Classification of Action Based Video using Heterogeneous Feature Extraction and SVM

Chandrawal kaur, Amit Doegar

Abstract*: Action recognition (AR) plays a fundamental role in computer vision and video analysis. We are witnessing an astronomical increase of video data on the web and it is difficult to recognize the action in video due to different view point of camera. For AR in video sequence, it depends upon appearance in frame and optical flow in frames of video. In video spatial and temporal components of video frames features play integral role for better classification of action in videos. In the proposed system, RGB frames and optical flow frames are used for AR with the help of Convolutional Neural Network (CNN) pre-trained model Alex-Net extract features from fc7 layer. Support vector machine (SVM) classifier is used for the classification of AR in videos. For classification purpose, HMDB51 dataset have been used which includes 51 Classes of human action. The dataset is divided into 51 action categories. Using SVM classifier, extracted features are used for classification and achieved best result 95.6% accuracy as compared to other techniques of the state-of- art.

***Index Terms*: Action Recognition, Classification, Convolution neural network, SVM.**

I. INTRODUCTION

Technology advancement has made researcher to work on video processing as it has gained greater attention. Action [1] Recognition is considered as a tough task because of large inter or intra class variation in video frames. This difficulty occurred because of texture of object, different illumination to the camera viewpoint and cluttered scene of backgrounds. Video processing is being done for different reasons and wide variety of applications e.g., robot tracking, video surveillance and human behavior detection. Modern evolutions in the Depth video technology have helped in minimizing the issues in the videos. The issues were created due to the texture and illumination changes and were nullified with the help of depth video technology because the depth information is unconcerned to such diversities. Moreover, the view point variations of camera are very much prone to affect the depth videos. We propose to heterogeneous features deep characteristics from RGB frames and Depth information from optical flow between frames and that resulting heterogeneous characteristics for action recognition to integrate to incorporate robustness in natural action videos against viewpoint differences. Traditionally, hand-crated features can be used for recognition [2], [3] of action. Recent decades, deep learning features acquired through convolutional neural networks (CNNs) have shown greater robustness and discriminative strength.

Revised Manuscript Received on September 05, 2019.

Chandrawal kaur, Computer Science and Engineering Department, NITTTR Chandigarh, India

Amit Doegar, Computer Science and Engineering Department, NITTTR Chandigarh, India

Greater success has been achieved in computer vision tasks like object tracking, image segmentation, and action recognition and so on with the help of famous deep learning models of CNNs (Convolutional Neural Networks) [4]. There are enormous types of advance algorithm and classifiers are used to achieve higher accuracy on different dataset of action recognition.

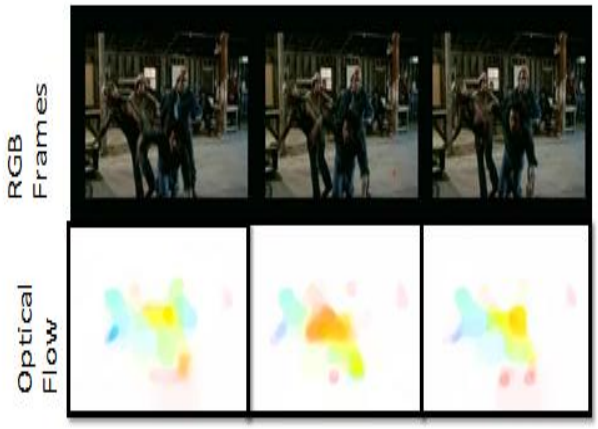


Fig. 1 The proposed approach work on RGB Frames for appearance and optical flow for motion on HMD51 dataset.

Usually, Realistic and non- Realistic actions are formed in videos. When actor performs some action scene with simple background such action is Non- realistic action, whereas action performed randomly by people with no simple background is realistic action. Our work is based on the analysis of realistic dataset and extracted frames characteristics are analyzed for AR and extract characteristics from RGB frames and optical flow frames using pre-trained Alex-Net network. After that create the database of heterogeneous features of RGB frames and motion features of frames for better classification. Apply Principle Component analysis on dataset to reduce the dimensions of data and classify data using Fine Gaussian Support Vector Machine achieved better accuracy as compared to other methods. The experimental outcome demonstrates that the Support Vector Machine classifier achieves more accuracy. The remainder of this research paper is explained further. Earlier work is being talked about in Section II. Section III method suggested the classification of action based videos. We present the results of a comprehensive set of data set experiments in Section IV. Finally, this paper is concluded by Section V.

II. EARLIER WORK

Over the past two decades, we will investigate some associated works on action recognition in videos. Representation of features and classification pattern is widely used in action recognition task. Previous action identification algorithm is approximately divided into two components: (1) Deep methods of learning, (2) Traditional techniques.

Handcrafted features are mostly traditional methods design to model the spatial-temporal structure and use the extracted features from frames to train an action classifier. For example, extracted features using Histogram of Optical Flow (HOF) [5] and used to train a classifier such as SVM [5]. Different types of techniques are used in action recognition iDTs [6], SIFT3D [7], ESURF [8], and HOF [9] are used to represent motion and appearance effectively across frames in videos. Several other techniques were proposed to model the temporal structure in an efficient way, such as the temporal action decomposition [10], ranking machines [11], and dynamic poselets [12]. Recently, due to the development of potent GPUs and huge action datasets, deep learning has been implemented broadly on the action recognition videos.

Several, methods have been proposed for action classification in deep learning. Classification of action videos implemented using various methods like SVM so on. In videos sequence of frames is quite different as compared to static images. By using Spatial-Temporal Convnet A. Karpathy *et al.* [13] discusses multiple strategies like late fusion, early fusion, and temporal fusion to extending the frame connectivity in temporal domain. Dong Li *et al.* [14] enhanced the accuracy of UCF101 dataset by applying temporal attention probability for each video segment in temporal sequence.

In videos, extraction of features from frames is very important to recognize the action. Many types of filters are applied on frames to features. Some methods for feature extraction from frames like Gabor, HoG, and CNN these all extract on the basis of texture, shape, color [15]. CNN pre trained Alex-Net network have many layers which are used for features extraction for recognition action in videos [16].

Currently, the most efficient type of learning machines and approaches of deep learning are used in recognition of action in videos. The Conv+LSTM method utilized by J. Donahue *et al.* showed 63.2% accuracy in classifying for activity recognition, video description, and image description [17]. The TDD method used to extract convolutional feature maps achieved 90.3% UCF101 dataset precision and HMDB51 dataset precision of 63.2% [18]. The 3D ResNet 101 method used by Kensho Hara *et al.* for recognition action in videos obtained 88.9% UCF101 dataset precision and HMDB51 dataset precision of 61.7% [19]. The Deep networks with Temporal Pyramid Pooling (DTPP) approach used by Jiagang Zhu *et al.* for recognition of action achieved 74.8% accuracy [20]. Temporal segment network (TSN) approach help in good practices in learning ConvNets on video data with an accuracy of 69.4% [21]. An approach to detect action performs in a video by ActionVLAD method was used by Rohit Girdhar *et al.* with an accuracy of 66.9% on HMDB51 dataset [22]. The Deep networks with Temporal Pyramid Pooling (DTPP) perform better than temporal segment network in case of videos sequence patterns [20]. Jue Wang *et al.* used Support Vector Machine Pooled (SVMP) descriptor for action classification obtained 81.3% accuracy on HMDB51 dataset [23]. VGG-16, fc6 layer functions used by pooling of kernelized block-diagonal correlation and ResNet-152 use features of pool5 and achieved 71.3% accuracy [24] for classification of action videos.

Table 1: Summary of some commonly used action datasets

Ref. no.	Methods	Dataset
[19]	3D ResNet 101	UCF101 HMDB51
[18]	Deep-convolutional Descriptor technique combined with trajectory	UCF101 HMDB51
[22]	Locally aggregated descriptors action vectors and Fisher Vector	HMDB51 UCF101 Charades
[24]	VGG-16 and ResNet-152	HMDB51 MPII JHMDB
[25]	CNN and Deep bidirectional LSTM	HMDB51 UCF101 YouTube action

Reference [25] described a method by using CNN and Deep bidirectional LSTM methods to process video frames and extracted features from Alex-Net network's fc8 layer and obtained 87.64 percent precision on the HMDB51 dataset. Hanling Zang *et al.* [26] improved the accuracy of action identification dataset by extracting the visual appearance and motion features of the video frames are extracted using the VGGNet network and classified using support vector machine classifier.

III. PROPOSED APPROACH

A. DATA COLLECTION

HMBD51 is a dataset based on Human Motion consisting of 6766 realistic videos with 51 classes of behavior. Datasets are more difficult owing to big differences in camera movement, appearance of objects and changes in the actors position, scale and point of view, as well as cluttered background. HMDB51 dataset collected from various sources like youtube, short movies clip, google videos etc. The dataset HMDB51 includes a number of general body motions, facial actions, and human body movements with human interaction and interacting with object. It is more complex by collecting the videos from each class for a multitude of subjects with separate illuminations. We proposed work on visual appearance and optical flow in frames to recognizing the motion in videos for better classification of action videos.

B. FLOW CHART

The proposed methodology is explained by flow chart in Figure 2.

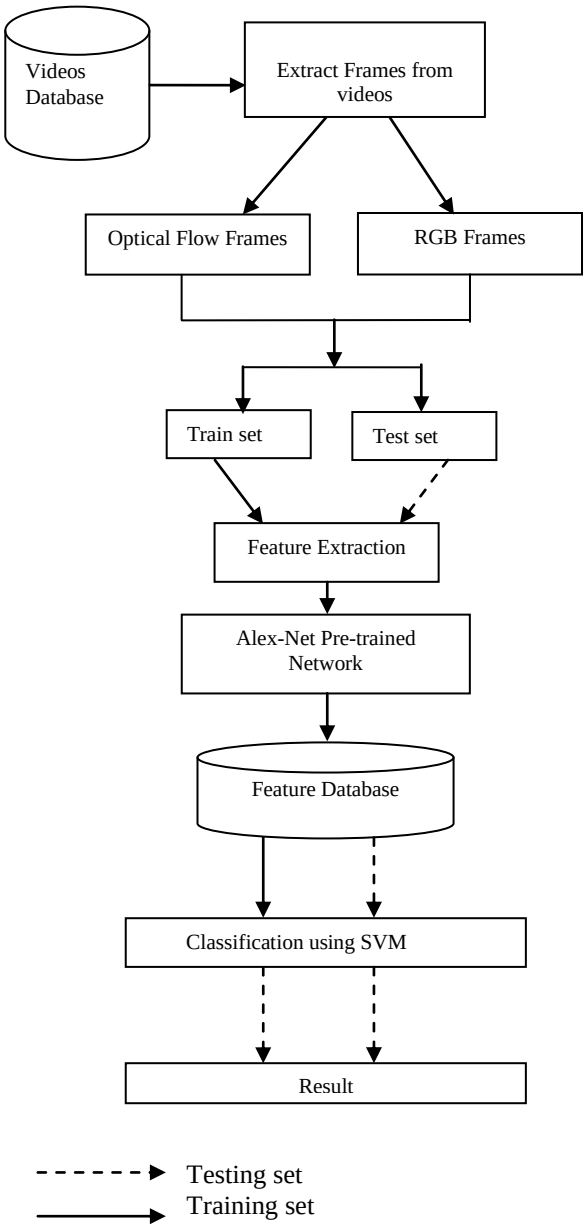


Fig. 2 Flow chart of proposed approach

C. IMPLEMENTATION

Firstly, extract the 1-10 frames from videos. Spatial and Temporal which include appearance and motion in frames of video. For visual appearance in frames RGB frames are used and optical flow frames are used for detect the motion between frames. The technique of optical flow is used to calculate the movement between two frames that are taken at times t and $t + \Delta t$. We use optical flow algorithms based on Horn – Schunck to detect the flow between frames because this method introduces a global constraint of smoothness. Flow is formulated as a global energy function. This function is given for two dimensional images.

$$E = \iint [\,(I_x u + I_y v + I_t)^2 + \alpha^2(||\nabla u||^2 + ||\nabla v||^2)]dxdy \quad (1)$$

Where the derivatives of image likeness strength evaluation along the x, y and the dimensions are I_x, I_y, and I_t, α is a constant of regularization.

$$Optical\ flow\ vector = [u(x,y), v(x,y)]^T \quad (2)$$

Flow of optics is a main component of the strategy to video classification because it encodes the pattern of obvious objects movement in a visual scene. Optical flow is very helpful in analyzing the movement in two or more frames in order to acknowledge the action in the video series of temporal frames.

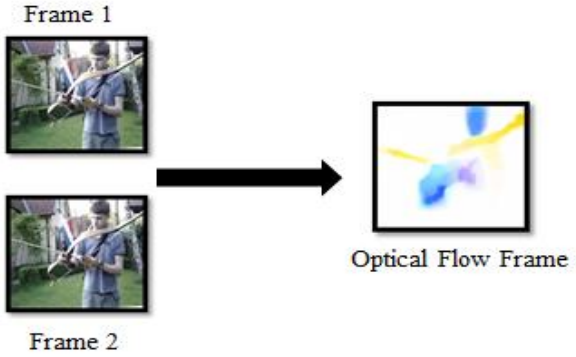


Fig. 3 Representation of Optical Flow between two frames

D. DEEP FEATURES EXTRACTION

Extraction of features includes extracting a greater amount of data from raw pixel values which can capture the difference between the categories concerned. This extraction of the feature is done in an unsupervised manner in which the image classes have nothing to do with pixel extracted information. Convolutional neural network play potent role in features extraction and it is unique types of networks intended for multi-layer neural designed to identify visual patterns with minimal pre-processing and directly from pixel images. We choose Alex-Net for features extraction. Alex-Net is used to extract features of RGB frames for appearance information and optical flow frames capture the motion information between frames. Alex-Net is a convolutional neural network that is pre-trained model. Alex-Net consists of 5 convolutional layers, 3 max layers for pooling and fc6, fc7, fc8 are layers completely linked which help to extract dense features of frames. Alex-Net is a helpful network for image categorization and object recognition because kernel size (11×11), pooling window (3×3) and profound structure layers are used. We extract 4096 dimensions features of RGB frames from fc7 layer and also extract features of optical flow frames with 4096 dimension at fc7 layer. Pre-trained model architecture displayed in figure 3.

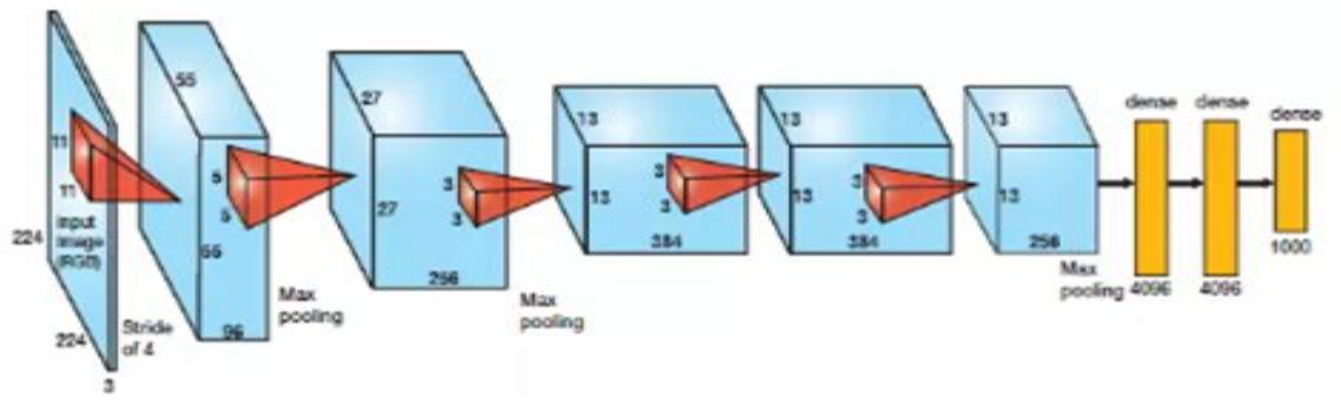


Fig. 3 Alex-Net pre-trained model architecture

After extracting the features from optical flow frames and RGB frames than concatenate these features. We get 8192 dimension features and created database of these features and apply principal component analysis on database to reduce the dimensions of features. Principal component analysis creates linear combinations of the features and compresses the dimensions for fast computation of data. Principal component analysis is unsupervised algorithm and identified the statistical pattern in the data.

E. CLASSIFIER

Classification is the process to distinguish the different classes of action dataset. In our proposed work we use SVM, based upon supervised learning technique. SVM is constructing a hyper-plane between the classes to classify them appropriately with maximum margin. We have used SVM with Gaussian kernel type and it gave best results as compared to other state of art methods with 95.6% accuracy.

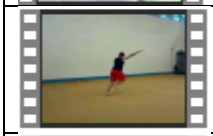
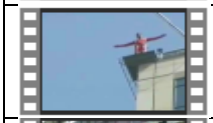
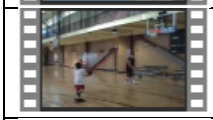
IV. EXPERIMENTAL RESULTS AND DISCUSSION

For experiment implementation MATLAB has been used and HMDB51 dataset is used for action classification. In our research work dataset contain 51 action categories of videos. The proposed work performance is evaluated on the basis of accuracy. Accuracy is defined as how often is the classifier correct and the matrix of confusion is a table which describe the implementation of a classifier model.

$$Accuracy = \frac{Tp + Tn}{Tp + Fp + Tn + Fn} \quad (3)$$

Where Tn, Fp, Tp, Fn is true negative, false positive, false negative, true positive respectively. Basically, confusion matrix of dataset is shown in figure 5. The vertical axes represent the true class and the horizontal line represents the predicted class. Table 2 shows some of the right and miss-classified visual outcomes. Our technique uses a test video as input and extracts 1-10 frames of video after that extracts features from RGB frame and optical flow frames. In Table 2, row 2 and row 4 are miss-classified, where ‘‘cartwheel’’ is classified as ‘‘somersault’’ and ‘‘shoot ball’’ is classified as ‘‘kickball’’. These inaccurate projections are attributed to the resemblance of visual material, camera movement, and shifts in actor body components in both action classifications.

Table 2: The suggested guess approach for AR for specimen clips. The red font shows the incorrect projection of our technique.

HMDB51 dataset videos	Predictions	Ground Truth
	**Brush hair**	**Brush hair**
	**Somersault**	**Cartwheel**
	**Dive**	**Dive**
	**kickball**	**Shoot ball**
	**Sword**	**Sword**

The action recognition accuracy of the suggested approach and other techniques are shown in Table 3. The suggested approach increased the accuracy on this dataset from 87.64% to 95.6% with an increase of 7.96%.

Table 3: Comparison with other state-of-art methods

Approach	Accuracy
Tingzhao Yu *et al.* [27]	60.2%
Hao Yang *et al.* [28]	65.4%
Yu Qian *et al.* [29]	75.7%
Zhigang Tu *et al.* [30]	80.9%
Amin Ullah *et al.* [25]	87.64%
Proposed	**95.6%**

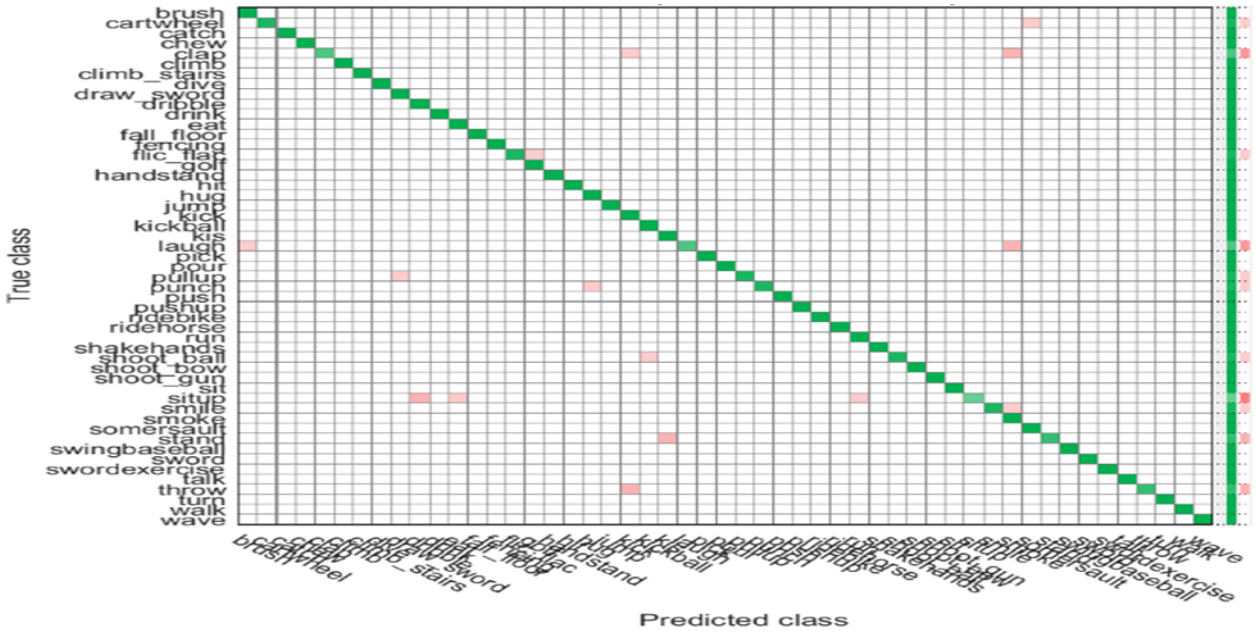


Fig. 5 Confusion Matrix of HMDB51 dataset for the proposed approach of AR

V. CONCLUSION

We have created a successful technique for classification of action video features for AR. The research involves RGB frames and optical flow frames are used for AR and feature extraction using Convolutional Neural Networks (CNNs) pre-trained model Alex-Net and classifying data using SVM classifier and achieved 95.6% accuracy. Experimental findings show that other state-of-art AR methods on HMDB51 action video datasets are effectively dominated by the accuracy of the suggested technique. Visual information of frames and motion data of video frames, these characteristics make our suggested model more suitable for AR. In future, can be work to extract only salient features of video frames.

REFERENCES

1. Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees G. M. Snoek, ''VideoLSTM Convolves, Attends and Flows for Action Recognition'', *Computer Vision and Image Understanding*, Vol. 166, Issue C, pp. 41-50, January 2018.
2. Zheheng Jiang, Danny Crookes, Brian D. Green, Yunfeng Zhao, Haiping Ma, Ling Li, Shengping Zhang, Dacheng Tao, and Huiyu Zhou, ''Context-Aware Mouse Behavior Recognition Using Hidden Markov Models'', *IEEE Transactions on Image Processing*, Vol. 28, no. 3, pp. 1133 – 1148, March 2019.
3. Himanshu S. Bhatt, Richa Singh, and Mayank Vatsa, ''On Recognizing Faces in Videos Using Clustering-Based Re-Ranking and Fusion'', *IEEE Transactions on Information Forensics and Security*, Vol. 9, no. 7, pp. 1056 – 1068, July 2014.
4. B. Li, R. Chellappa, Q. Zheng, and S. Der, "Model-based temporal object verification using video", *IEEE Trans. Image Processing*, vol.10, no.6, pp.897-908, 2001.
5. N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection", *IEEE Conference on Computer Vision Pattern Recognition*, pp. 886-893, 2013.
6. H. Wang and C. Schmid, "Action recognition with improved trajectories", *IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, pp. 1550-5499, Dec. 2013.
7. P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition" *In ACMMM*, Augsburg, Germany, pp. 357-360, Sept. 2007.
8. G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector", *In ECCV*, Berlin, Heidelberg, pp. 650 – 663, 2008.
9. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies" *In CVPR*, Anchorage, AK, USA, pp. 1063-6919, June 2008.
10. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies" *In CVPR*, Anchorage, AK, USA, pp. 1063-6919, June 2008.
11. J. C. Niebles, C.-W. Chen, and L. Fei-Fei, " Modeling temporal structure of decomposable motion segments for activity classification", *In ECCV*, Berlin, Heidelberg, pp 392-405, 2010.
12. B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, " Modeling video evolution for action recognition" *In CVPR*, Boston, MA, USA, pp. 1063-6919, June 2015.
13. L. Wang, Y. Qiao, and X. Tang, " Video action detection with relational dynamic-poselets", *In ECCV*, Cham, pp. 565-580, 2014.
14. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Feifei, "Large-scale video classification with convolutional neural networks*," IEEE International Conference on Computer Vision and Pattern Recognition. IEEE*, pp. 1725–1732, 2014.
15. Dong Li, Ting Yao, Ling- Yu Duan, Tao Mei, and Yong Rui, "Unified Spatio-Temporal Attention Networks for Action Recognition in Videos", *IEEE Transactions on Multimedia*, Vol. 21, no. 2, pp. 416 – 428, Feb. 2019.
16. Muhammad Sharif, Muhammad Attique Khan, Farooq Zahid, Jamal Hussian Shah, and Tallha Akram, "Human action recognition: a framework of statistical weighted segmentation and rank correlation based selection", *Pattern Analysis and Applications*, Springer London, pp. 1-14, Feburary 2019.
17. Xintong Han, Bharat Singh, Vlad I. Morariu, and Larry S. Davis, ''VRFP: On-the-Fly Video Retrieval Using Web Images and Fast Fisher Vector Products'', *IEEE Transactions on Multimedia*, Vol. 19, no. 7, pp. 1583 – 1595, July 2017.
18. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description" *In CVPR*, Boston, MA, USA, pp. 1063-6919, June 2015.
19. L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors", *In CVPR*, Boston, MA, USA, pp. 1063-6919, June 2015.
20. K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition" *In ICCV*, Venice, Italy, pp. 2473-9944, Oct. 2017.
21. Jiagang Zhu, Zheng Zhu, and Wei Zou, " End -to- end Video level Representation Learning for Action Recognition", *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, pp. 1051-4651, Aug. 2018.

22. L.Wang, Y.Xiong, Z.Wang, Y.Qiao, D.Lin, X.Tang, and L.V.Gool, " Temporal segment networks: towards good practices for deep action recognition", *In ECCV,* , Springer, Cham, pp 20-36, September 2016.
23. Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell, "ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 971-980, 2017.
24. Jue Wang, Anoop Cherian, Fatih Porikli, and Stephen Gould, "Video Representation Learning Using Discriminative Pooling", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1149-1158, 2018.
25. Anoop Cherian and Stephen Gould, " Second-order Temporal Pooling for Action Recognition", *International Journal of Computer Vision*, Vol. 127, no. 4, pp. 340–362, April 2019.
26. Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features", *IEEE Access*, vol.6, pp. 1155-1166, 28 November 2017.
27. Hanling Zhang, Chenxing Xia, and Xiuju Gao, "Action recognition based on multi-stage jointly training convolutional network", *Multimedia Tools and Applications*, vol.78, pp. 9919-9931, April 2019.
28. Tingzhao Yu, Lingfeng Wang, Cheng Da, Huxiang Gu, Shiming Xiang, and Chunhong Pan, "Weakly Semantic Guided Action Recognition", *IEEE Transactions on Multimedia*, pp. 1-1, March 2019.
29. Hao Yang, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, and Stephen J. Maybank, "Asymmetric 3D Convolutional Neural Networks for Action Recognition", *Pattern Recognition*, no. 85, pp. 1-12, January 2019.
30. Yu Qian, and Biswa Sengupta, "Pillar Networks: Combining parametric with non-parametric methods for action recognition", *Robotics and Autonomous Systems*, vol.118, pp. 47-54, August 2019.
31. Zhigang Tu, Hongyan Li, Dejun Zhang, Justin Dauwels, Baoxin Li, and Junsong Yuan, "Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition", *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2799 – 2812, June 2019.