Human Action Recognition based on Support Vector Machines

[1]R. Shreya, [2]Dr. S. Ravi[2*], [3]R. Jayasri[3]
[1]P.G. Student,[2]Assiatant Professor,[3]P.G. Student
[1, 2,3] Author Department of Computer Science, School of Engineering and Technology,
Pondicherry University, Puducherry - 605 014, INDIA,

Abstract: Accurate Human Activity Recognition (HAR) is very important activity in surveillance system and sports areas. Action classification is crucial part in HAR. Support vector machine (SVM) is one of the most powerful and robust algorithm in machine learning and it is commonly used for problems like classification, regression and pattern recognition. By separating the distinct class with a maximum possible wide gap, SVM tries to predict the respective class given a set of input data. Because of its higher prediction capabilities SVM has gained increasing attention in the remote sensing, image processing, robotics, pattern recognition and many others community. In this paper, the study and analysis is provided on human action recognition based on support vector machines by analyzing the advantages, disadvantages and accuracy. This survey will provide helping hand to the budding researchers who are started doing research in this field and that is the primary aim of this paper.

***IndexTerms* - Human Action Recognition, Support Vector Machines, depth motion map**

I. Introduction

Nowadays, Human Action Recognition (HAR) in video is playing very crucial role and particularly in the field of computer vision which attracts a lot of researchers over the past decades. HAR has a many applications such as Surveillance, human machine interfaces, sports video analysis, and video retrieval. The main problem in human action recognition by using sensors, is predicting the persons what they are doing and recognize action is. The movements of the people can be generally indoor activities such as sitting, jumping and standing. Accelerometer sensors are generally located on different subjects such as a smart phone or vest to record data generally in three dimensions. If the action of a person is detected and recognized, then artificial intelligence system can help us further. Once the data collected from sensors, the human action should be related.

Generally, it is not an easy task because relating sensor data clearly to specific actions is difficult. It became challenging task technically because the collection of data from sensors is large in quantity. Relating the large amount of data with specific actions is challenging task technically. Human action recognition approaches consist of two steps, namely, analysis and recognition of the actions from a particular given video input. The general structure of any HAR is described as three operation levels. The first level is low level in which feature extraction, detection kind of basic operations will be done and second is middle level which will be done after tracking and detection and finally third one is high level finding the actions using reasoning engine [1].

The selection of the right classification method plays a major role in human action recognition. Many pattern recognition techniques are available such as k-Nearest Neighbor (k-NN), Extreme Learning Machine (ELM) and Support vector machines will help in the detection of actions of people. The research community has strongly stated that the performance of support vector machine is good in action recognition [2]. SVM is a supervised machine learning algorithm and it can be used for the classification or regression purpose. By separating the distinct class with a maximum possible wide gap, SVM tries to predict the respective class, given a set of input data. The main purpose of Support Vector Machine is finding the hyperplane that in n-dimensional space which classifies the data points clearly, where n is number of features.

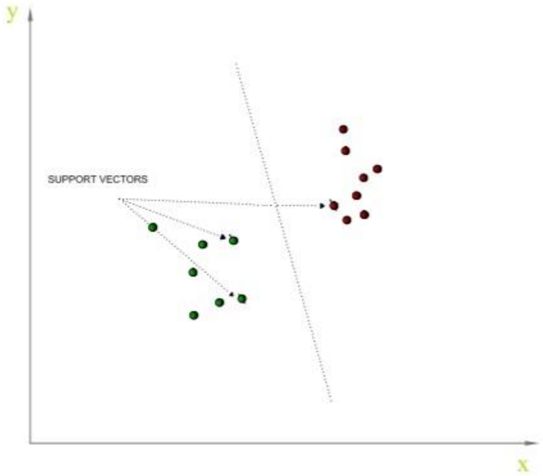

Fig.1. Support Vector Machine (https://www.analyticsvidhya.com/blog/2014/10/support-vector-machine-simplified/)[3]

II. HUMAN ACTION RECOGNITION BASED ON SUPPORT VECTOR MACHINES

Based on given training samples human action recognition system classifies a spatio-temporal feature descriptor in a video. Support Vector Machine (SVM) classifier solves the limitations of the large size of the feature vector and long training time. This paper is organized as follows. Section 2 describes various approached used for human action recognition based on support vector methods. Section 3 discusses advantage and disadvantages of those methods and finally Section 4 concludes this paper.

SVM is one of the best classifiers which have been regularly used in case of visual pattern recognition. [4] proposed a real time fast action recognition algorithm based on weighted hyper sphere support vector machine for recognizing different actions which can used for public safety in nontraditional manner and which can help us to find abnormal and suspicious actions, accidents and gives a warning in prior. [5] proposed a motion-based action recognition system with the help of motion-based features and SVM using 43 optical markers to detect 21 different kind of human actions and overall prediction rate is increased to 84%. [6] proposed a method which views action sequence not with time axis but with the space axis. Multi-class SVM is used as a classifier for classification of different actions with one against one approach. In this research work, LIBSVM is used for action classes but the action recognition rate is not that good as other best algorithms and it is affected because of image quality and filters used in MFCC. LIBSVM is machine learning library written in C++ which executes SMO algorithm and support for regression and classification.

The problem with the recent action recognition methods is that they are designed with limited view variation. Based on the available view information for recognition of human action, advanced methods are roughly grouped into two classes, viz., single view and multiview methods. Zeyu Liu [7] proposed a robust, fast action recognition method for sports video based on single view and multi-class linear SVM with shape, motion and pose features. The performance was boosted by using this poselet seed selection method. Detecting actions were hard in multi-view point as compared to single view point because when observing a static object, the actions may look different from different viewpoint angles cause fails in prediction. Appearance of action is affected by moving camera. Geometry based methods require the identification of body joints and corresponding point estimation in video. On contrast to this, Imran N. Junejo, [8] proposed a method which does not assume samples of multi view actions for testing and training. [9] also proposed very systematic and effective method for human action recognition which is independent of viewpoints. The method proposed by Yeyin Zhang [10] has a high stability over the seventeen viewpoints.

The computational time in processing the video frames is one of the main problem with the action recognition. The frames should be resized to the possible smallest sized frames because the goal is to detect humans in real-time. This problem was addressed by [10] which used the time weighted variance feature and temporal information in the feature in discriminating confusing actions like sit-down and stand-up and performed well on KTH dataset and on real-time videos gaining the accuracy of 70-80% [11]. Human action recognition is gaining its popularity due its wide applicability in automatic retrieval of videos of particular action using visual features [12]. [13] presented a new form of image object proposals, Temporally Enhanced Image Object Proposals (TE-IOPs), for the detection of the online video object/action. The proposed TE-IOPs augment the existing IOPs at every frame by their temporal dynamics in the past few frames. An adaptation method was proposed by [14] for enhancing the recognition of action videos by extracting knowledge from images. That knowledge is used to learn the correlated action semantics. A video Darvin method was provided by [15] for capturing the video-wide temporal information for action recognition. A novel method was presented by [9] for human action detection in crowded videos. The moving parts of each action are considered for training each action model and for copying the occlusion problem. But [13] method cannot handle data with fast or slow motion but solution is provided by [16] which can handle data with great changes of motion.

An object consists of several parts and each of them can be represented by a skeleton. But it remains to be a challenge when detecting the skeletons of individual objects in an image as it requires an effective part detector and a part merging algorithm to group parts into objects. [17] presented a new fully unsupervised learning framework for detecting the skeletons of human objects in a RGB-D video. [18] proposed a method for combining both depth map-based features and skeleton based features to improve human action recognition, which increases the robustness of the body pose estimation errors and takes the advantage of the additional discriminative data that the skeleton can provide. [19] proposed an approach for skeletal feature extraction and splitting, similarity measure for action classification and recognition for the cases of low-quality video, motion discontinuities and motion aliasing problems.

The local features-based fails to capture adequate spatial or temporal relationships. A trajectory-based local representation approaches have been proposed to overcome this problem [20]. [21] adopted trajectory-based representation that contain more discriminative information on motion and action in videos as trajectories are efficient in capturing object motions and actions in videos. But the problem with this approach is that computing descriptors of dense trajectories may spend lots of time and many trajectories which belong to the background trajectories may not be useful for the recognition.[16] proposed a trajectories-based motion neighborhood feature (TMNF) method for action recognition.

The Weizmann database consists of 90 low-resolution (180 × 144) videos; 10 types of human actions (Walk, Run, Jump, Gallop sideways, Bend, One-hand wave, Two-hands wave, Jump in place, Jumping Jack, Skip) performed several times by 9 subjects. This dataset uses a fixed camera setting and a simple background. [9] investigated the recognition accuracy for the human actions of the well-known Weizmann benchmark database and got accuracy of 100%. [22] provided action classification results on five different data sets in Weizmann database with the accuracy of 100%. [23] proposed a novel method using Boosted Exemplar Learning obtaining 100% accuracy.

III. RESULTS AND DISCUSSION

The table I shows the comparative analysis on human action recognition based on support vector machines. Temporally enhanced image object proposals (TE-IOPS) finds the detection of human action and online object. But it is not an optimal solution when it comes to extreme cases such as video contain fast or slow motion of action. Skeleton + trajectory can detect easily [1]. Trajectories based motion neighborhood can detect accurately with great changes of motion[2]. Image-to-video adaptation provides improved performance of recognition of human action in videos. Support vector machine describes a video in real time for blind people [4]. Hyper-sphere support vector machine is advantageous in providing efficient and scientific solutions [5]. Optical flow and SVM classifier provide accurate recognition result in case of controlled environmental background this method [6]. Depth map-based features (hon4d) and skeleton-based features (Fourier temporal pyramid) provides fusion of multiple features with the improved performance. Video Darwin is very easy to implement, effective to interpret and fast to compute in recognizing a broad variety of actions [7]. Depth motion map (DMM) provides better results in situation where by using two modality sensors together as compared with individual sensor [8]. Activity net method is a new large-scale benchmark providing huge varieties in terms of taxonomy, diversity activities [9]. Single-view key pose classification is a classification method takes more time in terms of computation time with a large clip length [24]. Poselets is not good when it comes to walking, basketball shooting [10]. Cumulative motion shapes (CMS), multi-training in k-nearest neighbor (k-NN) and multiclass support vector machines (SVM) having the advantage the effective interest point selection by using of nearest point bounding box [11].The time weighted variance feature embeds temporal information in the feature and helps in discriminating confusing actions such as sit-down and stand-up [12].Key-frame extraction and video skimming extracts key-frames and then skims the video clip by concatenating excerpts around the selected key-frames[13].Wavelet domain features use two-dimensional discrete wavelet transform offers improved space-frequency localization. It is helpful for analyzing images or video frames, where the information is localized in space [14]. The Fisher Kernel method describes how much the GMM parameters are modified to best fit the video trajectories [15]. ASIFT captures both action and identity information for human action recognition [9]. Affine moment invariants are derived from the 3D spatio-temporal action volume and the average image created from the 3D volume, and classified by an SVM classifier [16]. Context-constrained linear coding (CLC) not only considers spatio-temporal contextual information but also alleviates quantization error. The interest points are first detected in each action video and then the total local appearances are clustered into a codebook [17].Self- similarity matrix being possibly defined over a variety of image features, either static or dynamic, these descriptors can take different form and can be combined for increased descriptive power [18].Pyramid Histograms of Orientation Gradients (PHOG) of all invariant shapes in video are averaged to form a feature vector that captures the characteristic of human actions in this video sequence. Using Support Vector Machine (SVM), the method is tested on the KTH action dataset [19]. X-T slice based method is utilized to describe an action and MFCC is used to extract the feature. It treated an action in global level and found advantages over image-basedmethods [20]. Segmented skeletal features perform human action recognition system automatic skeletal feature extraction and splitting by measuring the similarity in the space of diffusion tensor fields, and multiple kernels Support Vector Machine based human action recognition [21]. Action detection based on masks is to handle the partial occlusion problem, only the moving body parts in each motion are involved in action training [16]. Computing invariants across frames made generalizations to cross ratio of four collinear points so that it could be applied to view-invariant representation of actions [25]. Semilatentdirichlet allocation (S-LDA) and semilatent correlated topic model (SCTM) obtains from largescale motion descriptors from a whole frame, rather than small space-time patches [22]. Boosted Exemplar Learning (BEL) approach recognize various actions in a weakly supervised manner [23].

Table I Comparative analysis on human action recognition based on support vector machines

Author	Advantage	Disadvantage	Accuracy
Jiong Yanga et al [13]	Motion information will be advantage in finding detection of human action	When it comes to extreme cases such as video contain fast or slow motion of action, this method is not an optimal solution	67%
Cheng, S., et al [17]	Image skeletons of discrete objects can detect easily	Need to increase the sizes of the training datasets	80%
Xiang Xiao et al [16]	Action can detect accurately with great changes of motion	It is not detecting interested region automatically	91.4%
Zhang, J. et al [14]	To improve the performance of recognition of human action in videos, this method will completely use the unlabeled videos heterogeneous features	When more labeled training videos are available performance of IVA is not good	60%
Shuang Liu, et al [4]	In the case of non-traditional public safety problems this method will provide efficient and scientific solutions	It cannot detect new abnormal action in videos	–
Jagadeesh B.et al., [24]	In case of controlled environmental background this method provides accurate recognition result	In real time action recognition invideo scenario the recognition accuracy is reducing greatly	90%
Li, K. e al.,[18]	Fusion of multiple features improved performance	Regarding the confusion matrix few action classes have worst recognition	94.14%
Fernando, B., et al [15]	Very easy to implement, effective to interpret, fast to compute in recognizing a broad variety of actions	From input video, frames are removing almost upto 20%	75%
Liu, Z. et al., [7]	Fast and robust using a quick feature extraction process	Taking more time in terms of computation time with a large clip length	92.3%
J. Wang and H. Lee et al [5]	Embedding temporal information and time weighted variance will help	The features along separate dimensions of x, y, z will cause drastic fall in the recognition	95%

		accuracy	
Azouji, N., et al[11]	Reduces run time drastically because input dimension is reduced and a smaller number of video frames are required	Recognition accuracy is not improved	88.00%
H. Imtiaz. Et al., [25]	Use of dimensional feature space is very less	When complex backgrounds and nonstable object scenario generally it is very difficult to conclude the features of foreground	100%
Atmosukarto, I., et al.[21]	The selected discriminative trajectories can be used to assist as instance prototypes in multiple instance learning frameworks	the Fisher Vector becomes impractical for large scale applications due to storage limitations	90%

Table I Comparative analysis on human action recognition based on support vector machines (Contd.)

Author	Advantage	Disadvantage	Accuracy
Junejo, I. N., et al[8]	Stable in view point variation scenario	Creates a mild guess of localization of person	65.%
Shan, Y., et al [6]	Efficiently utilize the body structure and also moving velocity	Make using of very simple mean and variance features, those are not sufficient to identify the variations in X-T sequences	92.99%
Yoon, S. M. et al., [19]	Very effective and efficient in recognition of actions	Cannot extract the robust features	-
Ping GUO [9]	For detection of human action in crowded videos using masks and segmentation is required for it	Detects action only when the body parts are moving or in motion	79.0%
Yeyin Zhang.et al.,[10]	High robustness to sampling intervals and changing viewpoints	Recognition rate is lesser than other methods	92.8%
Yang Wang and Greg Mori [22]	Achieves a great performance by using class labels information provided in training set	This method need support human figures, tracking in pre-processing stage	90%
Tianzhu Zhang, et al [23]	Recognize variety of actions in weakly supervised way	It is very hard to distinguish different kind of actions	94.33%

CONCLUSION

Support vector machine is one of the powerful and robust algorithms in machine learning and it is commonly used for problems like classification, regression and pattern recognition. By separating the distinct class with a maximum possible wide gap, SVM tries to predict the respective class given a set of input data. Because of its higher prediction capabilities SVM has gained increasing attention in the remote sensing, image processing, robotics, pattern recognition and many others community. When it comes to theoretical, SVM can produce robust, accurate classification results. In this survey paper, we provided a detailed analysis on human action recognition using Support Vector Machine algorithm and this paper can be used as guidance for fellow researchers who have started doing research in field of action recognition using the Support Vector Machine.

REFERENCES

[1]. S. R. Rashmi, S. Bhat and V. C. Sushmitha, (2017, August). Evaluation of human action recognition techniques intended for video analytics, In International Conference On Smart Technologies for Smart Nation (SmartTechCon), Bangalore, 2017.

[2]. Ayumi, Vina & Fanany, Mohamad Ivan. (2015, September). A Comparison of SVM and RVM for Human Action Recognition. In International Conference on Industrial Internet of Things, At Samosir Island, North Sumatra, Indonesia ,2015.

[3]. https://www.analyticsvidhya.com/blog/2014/10/support-vector-machine-simplified/

[4]. Liu, S., Chen, P., & Cui, X. (2017, July). Action recognition in videos based on weighted hyper-sphere support vector machine. In *Proceedings of International Conference on Machine Learning and Cybernetics*, *ICMLC 2017*.

[5]. J. Wang and H. Lee, (2009, May). Recognition of Human Actions Using Motion Capture Data and Support Vector Machine. In *WRI World Congress on Software Engineering*, Xiamen, 2009.

[6]. Shan, Y., Wang, S., Zhang, Z., & Huang, K. (2011, November). An X-T Slice Based Method for Action Recognition. *Signals*.In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops),2017.

[7]. Liu, Z., Miao, Z., &Huo, (2015, November). A Realtime Human Action Recognition Method Based on Single View Key Poses in Sports Video, 210–213.

[8]. Junejo, I. N., Dexter, E., Laptev, I., & Pe, P. (2011). View-Independent Action Recognition from Temporal Self-Similarities, *33*(1), 172–185.

[9]. Guo, P., & Miao, Z. (2010, August). Action detection in crowded videos using masks. *Proceedings. In International Conference on Pattern Recognition*, 2010.

[10]. Yeyin Zhang, Kaiqi Huang, Yongzhen Huang and Tieniu (2009, November) View-Invariant Action Recognition Using Cross Ratios Across Frames, Tan National Laboratory of Pattern Recognition in International Conference on Image Processing (2009).

[11]. Azouji, N., &Azimifar, Z. (2013, December). *A new Approach to Speed up in Action Recognition* Based on Key-frame Extraction. In 8th Iranian Conference on Machine Vision and Image Processing (MVIP) ,2013.

[12]. C. J. Dhamsania and T. V. Ratanpara, (2016, November) "A survey on Human action recognition from videos. In *Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, 2016.

[13]. Yang, J., & Yuan, J. (2018). Temporally enhanced image object proposals for online video object and action detections. *Journal of Visual Communication and Image Representation*, *53*(September 2017), 245–256.

[14]. Zhang, J., Han, Y., Tang, J., Hu, Q., & Jiang, J. (2017). Semi-Supervised Image-to-Video Adaptation for Video Action Recognition. *IEEE Transactions on Cybernetics*, *47*(4), 960–973.

[15]. Fernando, B., Gavves, E., José Oramas, M., Ghodrati, A., &Tuytelaars, T. (2015, June). Modeling video evolution for action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2015.*

[16]. Xiao, X., Hu, H., Wang, (2017, September). Trajectories Based Motion Neighborhood Feature for Human Action Recognition.In 2017 IEEE International Conference on Image Processing (ICIP) ,2017.

[17]. Cheng, S., & Hsiao, K. (2018). A novel unsupervised 3D skeleton detection in RGB-D images for video surveillance. Multimedia Tools and Applications. 1-29. 10.1007/s11042-018-6292-y.

[18]. Li, K., Liu, Z., Liqin, L., &Yanan, S. (2016, October). Human action recognition using associated depth and skeleton information. In *2nd IEEE International Conference on Computer and Communications, ICCC 2016*

[19]. Yoon, S. M., &Kuijper, A. (2010). Human Action Recognition using Segmented Skeletal Features Sang Min Yoon. An International Journal. 40. 6848-6855. 10.10

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

[21]. Atmosukarto, I., & Ghanem, B. (2012, December). Trajectory-based Fisher Kernel Representation for Action Recognition in Videos. In ICPR 2012 - 21st International Conference on Pattern Recognition (ICPR),1990.

[22]. Wang, Y. (2009). Human action recognition by semi-latent topic models. *IEEE Transactions on Patttern Analysis and Machine Intelligence*, *31*(10), 1762–1774.

[23]. Tianzhu Zhang, Liu, J., Si Liu, Yi Ouyang, &Hanqing Lu. (2009 October). Boosted Exemplar Learning for human action recognition. *In IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*,2009.

[24]. Jagadeesh, B., & Patil, C. M. (2016, May). Video based action detection and recognition human using optical flow and SVM classifier. *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* ,2016.

[25]. H. Imtiaz, U. Mahbub, G. Schaefer, and M. A. R. Ahad (2013, November), "A Multi-resolution Action Recognition Algorithm Using Wavelet Domain Features," In 2nd IAPR Asian Conference on Pattern Recognition, 2013.

[26]. https://machinelearningmastery.com/how-to-load-and-explore-a-standard-human-activity-recognition-problem/.

[27]. https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf