Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 2, 2023

A Review of Video Action Detection Based on Deep Learning

Zhuofan Zeng

Michigan State University, East Lansing, Michigan State, 48824, USA

Abstract: Currently, the application of deep learning to solving problems associated with traditional surveillance video analysis has become one of the research hot topics. The video action detection is referred to as detecting the temporal segments containing the action in the video as temporal action proposals. The existing work is mainly classified into two categories: one is to use the low-level details of video to generate action proposals; the other is to use the high-level semantics of video to generate action proposals. By deeply researching the video action detection methods based on deep learning, this paper is an attempt to find out problems with the existing methods and put forward some suggestion for improvement.

Keywords: Video action detection; Deep learning; Dataset.

1. Introduction

The majority of meaningful information in video is associated with human activities, therefore, the detection of human action in a video is the foundation and important part of video analysis, and is also one of the hot research directions in the field of image processing, pattern recognition and computer vision. The human action detection is a crucial video understanding method, which aims to use machine learning and image processing techniques to extract meaningful information from the collected video sequences (surveillance video, etc.), and automatically locate and identify specific action contained in a video, so that the computer can have an understanding of human action and consequently make analysis of the video. The rapid development of action recognition has spurred the emergence of action detection technology. The action detection includes classification of action recognition, localization, identification of the categories of action contained in a video, as well as the determination of temporal intervals (starting time and ending time) when the action occurs, which makes human action detection more in line with the needs of practical application. The accuracy of video action detection has been difficult to improve due to various and constantly changing video backgrounds, intrinsically complicated and changeable human action and other factors. In order to solve the above-mentioned problems, researchers have put forward a variety of solutions, of which deep learning is the most effective. The concept of deep learning was proposed by Hinton et al. [1] in 2006, and has been remarkably used in image processing, speech recognition, machine translation, natural language processing and other fields. Inspired by the huge success of deep learning in image and natural language processing, some scholars have attempted to apply it to video understanding tasks and others. Various indicators have been significantly improved, however, greater improvement cannot be made by simply applying image processing to the video field, since the video not only includes spatial information, but also has temporal correlation between frames. Therefore, the focus of current research has been on how to accurately express and integrate the temporal and spatial information in a video.

Figure 1. Schematic diagram of traditional manual surveillance video system

2. The general framework of video behavior detection

The general framework of video action detection is shown in Figure 2, which can be divided into four stages, namely, video pre-processing, feature extraction, feature learning and anomaly judgment. Video pre-processing is usually referred as to the morphological processing of video frames, such as image enhancement, image noise reduction, etc. Feature extraction is referred as to the extraction of key appearance or motion features of the input video sequences to represent a video. Feature learning is referred as to using extracted features to establish an action model of normal or abnormal event patterns and their changing rules. Anomaly detection is used to measure the degree of matching between basic events in the test video and the established action model, with the detection results returned in the form of anomaly scores or classification labels.

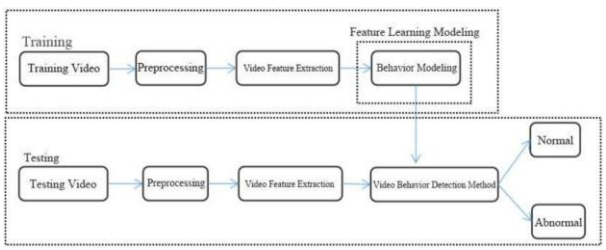

Figure 2. General framework of video action detection

The earliest research on human action can be traced back to 1975, when Johansson et al. proposed that human action could be described by 12 points [2], laying a foundation for the development of human action detection. Recently, with the rapid development of deep learning and the enhancement

of computing power, video processing capabilities have been remarkably improved, significantly promoting the development of action detection and contributing to many important algorithms and ideas. Currently, the first step in the general process of action detection algorithms based on deep learning is to extract video features using some modelling for action recognition, after which the extracted features remain fixed. On this basis, recognition and localization are subsequently carried out. Therefore, the following will make an introduction to relevant research progress of action recognition and action detection successively.

Prior to 2015, the majority of frameworks of action recognition and detection followed a fixed process, that is, the spatial and motion features of videos are extracted using manual methods, and the action is subsequently detected by independently using classifiers such as the support vector machine (SVM) [3]. Of which, human features are generally comprised of global features and local features. Global features are used to extract features by regarding the whole human action as the description object, for example, the acceleration vector and motion trajectory information were used in literature [4] to make a judgement whether any action was generated in a video. However, global features are sensitive to changes in environment background, shooting angles and techniques, resulting in the relatively low detection accuracy rate. Local features are used to describe partial motion features of an action, and generally represent motion features of video using local feature descriptors. In this study, Yang et al. [5] introduced a multi-scale feature descriptor based on the feature representation method of Histograms of Optical Flow (HOF), which can simultaneously obtain the displacement and spatial features of video frames. Wang et al. proposed the improved Dense Trajectories (iDT), which carries out intensive sampling and tracks pixel points on video frames to construct local feature descriptors [6][7], and can estimate the motion of the camera by matching feature points between frames. These methods have achieved satisfactory results in specific scenarios, however, the extraction of human features requires huge computing cost, and feature computing is generally aimed at a specific video type. Therefore, it is difficult to achieve its expansion and deployment, and there are still some limitations on video processing application.

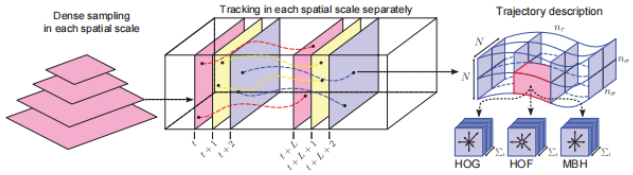

Figure 3. iDT algorithm and three kinds of histograms

3. The Convolutional Neural Network (CNN) applied to video behavior detection

3.1. 3D Convolutional Networks

With the rise of deep learning, some scholars have started to attach attention to how to achieve the application of Convolution Neutral Network (CNN) to video processing. Li Feifei's team[8] first proposed the application of a single 2D CNN to each video frame in a successive manner, and researched some temporal connectivity patterns to learn the temporal and spatial features of videos. The 3D convolution was first proposed by JiS et al. [9] in 2013, which adds a dimension on the basis of 2D convolution in order to extract the features in the temporal dimension. On this basis, Facebook Trans D et al. used the 3D convolution and 3D pooling to construct 3D Convolutional Networks (C3D)[10], which can be applied to action recognition, scenario recognition, video similarity analysis and other fields. With C3D networks combined with the traditional machine learning classification, the support vector machine (SVM) can achieve 85.2% accuracy rate on the UCF101 dataset. The accuracy is not high, but the speed of C3D network is very fast, reaching 314FPS in the paper. In 2017, Facebook Carreira J et al. proposed Inflated 3D Convolutional Networks (I3D) [11] (which adopted RGB images as input, and input the optical flow into the networks), and a two-stream 3D convolution model by combining with the fusion model of two-stream network, which can achieve 98% accuracy rate on the UCF101 dataset. Based on the video spatio-temporal features extracted by the 3D convolutional networks, the spatial features are aggregated by subsampling in the spatial domain. Subsequently, a temporal up-sampling operation is performed in the temporal domain to restore the features to the original lengths of the input video sequences as much as possible in the temporal dimension. The advantage of this method is that more precise temporal boundaries can be detected in the temporal domain, thus improving the performance of temporal action detection. The 3D convolution can learn the spatial information about each frame and the temporal correlation between frames, however, the 3D convolution requires a large amount of computation and needs to be trained from the start without pre-training parameters, which makes it difficult to achieve convergence and to deploy the 3D convolution on embedded terminals, thus limiting the practicability of 3D convolution. In order to solve these problems, Qiu Z et al. [12] proposed Pseudo-3D convolutional Networks (P3D) in 2017, which decomposes the 3D convolution into pseduo-2D convolution and pseudo-1D convolution, for example, 1×3×3 spatial convolution and 3×1×1 temporal convolution are used to replace 3×3×3 3D spatio-temporal convolution, which can extract video spatio-temporal features using 3D structure, and reduce the amount of model computation, laying a foundation for the practical application of 3D convolution.

3.2. Two-stream convolutional networks

In 2014, Simonyan K et al. [13] proposed the two-stream method, which uses spatial and temporal information of videos to train models respectively, and subsequently fuses its output to obtain video-level feature representation, and eventually inputs the video features into traditional machine learning classifiers to obtain recognition results. In this method, the spatial information uses RGB pictures directly extracted from the video, and the temporal information uses the dense optical flow calculated every two frames in the video sequences. The method can achieve 88% accuracy rate on the UCF101 dataset. Hereafter, the focus of research has been on the improvement in the two-stream network. In 2016, Feichtenhofer C et al. [14] made an improvement in the feature extraction network and fusion method of the two-stream network, using VGG16[15] to extract video features and using the Convolutional Neural Networks (CNN) to obtain fusion weights, which improved classification accuracy rate to 92.5% on the UCF101 dataset. Yue-Hei Ng J et al. [16] adopted Long Short-Term Memory (LSTM) as the

fusion network of temporal and spatial stream, with the accuracy rate reaching 88.6% on the UCF101 dataset. Limin Wang et al. [17] proposed Temporal Segment Networks (TSN) in 2016 and comprehensively discussed how to improve the effect of the two-stream network from the aspects of input data mode, video sampling strategy and basic network for extracting spatial features. It is currently the most popular action recognition model, with the accuracy rate reaching 94.2% on the UCF101 dataset.

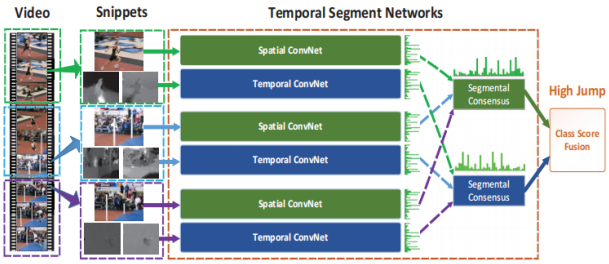

Figure 4. Feature extraction diagram of two-stream network

3.3. The action detection algorithms

After obtaining the spatio-temporal feature representations of a video, the current action detection algorithms can be divided into two categories: two-step method and one-step method. The research focus is laid on temporal boundary proposal technology, which is mainly composed of three methods, namely, sliding window [18], temporal actionness grouping [19] and temporal unit regression [20].

3.3.1. The method of sliding window

The method of sliding window was proposed by Shou Z et al. [18] in 2016 in the Segment Convolutional Neural Network (SCNN). Just as the name implies, windows of different sizes are used to slide on the video, and windows of more diverse sizes can achieve a better detection effect. SCNN is a three-stage model. In addition to temporal boundary proposal, SCNN also includes two stages: binary classification and multi-classification. Before multi-classification, the binary classification is employed to filter the candidate temporal intervals, which can reduce redundancy to a certain extent. The feature extraction network used in SCNN is C3D, with the detection accuracy rate reaching 19.0% on the THUMOS14 dataset. On the basis of the sliding window, the Convolutional De-Convolutional networks (CDC) [21] uses the structures of convolutional and deconvolution to make fine adjustment of the interval boundaries, with the detection accuracy rate reaching 23.3% on the THUMOS14 dataset. The candidate temporal intervals obtained by the sliding window create much redundancy, which increases the computation, and cannot adapt to the flexible action length due to the fixed interval boundary. In order to solve above-mentioned problems, the temporal actionness grouping algorithm and the temporal unit regression algorithm are proposed successively.

3.3.2. The temporal action grouping algorithm

In 2017, Xiong Y et al. [19] proposed a temporal action grouping algorithm, which uses the trained action recognition model to perform binary classification of each segment in the video to obtain the action probability sequences, and subsequently uses different threshold values and grouping methods to obtain multiple candidate temporal intervals. Using the TAG Network, Zhao Y et al. [22] proposed the Structured Segment Network (SSN), which adds contextual information to the candidate temporal intervals, constructing a pyramid structure, to represent the spatio-temporal features of each action. Inspired by this, in 2018, Qiu et al. [23] used the Recurrent Neural Network (RNN) structure to capture contextual information of video action and added non-local features to the top of the pyramid structure of SSN, with the detection accuracy rate reaching 34.2% on the THUMOS14 dataset.

3.3.3. The temporal unit regression algorithm

The temporal unit regression algorithm calibrates with the label intervals by regression of the boundaries of the candidate temporal interval, based on the boundary regression method in literature [24]. On this basis, Gao et al. [25] proposed a Cascaded Boundary Regression (CBR) in 2017, with an aim to solve the shortcoming that the temporal proposal boxes obtained by the sliding window may merely contain part of action segments by means of cascaded boundary regression. Meanwhile, the video is divided into several small units by the method of extracting video temporal and spatial features in the Temporal Unit Regression Network for Temporal Action Proposals (TURN TAP). The spatial and temporal features of each unit are computed, and subsequently the sliding window is operated based on the computed unit features. This feature extraction method avoids the double computation caused by the method of sliding over the video image sequences and then extracting the features, thus improving the efficiency of feature extraction. Using temporal unit regression, a candidate temporal interval is found for each action, and the highest score is taken as the final temporal interval, with the detection accuracy rate reaching 31.0% on the THUMOS14 dataset.

In order to obtain more flexible boundaries, in 2017, Xu H et al. [26] proposed the Region 3D Convolutional 3D Network (R-C3D), which combines temporal boundary proposal and classification for training. In 2018, Chao Y W et al. [27] made an improvement in this structure by adding optical flow information and contextual information of action, making it more adaptable to the action detection. In addition, Gao J et al. [28] made a combination of sliding window and temporal actionness grouping algorithm to generate candidate temporal intervals, with the detection accuracy reaching 29.9% on the THUMOS14 dataset. Lin T et al. [29] proposed a Boundary Sensitive Network (BSN), which uses CNN model to predict the probabilities that action begins, ends and is ongoing, for each temporal position in a video, to generate the starting, ending and ongoing probability sequences. Subsequently, these three probability sequences are processed to obtain a more accurate candidate temporal interval. Combined with the classification method in SCNN, the detection accuracy rate reaches 29.4% on the THUMOS14 dataset.

In 2019, Lin T et al. [30] improved the matching method of the three probability sequences, and proposed the Boundary Matching Network (BMN). Combined with the classification method in SCNN, the detection accuracy rate is improved to 32.2% on the THUMOS14 dataset. The one-step method can simultaneously classify, locate and directly output the results of action prediction. In 2016, Li Fei fei's team used reinforcement learning to train an agent based on RNN, which can directly generate action prediction [31]. In 2017, Lin T et al. [32] drew on the framework's structure of Single Shot MultiBox Detector (SSD) [33], and first proposed the Single Shot Temporal Action Detection (SSAD) model, which performs classification and regression only once on the whole network, however, the detection accuracy on the THUMOS14

data set was only 24.6%. In 2019, Huang Y et al. [34] made improvement of this structure by adding classification branches and regression branches, with the detection accuracy increasing to 35.8%.

In addition to the two-stage method mentioned above, some scholars have carried out exploration of methods based on one-stage temporal action detection. The one-stage temporal action detection method does not generate temporal proposal boxes, but directly uses the network to detect all candidate windows. Lin et al. [35] used one-dimensional convolution to model multi-scale features to generate predefined anchor boxes at each temporal position with multi-scale features, to directly classify the action category of each anchor box, and simultaneously to perform a regression of the boundary box offset. Based on the Recurrent Neural Network (RNN), Buch et al. [36] performed the detection of temporal action boundaries, while carrying out classification of action in a video. Yeung et al. [37] explored the usage of RNN to predict the starting and ending temporal points of action in an end-to-end manner. Huang et al. [38] explored the detection and classification methods in the one-stage detection algorithm of decoupling. In this method, two parallel branches are designed, each having a separate feature representation to decouple the detection and classification processes.

4. Future expectations

Currently, research at home and abroad has basically achieved the action detection in unsegmented video, however, the detection efficiency is low, and the application of mobile and embedded platforms has not been popularized. In terms with detection accuracy rate, in 2017, the average detection accuracy rate reached about 25%, and since 2018, the average detection accuracy has been merely 35% or so. In terms of detection speed, researchers generally use the 3D convolution to improve the computing speed. For example, the detection speed of R-C3D networks proposed by Xu H et al. reaches 569FPS, however, due to the large amount of 3D convolution computation, this method is difficult to optimize, with its detection accuracy rate merely reaching 29.8%. Therefore, it is urgent to find an efficient action detection algorithm. In the mobile and embedded field, video-based human action detection generally adopts traditional image and video processing methods. For example, Ding Chan achieved the moving object detection on embedded platform using foreground and background extraction methods; Xi Lu achieved a smart home monitoring system using iDT based on the ARM platform.

Since the action detection models based on deep learning have a high requirement for computing capacity, there is a severe delay in developing and deploying the embedded platforms, causing such models cannot step into the practice. Therefore, I believe that the research of video action detection algorithm based on deep learning is worth further researching in the direction of deployment on embedded devices.

References

[1] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[2] Johansson G. Visual motion perception[J]. Scientific American, 1975, 232(6):76-89.

[3] Burges C . A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2):121-167.

[4] Datta A, Shah M, Lobo N D, et al. Person-on-person violence detection in video data[C]//In the International Conference on Pattern Recognition. 2002: 433-438.

[5] Cong Y, Yuan J, Liu J, et al. Abnormal event detection in crowded scenes using sparse representation[J]. Pattern Recognition, 2013, 46(7): 1851-1864.

[6] Wang H, Klaser A, Schmid C, et al. Action recognition by dense trajectories[C]//In the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2011: 3169-3176.

[7] Wang H, Schmid C. Action recognition with improved trajectories[C]//In the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2013: 3551-3558.

[8] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks[C] //In the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2014.

[9] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221-231.

[10] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE international conference on computer vision. IEEE, 2015: 4489-4497.

[11] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 6299-6308.

[12] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]. proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017: 5533-5541.

[13] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]. Advances in Neural Information Processing Systems. IEEE, 2014:568-576.

[14] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016:1933-1941.

[15] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J/OL]. arXiv preprint arXiv:1409.1556, 2014.

[16] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2015: 4694-4702.

[17] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. European Conference on Computer Vision. IEEE, 2016: 20-36.

[18] Shou Z, Wang D, Chang S F. Temporal action localization in untrimmed videos via multistage cnns[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 1049-1058.

[19] Xiong Y, Zhao Y, Wang L, et al. A pursuit of temporal accuracy in general activity detection[J/OL]. arXiv preprint arXiv:1703.02716, 2017.

[20] Gao J, Yang Z, Chen K, et al. Turn tap: Temporal unit regression network for temporal action proposals[C]. Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017: 3628-3636.

[21] Shou Z, Chan J, Zareian A, et al. CDC: Convolutional-De-Convolutional networks for precise temporal action localization in untrimmed videos[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2017: 5734-5743.

[22] Zhao Y, Xiong Y, Wang L, et al. Temporal action detection with structured segment networks[C]. Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017:2914-2923.

[23] Qiu H, Zheng Y, Ye H, et al. Precise Temporal Action Localization by Evolving Temporal Proposals[C]. Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. ACM, 2018:388-396.

[24] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. Advances in neural information processing systems. IEEE, 2015: 91-99.

[25] Gao J, Yang Z, Nevatia R. Cascaded boundary regression for temporal action detection[J/OL]. arXiv preprint arXiv:1705.01180, 2017.

[26] Xu H, Das A, Saenko K. R-C3D: Region convolutional 3d network for temporal activity detection[C]. Proceedings of the IEEE international conference on computer vision. IEEE, 2017: 5783-5792.

[27] Chao Y W, Vijayanarasimhan S, Seybold B, et al. Rethinking the Faster R-CNN Architecture for Temporal Action Localization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018:1130-1139.

[28] Gao J, Chen K, Nevatia R. Ctap: Complementary temporal action proposal generation[C]. Proceedings of the European Conference on Computer Vision. IEEE, 2018: 68-83.

[29] Lin T, Zhao X, Su H, et al. BSN: Boundary sensitive network for temporal action proposal generation[C]. Proceedings of the European Conference on Computer Vision. IEEE,2018:3-19.

[30] Lin T, Liu X, Li X, et al. BMN: Boundary-matching network for temporal action proposal generation[C]. Proceedings of the IEEE International Conference on Computer Vision.IEEE, 2019:3889-3898.

[31] Yeung S, Russakovsky O, Mori G, et al. End-to-end learning of action detection from frame glimpses in videos[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 2678-2687.

[32] Lin T, Zhao X, Shou Z. Single shot temporal action detection[C]. Proceedings of the 25th ACM international conference on Multimedia. IEEE, 2017: 988-996.

[33] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot Multibox Detector[C]. European conference on computer vision. IEEE, 2016: 21-37.

[34] Huang Y, Dai Q, Lu Y. Decoupling Localization and Classification in Single Shot Temporal Action Detection[C]. IEEE International Conference on Multimedia and Expo. IEEE, 2019: 1288-1293.

[35] Lin T W, Zhao X, Shou Z. Single shot temporal action detection[C]. Mountain View: 25th ACM International Conference on Multimedia, 2017: 988-996.

[36] Buch S, Escoricia V, Ghanem B, et al. End-to-end, single-stream temporal action detection in untrimmed videos[C]. London: 28th British Machine Vision Conference, 2017: 213-225.

[37] Yeung S, Russakovsky O, Mori G, et al. End-to-end Learning of Action Detection from Frame Glimpses in Videos[C]. Las Vegas: 29th IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2678-2687.

[38] Huang Y P, Dai Q, Lu Y T. Decoupling localization and classification in single shot temporal action detection[C]. Shanghai: 2019 IEEE International Conference on Multimedia and Expo, 2019: 1288-1293.