Study of Various Techniques Used for Video Retrieval

DIGVIJAY PANDEY,
RESEARCH SCHOLAR, IET LUCKNOW.

BINAY KUMAR PANDEY
ASSISTANT PROFESSOR, GBPUAT PANTNAGAR U.K.

SUBODH WARIYA
HOD AND PROFESSOR, IET LUCKNOW INDIA.

***Abstract*— In present time a number of image processing and neural network techniques are being utilized in the analysis of various frames of videos that corresponding to different-different human actions. This paper performs survey on various method of classification of all human action that are stored in dataset. Different types of algorithms and methods are used to retrieve action video from large dataset. In our study, it is found that CNN (Convolution Neural Networks) famous deep learning models has achieved great success in action retrieval like object tracking, image segmentation, action recognition and so on.**

1.INTRODUCTION

The coming flooded requirement of internet, multimedia, storage and retrieval of Big data creates the challenge for efficiently retrieving the relevant content. The advancement of the technology makes video processing greater attention in the field of research. Now a days Action recognition [1], [2], [3] has become a challenging task in the computer vision research. It is well known that, Video is a collection of frames which are moved at a fast speed, so that they appear to be actually moving and quality of video depends upon the color quality for each bitmap in the frame sequence. There are various famous video formats exits but in this paper used only AVI (audio video interleaved) format is studied. In simple words action is defined as the trouble in system. A physical motion and something that individual do, all are called action. Recognition of action means to capture the particular action in the video sequence and used widely in various applications such as security, education, human robot interaction, sports, video surveillance and so on. As a bulge of the videos are stored by the end-user. previously the manual system was deployed to obtained a particular video from the large database. But, it is most time consuming and hectic task..Intelligent content retrieval [4] system has been solved this issue to up to some level.As they quickly find the requested video from the database. This paper is targeted on the survey of classification of action based videos by extracting hybrid features of videos frames and trained on support vector machine to recognizing the action.

Action recognition many times creates problem in computer vision due to visual effects [5]. Large variations in the action caused by the high dimensions, cluttered backgrounds, viewpoint variations, and low quality of video data are among the main challenges for classify the action data for recognition .In videos sequence, some motion of body part while interacting with environment in humans is called action and is normally represented using number of frames arranged sequentially, that can be easily understood by analyzing multiple frames in sequence. One of the common characteristics of action based video retrieval is focused on specific action. But this paper explains method of classification of all human action that are stored in HMDB51 [6] dataset. Different types of algorithms and methods are used to retrieve action video from large dataset. In the recent years, CNN (Convolutional Neural Networks) [7], famous deep learning models has achieved great success in computer vision tasks like object tracking, image segmentation, action recognition and so on. There is an various types of advance algorithm and classifiers are used to achieve higher accuracy on different dataset of action recognition. Convolutional neural network [8],[9] mostly applied for action recognition.

The rest of this paper is prepared as follows. Section II discusses related work. Section III present is conclusion IV Future Scope, Section V includes references.

RELATEDWORK

In this sectione some related works about action recognition in videos over the last two decades has been explore. Normally Feature representation and classification of pattern is used in recognition of action in task. Previous algorithm of action recognition is roughly split into two parts: (1) Deep learning-based methods, (2) traditional methods. 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 Histogramof Optical Flow (HOF) [10] and used to train a classifier such as SVM [11]. Different types of techniques are used in action recognition iDTs [12], SIFT3D [13],ESURF [14], andHOF [15] are used to represent motion and appearance effectively across frames in videos.Several other techniques were proposed to model the temporal structure inan efficient way, such as the temporal action decomposition [16], ranking machines [17], and dynamic poselets [18].Several, methods have been proposed for action recognition and 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. [19] discusses multiple strategies like late fusion, early fusion, and temporal fusion to extending the frame connectivity in temporal domain. Dong Li et al. [20] enhanced the accuracy of UCF101 dataset by applying temporal attention probability for each video segment in temporal sequence.

Recently, due to the development of potent GPUs and huge action datasets, deep learning has been implemented broadly on the action recognition videos.Usually, in videos two types of action form realistic and non-realistic. Non- realistic action where an actor perform some action in a scene with simple background. Handcraft methods are used for non-realistic videos to extract low level features and then trained using decision tree, SVM and for action recognition KNN. In videos, extraction of features from frames is very important to recognize. Many types of filters are applied on frames to extract low level features. Some methods for feature extraction from frames like Gabor, HoG, and CNN these all extract on the basis of texture, shape, color [21]. With the help of CNN model achieve great success in computer vision . Visin et al. [22] proposed ReNet which used for object recognition. ReNet consider four recurrent neural network that swept across the image horizontally and vertically directions.

Currently, the most efficient type of machine learning and deep learning approaches are used in classification 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 [23]. The TDD method used to extract convolutional feature maps achieved 90.3% accuracy on UCF101 dataset and 63.2% accuracy on HMDB51 dataset [24]. The 3D ResNet 101 method used by Kensho Hara et al. for recognition action in videos obtained 88.9% accuracy on UCF101 dataset and 61.7% accuracy on HMDB51 dataset [25]. In this network they used 3D frames but in our proposed method we used 2D frames. The Deep networks with Temporal Pyramid Pooling (DTPP) approach used by Jiagang Zhu et al. for recognition of action achieved 74.8% accuracy [26].Temporal segment network (TSN) approach help in good practices in learning ConvNets on video data with an accuracy of 69.4% [27]. An approach to detect action perform in a video by ActionVLAD method was used by Rohit Girdhar et al. with an accuracy of 66.9% on HMDB51 dataset [28]. Jue Wang et al. used Support Vector Machine pooled (SVMP) descriptor for action classification obtained 81.3% accuracy [29]. FC6 layer features of VGG-16 used by block-diagonal kernelized correlation pooling (BKCP) and ResNet-152 model achieved 71.3% accuracy [30].

Hangling Zhang et al. [31] classify action videos using multi stage training convolutional neural network and achieved 52.0% accuracy through temporal network on HMBD51 dataset. They classify videos by collecting 1 to 10 frames of videos and also classify action on the basis of spatial network obtained 51.4% accuracy. According to literature review Convolutional neural network is more convient as compared to other handcrafted features method. In deep learning CNN play a potent role because CNN consists of many cascading layers which extract features automatically from computational .intensive image[32].

CONCLUSION

Different types of algorithms and methods are used to retrieve action video from large dataset. In our study, it is found that CNN (Convolution Neural Networks) famous deep learning models has achieved great success in action retrieval like object tracking, computational intensive image segmentation, action recognition and so on.. Also, the small size of the model makes it convenient to be run at machines with low computational resources.

FUTURE SCOPE

This work can be further extended to increase the efficiency of CNN. Also, the performance of a CNN

depends greatly on the size of training dataset so by acquiring a larger dataset, performance and accuracy of the model can be enhanced.

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] Zahid Mehmood, Fakhar Abbas, Toqeer Mahmood, Muhammad Arshad Javid, Amjad Rehman, and Tabassam Nawaz, ‘‘Content-Based Image Retrieval Based on Visual Words Fusion Versus Features Fusion of Local and Global Features’’, Arabian Journal for Science and Engineering, Vol. 43, no. 12, pp. 7265–7284, December 2018.

[5] A. Nanda, D. S. Chauhan, P. K. Sa, and S. Bakshi, ‘‘Illumination and scale invariant relevant visual features with hypergraph-based learning for multi-shot person re-identification,’’ Multimedia Tools Appl., pp. 1–26, Jun. 2017

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, “HMDB: A large video database for human motion recognition”, IEEE, Barcelona, Spain, pp. 2556-2563, 2011.

[7] Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung wook Baik, “Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non- stationary environments, Future Generation Computer Systems, Vol. 96, pp. 386-397, July 2019.

[8] Glorot X, Bordes A, and Bengio Y , “ Deep sparse rectifier neural networks”, International conference on artificial intelligence and statistics, pp.315-323, 2011.

[9] Aasma Aslam, Babar Hussain, Ahmet Enis Cetin, Arif Iqbal Umar, and Rashid Ansari, “ Gender classification based on isolated facial features and foggy faces using jointly trained deep convolutional neural network”, journal of Electronic Images, sept. 2018.

[10] N. Dalal, and B. Triggs, “Histograms of oriented gradients for human detection**”,** IEEE Conference on Computer Vision Pattern Recognition, pp. 886-893, 2013.

[11] Zi Hau Chin1, Hu Ng1, Timothy Tzen Vun Yap1, Hau Lee Tong1, Chiung Ching Ho1, and Vik Tor Goh2, “Daily Activities Classification on Human Motion Primitives Detection Dataset”,Computational Science and Technology,Vol. 481, pp. 117-125, August 2018.

[12] H. Wang and C. Schmid, “Action recognition with improvedtrajectories”, IEEE International Conference on Computer Vision,Sydney, NSW, Australia, pp. 1550-5499, Dec. 2013.

[13] 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.

[14] G. Willems, T. Tuytelaars, and L. Van Gool, “An efficientdense and scale-invariant spatio-temporal interest point detector”, In ECCV,Berlin, Heidelberg,pp. 650 – 663, 2008.

[15] 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.

[16] 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.

[17] 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.

[18] L. Wang, Y. Qiao, and X. Tang, “ Video action detection with relational dynamic-poselets”, In ECCV, Cham,pp. 565-580,2014

[19] 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.

[20] 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.

[21] 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.

[22] Visin F, Kastner K, and Cho K, "Renet: a recurrent neural network based alternative to convolutional networks", Computer Vision pattern recognition, May 2015.

[23] 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.

[24] L. Wang, Y. Qiao, and X. Tang, "Action recognition withtrajectory-pooled deep-convolutional descriptors", In CVPR,Boston, MA, USA, pp. 1063-6919, June 2015.

[25] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporalfeatures with 3d residual networks for action recognition" In ICCV,Venice, Italy, pp. 2473-9944, Oct. 2017.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Hangling Zhang, Chenxing Xia and Xiuju Gao, "Action recognition based on multi- stage jointly training convolutional network", Multimedia Tools and Applications, vol. 78, no. 8, pp. 9919-9931, April 2019.

[32] Binay Kumar Pandey,Sanjay Kumar Pandey,and Digvijay Pandey. "A Survey of Bioinformatics Applicaion on Parallel Architectures. " Inrernational Journal of Computer Applications,vol. 23, pp.21-25, June 2011.