Academic Journal of Engineering and Technology Science
ISSN 2616-5767 Vol.6, Issue 12: 39-44, DOI: 10.25236/AJETS.2023.061206

Research on Human Action Analysis and Recognition Methods Based on Deep Learning

Jiale Zhang

School of Information Engineering, Heilongjiang University of Finance and Economic, Harbin, 150000, China

Abstract: *With the vigorous development of the Internet and multimedia technology and the large-scale popularization of video capture devices such as smartphones and surveillance cameras, video data has shown explosive growth. The goal of action recognition is to recognize the action being performed by the human in a video. As a basic but extremely challenging task in computer vision, it has a broad application prospect in many fields, such as human-machine interaction, virtual reality, intelligent video surveillance, and social public security. In this paper, human action analysis and recognition methods based on deep learning are summarized, and several prevailing action recognition algorithms are introduced and categorized in detail. Different from the traditional classification methods, we survey the currently popular algorithms into 2 series from the perspective of feature fusion: 2D convolutional series based action recognition algorithms and 3D convolutional series based action recognition algorithms.*

Keywords: *deep learning, human action analysis, human action recognition*

1. Introduction

In recent years, with the vigorous development of the Internet and multimedia technology, video capture devices such as smartphones and surveillance cameras have become widely used. According to statistics, the total number of cell phones in China now exceeds 1.3 billion. More than 35 million cameras are deployed and a large number of videos are recorded and shared every day. As shown in Fig. 1, as of 2020, about 800 hours of video were uploaded to the YouTube website per minute, which is the world's largest video platform. It is almost impossible to analyze and process these massive influxes of video data only relying on traditional manual ways.[1] Therefore, how to use computers to achieve fast and automated video analysis and understanding has become one of the fundamental problems in computer vision. Since human is the main part of the videos, human actions have become important features to describe the video content, so the human action analysis and recognition is the key and difficult problem in the field of video understanding. It analyzes human actions in video sequences and creates corresponding relations between video content and action categories, so that computers can "read" and "understand" human actions in videos as human does. With the gradual increase of video data size, the wide application of high-performance computing equipment and the development of information technologies related to neural network, human action analysis has changed from the traditional action recognition task towards short video classification to tasks towards long video understanding in present stage, including action prediction, online action recognition, temporal action detection, and cross-modal video action localization. All these tasks put forward stricter demands of human action analysis and recognition technology.

Figure 1. The explosive growth of multimedia data in the Internet era

ISSN 2616-5767 Vol.6, Issue 12: 39-44, DOI: 10.25236/AJETS.2023.061206

2. Main Types of Human Action Analysis and Recognition

The aim of action recognition is to identify the actions that appear in a video, usually the actions of the person in the video. A video can be regarded as a data structure consisting of a set of image frames arranged in temporal order, which owns an extra time dimension compared with an image. Action recognition would not only analyze the content of each image frame in a video, but also mine clues from the time-series information between video frames. Action recognition is a core field of video understanding. Although action recognition focuses on recognizing human actions in videos, most of the algorithms developed in this field are not specified for humans and could also be used in other video classification scenarios as well.

2.1 Offline Action Recognition and Online Action Recognition

Action recognition includes offline action recognition and online action recognition. As shown in Fig. 2, offline action recognition requires the determination of the category of human actions occurring in a video after observing the entire video sequence. Usually, the task assigns an action category label to each video based on a predefined action list. Meanwhile, the model needs to learn the mapping relationship from video content to action categories so as to achieve accurate classification of actions, which is essentially a video classification task. Meanwhile, Online action recognition is oriented to the needs of practical application scenarios. It would require a real-time processing of online video streams, and achieve action recognition on each frame of the video, trying to detect the beginning of the action before it really occurs.

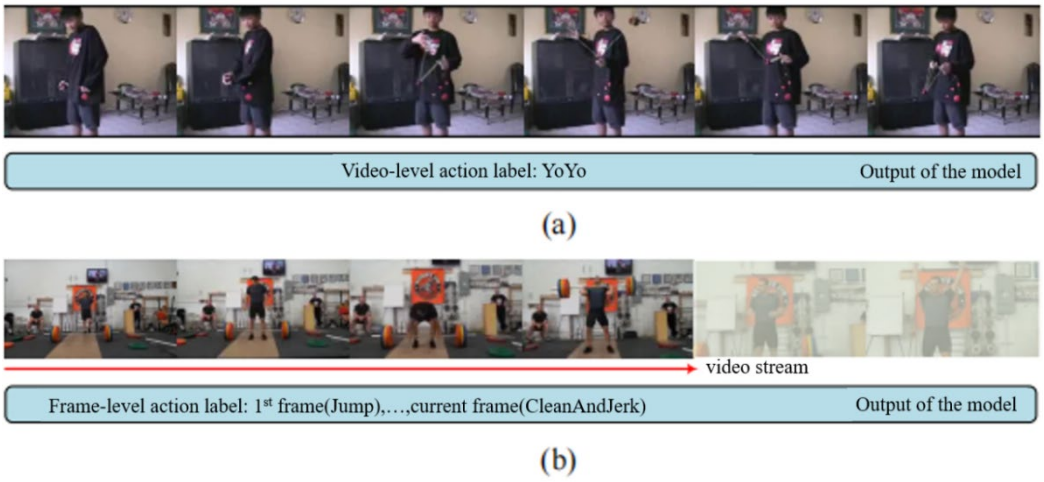

Figure 2. (a) Offline action recognition (b) Online action recognition

In traditional video action recognition tasks, the processed videos are usually cropped short video that contain only the action segments of interest. However, in actual scenarios, long uncropped videos are more common. Online action recognition mainly deals with long uncropped videos, which usually contain a large number of irrelevant background clips, making it a challenging task to accurately determine the start of action in them.

2.2 Action Detection

Action detection includes temporal action detection and spatio-temporal action detection. As shown in Fig. 3, for a long uncropped video sequence, the main purpose of the temporal action detection task is to locate the start and end time points of the target action and the corresponding action categories. Meanwhile, the spatio-temporal action detection would additionally predict the spatial location where the action occurs. The temporal action detection algorithm generally consists of two steps. First, a candidate action segment with precise temporal boundaries is generated. Second, action categories for the candidate temporal action segment are classified.

ISSN 2616-5767 Vol.6, Issue 12: 39-44, DOI: 10.25236/AJETS.2023.061206

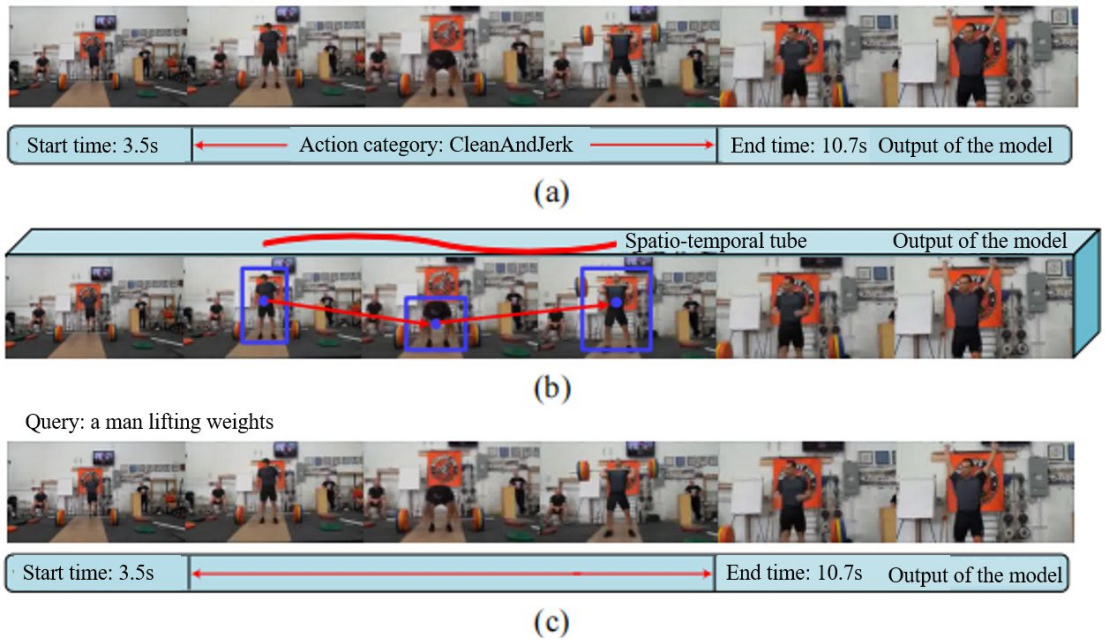

Figure 3. (a) Temporal action detection (b) Spatio-temporal action detection (c) Video action localization based on natural language description

2.3 Action Prediction

Action prediction includes early action prediction and long-term action prediction. As shown in Fig. 4, early action prediction forecasts the category of actions that are occurring in a video clip at the early stage of action being executed according to that observed video clip. Such a task is mainly targeted at cropped short videos, which are generally short in duration and contain only one action category. Therefore, this type of action prediction task is aimed at determining "what is happening now" at the early stage of action completion. Long-term action prediction is to predict possible action categories in the future by observing human actions at current moment the in the video, i.e., to answer the question of "what is going to happen". Such a task would be applied into long and uncropped videos that are several minutes or even more than ten minutes long and often involve complex action categories with multiple human targets. Due to the high spatial and temporal uncertainty of future human actions, long-term action prediction is extremely challenging.

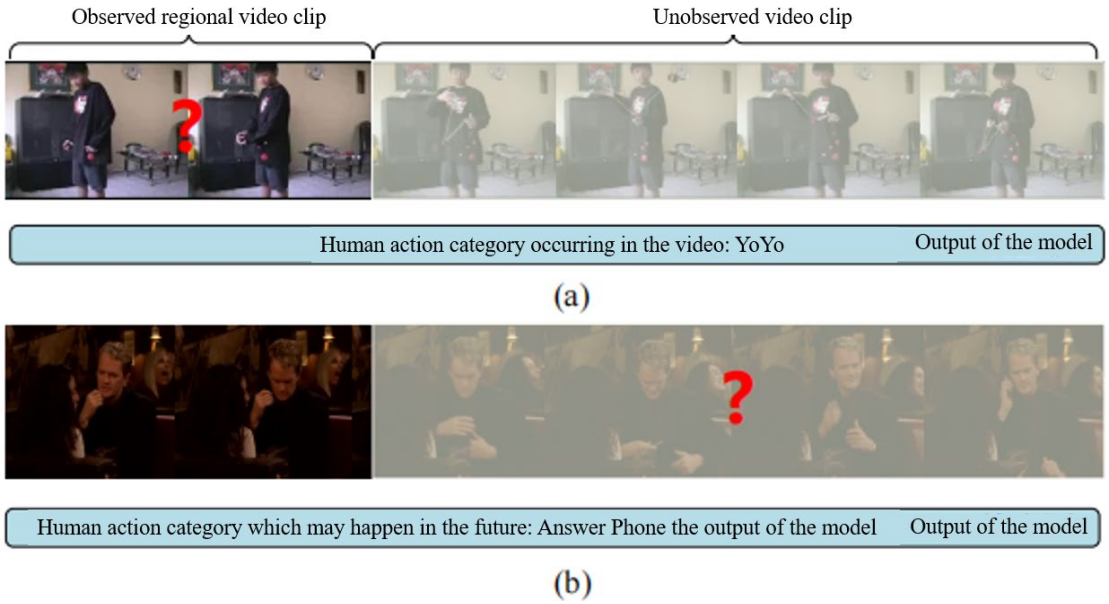

Figure 4. (a) Early action prediction (b) Long-term action prediction

ISSN 2616-5767 Vol.6, Issue 12: 39-44, DOI: 10.25236/AJETS.2023.061206

3. Typical Methods for Human Action analysis and Recognition

Human action recognition research involves the crossing and integration of various disciplinary fields, including image processing, artificial intelligence, pattern recognition, optical technology, etc. In addition, due to the influence of dynamic information in the background, the difficulty of research on human action recognition has been significantly increased because of lighting interference, video processing complexity and other factors. Hence, the implementation of video-based human action recognition is extremely challenging. At present, typical methods for human action analysis and recognition that have been proposed in related research are as follows:

3.1 Traditional Manual Feature Extraction Methods

Traditional manual feature extraction is the manually designed method in which reasons for each step could be outlined. Prevailing feature extraction methods include Harris, SIFT, SURF, LBF, HOG, etc. Vinegar *et al.* proposed a new set of corresponding relations among joints (Trisarea feature)[2], which is defined as the area of triangles formed by the three joints. It could be used to recognize relevant triangles which describe human poses, and show changes in the selected Trisarea features over time so as to constitute the descriptors of human actions. Yan *et al.* achieve human action recognition and classification based on the human skeleton dataset by the local joint structure (LJS) and the histogram of three dimensional joints (hoj3D) methods[3]. Firstly, two kinds of features are extracted from the 3D positions. Then, the linear discriminant analysis is used to reduce the dimensionality and K-means clustering is used to generate pose code words. Particularly, the local joint structure is used to describe the local relationship among different joints, and the 3D joint histogram is used to describe the global distribution of joints in 3D space.

3.2 Action Recognition Based on 2D Convolutional Neural Networks

(1) Two-stream network

Since convolutional neural networks have achieved high accuracy in image classification, some researchers have started to try to apply convolutional neural networks into action recognition tasks for video understanding. However, although both action recognition and image classification belong to classification tasks, there are many challenges in dealing with video features due to the different modalities of the data objects. In action recognition tasks, in addition to the static information contained in the video (e.g., human positions, human poses, human sizes, scenarios, etc.), the dynamic information between video frames is also important for categorizing action labels. However, the neural network structures migrated from the image classification task do not have the ability to understand dynamic information. To solve this problem, optical flow is proposed to represent dynamic features.

Optical flow represents information about the change in luminance at a corresponding location in an image. The luminance features of the corresponding pixels would change according to the movement of a human or an object in a video. In 2014, Simonyan *et al.* first proposed a two-stream network approach.[4] The network is based on the inspiration of the two-stream hypothesis. Two independent convolutional neural networks are used to calculate the action prediction scores of RGB images and stacked optical flows, respectively. Then the average respective scores are calculated as the recognition accuracy of the model. Fig. 5 shows the two-stream architecture for video classification. In particular, RGB images represent static feature signals, while stacked optical flows represent the varied information of pixel points between consecutive video frames. By using the complementary of the two visual information, i.e., static features and temporal features, the two-stream network could effectively integrate the classification results of spatial and temporal features to improve the recognition accuracy of the model, and avoid the error of recognition accuracy caused by single vision.

ISSN 2616-5767 Vol.6, Issue 12: 39-44, DOI: 10.25236/AJETS.2023.061206

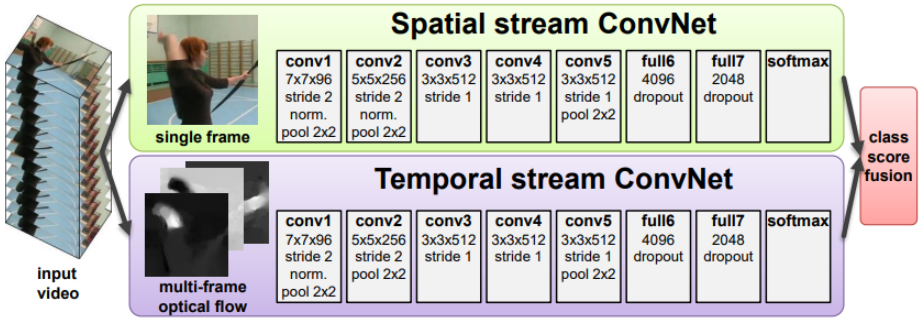

Figure 5. Two-stream architecture for video classification

(2) Network of dependency relation in temporal feature

Although optical flow can represent the instantaneous dynamic information of the human body, it still has some drawbacks. For example, the optical flow could be easily interfered with changes in light intensity and could not represent the dynamic connection between frames over a long duration of a video. Therefore, in order to solve these defects, some researchers tried to improve the model's ability to learn dynamic features from the structural design of neural networks. One of the prevailing ideas in recent years is to use CNN to learn the static features of single-frame images, and then use RNN to learn the dynamic features of the video. In 2015, Joe *et al.* proposed to use AlexNet and GoogLeNet as the backbone of the CNN to compute the image features of each frame, and then encode the features by feature pooling and feature aggregation. Furthermore, the temporal dependency information is obtained by putting the encoded information into the LSTM structure[5]. By combining models in such a manner, the drawbacks of CNN, i.e., the ability to learn dynamic features in the temporal dimension, could be largely bridged.

3.3 Action Recognition Algorithm Based on 3D Convolutional Neural Networks

3D convolution was first proposed by Ji *et al.*[6] The idea is to add the original 2-dimensional convolutional kernel operator with a one-dimensional depth channel, which could be used to represent a continuous frame in a video or different slices in a stereoscopic image. The motivation is to allow the convolutional neural network to compute spatial features and temporal features at the same time. In the early days, Tran *et al.* were the first to experimentally explore the size of the 3D convolutional kernel that best fits the video feature extraction. Several experiments were conducted on the UCF-101 dataset with four fixed sizes of 3D convolutional kernels. Experimental results showed that a convolutional kernel size of 3 × 3 × 3 could give the best results. The classical convolutional 3D (C3D) neural network (shown in Fig. 6) was built by using this size of convolutional kernel. C3D contains 8 convolutional layers, 5 pooling layers, 2 fully connected layers and 1 softmax output layer. All convolutional layers have a step size of 1×1×1. The first pooling layer has a size of 1×2×2 and a step size of 1×2×2. The rest of the pooling layers have a size of 2×2×2 and a step size of 2×2×2.

The size of the video slices input to C3D is 3×16×112×112, where 3 denotes the number of channels while 16 denotes the number of frames. Besides, 112×112 denotes the size of the image. During the process of the video slices entering into the C3D for feature computation, the temporal features of the video slices will also be aggregated continuously with the forward propagation of the network. After going through all the convolutional and pooling layers, the final feature map is input to the fully connected layer with 4096 output units for prediction.

3D convolution could be a good way to merge spatial and temporal features for parallel learning in video processing tasks, which could significantly improve the efficiency of spatio-temporal feature fusion and strengthen the spatio-temporal modeling capability of neural networks. However, the disadvantage of expanding from 2D to 3D models is that the number of parameters of the network would also be geometrically increased. At the same time, since 3D convolutional networks need to input multiple frames at the same time, the computation cost of 3D convolutional network is also much higher than that of 2D convolutional networks. The foremost advantage of 3D convolutional model is that it could solve the problem of temporal modeling and can extract video features comprehensively, but it also brings the problem of an excessive amount of parameter computation, which would make the model easy to overfitting and difficult to optimize. Currently, some variants try to reduce the computation cost of 3D convolution by decomposing 3D convolution into 2D convolution and 1D time convolution, such as decomposing 3D convolution into 2D spatial convolution and 1D time convolution, or mixing 2D CNN

ISSN 2616-5767 Vol.6, Issue 12: 39-44, DOI: 10.25236/AJETS.2023.061206

and 3D CNN[7].

4. Conclusion

Human action recognition is a popular research direction in the field of computer vision, which has significant application value in video surveillance, human-machine interaction, motion detection and other fields. In this paper, a comprehensive overview of the feature extraction methods involved in the field of action recognition is provided. We summarize the current research and analyze the advantages and disadvantages of the traditional manual feature extraction methods and deep learning-based feature extraction methods. Moreover, the existing challenges and possible future research directions in the field of human action recognition are summarized so as to provide references for the relevant researchers.

References

[1] Information on: https://www.groupisd.com/what-happens-online-in-60-seconds/
[2] VINAGRE M, ARANDA J, CASALS A. A New Relational Geometric Feature for Human Action Recognition [C]. proceedings of the Lecture Notes in Electrical Engineering, 2015. 263-278.
[3] YAN L, LU W, WEI L, et al. Action Recognition Using Local Joints Structure and Histograms of 3D Joints [C]. proceedings of the 2014 Tenth International Conference on Computational Intelligence and Security, 2015. 185-188.
[4] Simonyan, Karen, Zisserman, et al. Two-Stream Convolutional Networks for Action Recognition in Videos[C]. Advances in Neural Information Processing Systems, 2014: 568–576.
[5] Ng Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015: 4694-4702.
[6] 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
[7] Diba A, Fayyaz M, Sharma V, et al. Temporal 3d convnets: New architecture and transfer learning for video classification[EB/OL]. 2017, 11 22. arXiv preprint arXiv:171108200. https://doi.org/10.48550/arXiv.1711.08200.