures.	Classification is difficult as it ignores spatial relationships among the patches type data important for representation.	Images, videos and documents classification
KNN	Instance based classification based upon similarity and distance measures between image representation from an observed sequence and the training set data.	Simplest classification technique and robust to almost all types of pattern classification.	It is comparatively costlier and for large training set, computing distances can be expensive	Human actions recognition, Text categorization, human gait pattern recognition, image classification

5. CONCLUSION

In this paper, different classification techniques for human activities recognition have been discussed. Each technique is better suited than the other for different types of activities in different application areas. On an average SVM performs better classification when we need a linear classification but the size of data is quite large. KNN, as discussed provides higher level of abstraction with high accuracies but time and complexity increase as compared to SVM. We also saw one more classifier 'visual words', which builds up histograms of extracted features and the classification is based on probabilistic model. Each technique has its own accuracy rate. For further work, to achieve better accuracies, more than one classifier can be combined together for performing better classification and recognizing activity in the videos.

REFERENCES

[1].Aggarwal J. K., Ryoo M. S., "Human Activity Analysis: A Review", Journal on ACM Computing Surveys (CSUR), 2011, Vol.43, No. 3, pp.1-47.

[2].Brendel W., Todorovic S., "Activities as time series of human postures", Proceedings of ECCV, Crete, 2010, Vol.6312, pp.721-734.

[3].ChuaT. W., Leman K., Pham N. T., "Human Action Recognition via Sum-Rule Fusion of FuzzyK-Nearest Neighbor Classifiers", IEEE International Conference on Fuzzy Systems (FUZZ)", Taiwan, 2011, pp.484-489.

[4].Gorelick L., Blank M., Shechtman E., Irani M., Basri R., "Actions as Space-Time Shapes", IEEE Tenth International Conference on Computer Vision (ICCV), Beijing, 2005, Vol.2, pp.1395-1402.

[5].Kaghyan S., Sarukhanyan H., "Activity Recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer", International Journal on Information Models and Analyses", 2012, Vol 1, pp. 146-156.

[6].Ke S. R., Thuc H. L. U., Lee Y. J., Hwang J. N., Yoo J. H., Choi K. H., "A Review on Video-Based Human Activity Recognition", 2013, Vol. 2, pp.88–131.

[7].Maji S., Bourdev L., Malik J., "Action Recognition from a Distributed Representation of Pose and Appearance",IEEE Conference on Computer Vision and Pattern Recognition (CVPR),Providence, 2011, pp.3177-3184.

[8].Paul M., Haque S. M. E., Chakraborty S., "Human detection in surveillance videos and its applications -a review", Springer EURASIP Journal on Advances in Signal Processing, 2013, pp.1-16.

[9].Ramanathan M., Yau W.Y., Teoh E.k., "Human Action Recognition with video data:Research and Evaluation Challenges ", IEEE Transaction on Human-Machine Systems, 2014, vol.44, no.5, pp.650-663.

[10]. Rani T.J., Priyadharsini S.S., "Region of Interest Tracking In Video Sequences", International Journal of Computer Applications ,2010, Vol. 3, No.7, pp. 32-36.

[11]. Ryoo M.S., "Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos", IEEE International Conference on Computer Vision, 2011, pp.-1036-1043.

[12]. Sadanand S., Corso J. J., "Action Bank: A High-Level Representation of Activity in Video", IEEE Conference on Computer Vision and Pattern Recognition (CVPR),Providence,2012, pp.1234 –1241.

[13]. Schuldt C., Laptev I., Caputo B., "Recognizing Human Actions: A Local SVM Approach", IEEE International Conference on Pattern Recognition, 2004, Vol 3, pp. 32-36

[14]. Vemulapalli R., Arrate F., Chellappa R., "Human Action Recognition by Representing 3D Skeltons as points in a lie group ", IEEE Conference on computer vision and

[15]. Wang H., Yuan C., Hu W., Ling H., Yang W., Sun C.,"Action Recognition Using Nonnegative Action Component Representation and Sparse Basis Selection", IEEE transactions on image processing, 2014, Vol. 23, No. 2, pp.570-581.

[16]. Yang M., Lv F., Xu W., Yu K., Gong Y., "Human Action Detection by Boosting Efficient Motion Features" IEEE 12th International Conference on Computer Vision Workshops (ICCV), Kyoto,2009, pp. 522-529.