Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230806

Video action recognition: A survey

Qingqing Duan

Beijing University of Posts and Telecommunications, Beijing, 100876, China

18830389008qing@bupt.edu.cn

Abstract. In recent years, as computer vision technology advances quickly, video action recognition has become a research hotspot. However, the field of video action recognition still faces many problems and challenges, such as the complex temporal dynamics of video data and the high computation cost caused by the explosive growth of video data. This paper summarizes the algorithms of video action recognition, and introduces them in two parts, namely, traditional handcrafted representation methods and deep learning representation methods. Among them, the traditional handcrafted representation methods are divided into two categories: Holistic Representation Methods and Local Representation Methods; The deep learning representation methods can be divided into RGB data based methods and Kinect-Based methods according to the type of input data modality. This paper focuses on the methods based on RGB data. This paper introduces the representative achievements according to different research directions, compares the advantages and disadvantages of various algorithms, and finally puts forward the prospect of the future development direction.

Keywords: action recognition, Two-Stream, 3D CNN, LSTM, transformer.

1. Introduction

Before starting the research, we first need to determine what is an action. According to Wang et al. [1], an action's actual significance is found in the change or transformation it causes in the environment, which is in line with the goals of our research.

Action representation, also referred to as feature extraction, and classification are the two phases that often used to accomplish action recognition. The basis of video action recognition is action representation, which includes extracting discriminative posture and motion information from recordings of human movements. Effective action representation should be discriminative, efficient and low dimensional. Action learning and classification are processes that involve extracting information, learning statistical models from those features, and utilizing those models to categorize newly observed features [2].

In this paper, we have sorted out some typical video action recognition algorithms in recent years, and classified them as shown in Figure 1.

DOI: 10.54254/2755-2721/6/20230806

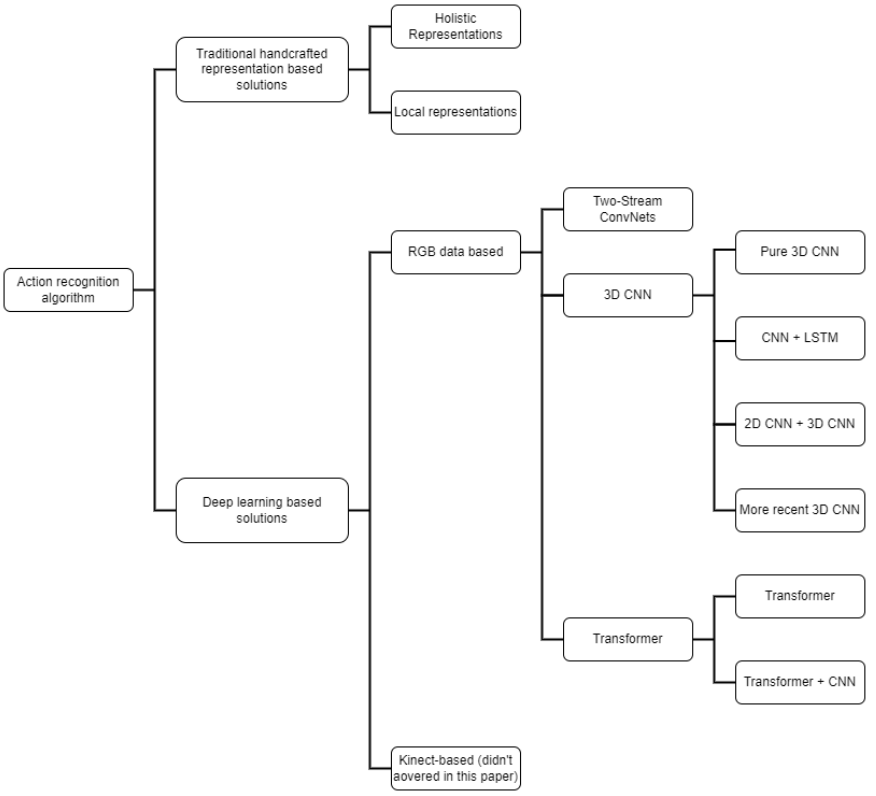

Figure 1. Classification of video action recognition methods.

2. Traditional handcrafted representation methods

Traditional handcrafted representation methods take the artificial design feature as the input of machine learning model, and then output the predicted category. The quality of machine learning model depends on the representation ability of features. Extracting good features is the key to building a machine learning model. It's necessary to design a good feature to clearly distinguish different targets to improve the accuracy of machine learning algorithm. Then, these features are input to the classifier, e.g., Support Vector Machine (SVM), Adaboost, etc., and then the predicted categories are output. The whole process is displayed in Figure 2.

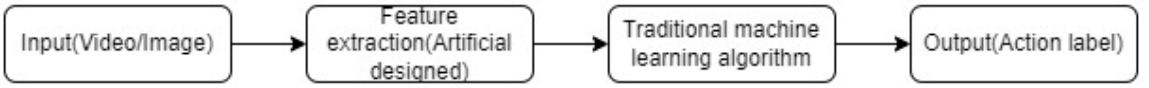

Figure 2. Video action recognition flow based on traditional machine learning method.

Traditional handcrafted representation methods can be split into two categories: global representations (or holistic representations) and local representations.

2.1. Holistic representation methods

Global representations represent the entire visual observation. Global representations are created top-down, starting with the localization of an individual through background removal or tracking. The image descriptor is then produced by encoding the region of interest as a whole. Since they encode most of the information, the representations are effective. But they depend on precise localization, backdrop removal, or tracking. Additionally, they are more sensitive to occlusions, noise, and

DOI: 10.54254/2755-2721/6/20230806

viewpoint. Global representations typically perform well if the domain can provide good control over these variables [3].

There are several classical global representation methods, such as the MEI and MHI [4], HOG [5], 3D HOG [6] and STV [7]. MEI and MHI are appearance-based methods, using silhouettes and contours of the human conducting the action, while HOG and 3D HOG are based om gradients, and STV namely spatio-temporal volume provide another method.

The MEI (Motion Energy Image) and MHI (Motion History Image), which Bobbick and Davis [4] suggested, are one of the most influential methods. It is aimed at recognizing the motion itself directly. Where motion has happened in a series of images is represented by a MEI (motion-energy image). The intensity of a MHI (motion-history image), which has scalar-values, depends on how recently a motion occurred. The MHI (motion-history image) can depict how the image is moving rather than where it is moving.

Dalal and Triggs [5] proposed HOG (Histograms of Oriented Gradients) descriptor. The main premise is that, even in the absence of precise knowledge of the edge directions or associated gradient, the distribution of local intensity gradients or edge orientations may typically provide a good description of the appearance and shape of local objects. In addition to HOG, 3D HOG proposed by Klaser et al. [6] is also a gradient based method. They presented a 3D HOG feature to depict actions, 3D HOG expanded the gradient histogram (HOG) characteristic of static images to the space-time dimension. This approach directly quantifies in the gradient direction, which has strong robustness and low computational cost.

Yilmaz and Shah [7] suggest that activities be identified based on the Spatio-Temporal Volume's differential features (STV). The steps of this approach are as follow: (1) Generate the STV. (2) Obtain action descriptors from STV. (3) Perform action recognition.

2.2. Local representation methods

Local representations use a collection of independent patches to describe the observation. In order to calculate local representations, local patches are first calculated around spatiotemporal interest locations that are first identified. A final representation is created by combining the patches. After the bag-of-feature techniques' first success, correlations between patches are now receiving greater attention. Local representations do not technically need background monitoring or subtraction because they are much more resistant to noise and partial blockage. Nevertheless, local representations require sufficient pertinent interest points to be extracted. Therefore, they may need pre-processing occasionally [3].

2.2.1. Space-time interest point detectors. The Space-time interest point based method uses detectors to detect interest points in video. The most dramatic areas in the spatio-temporal dimension of video are called interest points. Extracting relevant regional features based on interest points is a bottom-up local research method.

The idea of spatial interest points was initially applied to the space-time domain by Laptev [8], proposed Harris 3D spatiotemporal interest points, and showed that the local spatiotemporal features obtained often related to interesting events in video data, and these feature types are useful for sparse coding, which was obtained using standard optimization methods, without strict manual initialization or tracking.

In their review and comparison of methods using local spatiotemporal features and visual dictionary interpretation, Peng et al. [9] proposed a straightforward and efficient representation called a hybrid supervector that outperforms other conventional BoVW (Bag of visual words model) pipelines and performs well on a variety of action datasets.

2.2.2. Local descriptors. With the human motion, a trajectory will be generated. The trajectory based method uses the key points of the human skeleton or the motion trajectory of the joints to represent the action.

DOI: 10.54254/2755-2721/6/20230806

The DT (dense trajectories) algorithm is first proposed by Wang et al. [10, 11]. And on the basis of this DT algorithm, they then proposed an improved iDT (improved dense trajectories) algorithm.

A representation of a video that uses motion boundary descriptors and dense trajectories is presented by Wang et al. [10]. Trajectories are responsible for capturing video's local motion information. DT algorithm adopts grid division to intensively sample feature points on various image sizes, and then, in the sampled feature point field, calculates the optical flow median to determine the trajectory of feature points. After obtaining a large number of track features, DT algorithm uses SVM classifier to classify the video.

The iDT [11] algorithm and the DT algorithm have the same basic framework, but the optimization of optical flow image, regularization method and feature encoding method are improved and optimized. This method can basically overcome the change of camera angle and analyze the local motion of people.

3. Deep learning based representation methods

Deep learning has been developed vigorously in recent years, and is gradually introduced into video action recognition tasks. The deep learning-based approach automatically learns the trainable feature in video, while the handmade representation method need to construct the feature manually [2]. Video data contains both spatial information and temporal information. How to extract and classify the spatio-temporal features of video is the key point of designing video action recognition algorithm based on deep learning.

As shown in Figure 3, the neural network receives the video data as input, and when the data passing through the network layer, the pattern in the video will be recognized and features will be created. Neural networks automatically extract useful features from video data sets. Generally speaking, the deeper the neural network is, the richer the features can be learned. The last layer is usually used as a classifier to output classification labels.

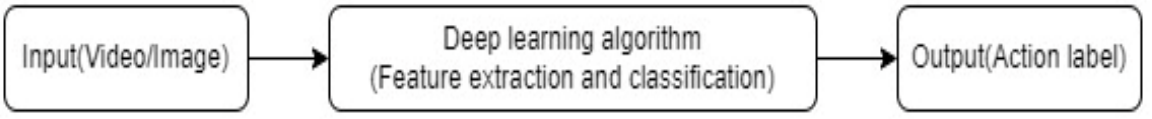

Figure 3. Video action recognition flow based on deep learning method.

This paper divides the deep learning representation methods into 3 categories, i.e., Two-Stream ConvNets, 3D CNN and Transformer based methods.

3.1. Two-Stream ConvNets

RGB and optical flow are the two inputs used in Two-Stream CNNs [12] or its derivatives with the purpose of representing appearance and motion information in late-fusion videos separately.

In 2014, Simonyan et al. [12] first proposed Two-Stream Convolutional Network. The whole network consists of two convolutional neural networks. Spatial stream ConvNet uses a single frame RGB image as its input, primarily to extract the spatial information of the image. Optical flow ConvNet takes the optical flow image between consecutive frames as the input, which is useful for extracting the video's time information, as shown in Figure 4 [12].

DOI: 10.54254/2755-2721/6/20230806

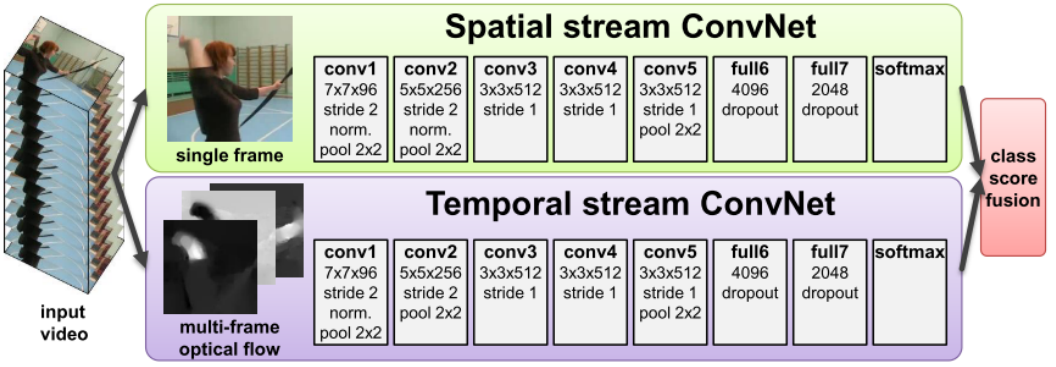

Figure 4. Architecture of Two-stream CNNs. [12]

Simonyan et al. [12] also proved that temporal and spatial recognition streams are complementary, the accuracy of either the spatial network or the temporal network is higher than that of the most advanced deep architectures, and the combination of the two networks can further improve the accuracy.

In 2015, Wang et al. [13] improved the Two-Stream model. They realized that the concept of human behavior is much more complex than objects, and it often involves high-level visual content such as interactive objects, human posture, scene context, etc., so a more complex model is needed to model complex video features. They chose GoogLeNet and VGGNet to build a very deep Two-Stream ConvNets, and then designed some training strategies to solve the overfitting problem caused by small datasets: (1) Pre-training for both spatial and temporal nets; (2) Smaller learning rates; (3) More data augmentation techniques; (4) High dropout ratio; (5) Multi GPU parallel training improves computing efficiency and reduces memory consumption.

Wang et al. [14] further improved the Two-Stream ConvNets model in 2016 based on the research results of [13]. They provide a brand-new framework for video-based action identification called the temporal segment network (TSN), which is based on the notion of long-range temporal structure modeling. This framework utilizes a sparse sampling technique to extract short snippets from a lengthy video clip, with samples distributed equally along the temporal dimension. The Inception with Batch Normalization (BN-Inception) method is used to construct the architecture of the TSN network in order to achieve a suitable balance between accuracy and efficiency.

Feichtenhofer et al. [15] conducted in-depth research on the fusion of spatial network and temporal network, and found that the best fusion location is in the last convolution layer. After fusion, using 3D pooling instead of 2D pooling can further improve the performance.

The above TSN [14] uses late fusion on a 2D CNN-based baseline for long-range temporal modeling. In recent years, some 2D CNN based methods which have temporal modules for all stages are proposed, e.g., R(2+1)D [16], TSM [17] and TEINet [18].

3.2. 3D CNN

Three-dimensional convolution and pooling are proposed by 3D-CNNs [19, 20] to directly learn spatiotemporal features from videos.

3.2.1. Pure 3D-CNN model and some variants of 3D-CNNs. 3D convolution was first proposed by Ji et al. [19]. Its idea is to expand the one-dimensional depth channel on the original 2D convolution kernel operator. This depth channel can be represented as a continuous frame on the video or different slices in the stereo image. Its purpose is to realize the synchronous computation of spatial and temporal characteristics of convolutional neural networks.

DOI: 10.54254/2755-2721/6/20230806

In the early days, Tran et al. [20] presented a straightforward and efficient method, which adopts deep 3D ConvNets that were trained on a supervised large scale video dataset to learn spatiotemporal features. They took the lead in exploring the best 3D convolution kernel size that most conforms to the video feature extraction through experiments. At that time, they carried out many experiments on the UCF-101 dataset and set up four kinds of 3D convolution kernels with fixed size. Finally, it was demonstrated that one of the top performing architectures for 3D ConvNets is a homogenous architecture in which all layers contains a small $3 \times 3 \times 3$ convolution kernel. The classical C3D convolution neural network is built using the convolution kernel of this size. A novel Two-Stream Inflated 3D ConvNet (I3D) (pure 3D-CNN model), based on 2D ConvNet inflation, was introduced in 2017 by Carreira et al. [21]. The design principle of the I3D network is similar to that of C3D, which is to expand the time sequence dimension on the original 2D convolution kernel. However, unlike C3D, although I3D is composed of 3D convolutions, it does not need to be trained from scratch. Instead, it directly loads the weight parameters of the Inception network on the image classification dataset ImageNet and fine tunes them. Compared with C3D, I3D has a better parameter initialization mode and a more efficient training process.

There are also some variants of 3D CNNs. For instance, Wang X et al. [22] presented Non-local I3D in 2018, which uses non-local procedures to capture long-range dependencies. Additionally, any current architectures can be merged with these non-local blocks.

3.2.2. CNN+LSTM. The method of action recognition based on Long Short-Term Memory Network is to add a recursive layer such as LSTM to the CNN (convolutional neural network) to construct a hybrid network.

Long-term recurrent convolutional networks (LRCNs), which combines convolutional layers and long-range temporal recursion, is proposed by Donahue et al. [23]. It is an innovative architecture aims at visual recognition and description and is end-to-end trainable. As shown in Figure 5 [23], LRCN takes variable-length input, then uses a CNN to process the input, CNN's outputs are fed into LSTMs, and finally generate a variable-length prediction.

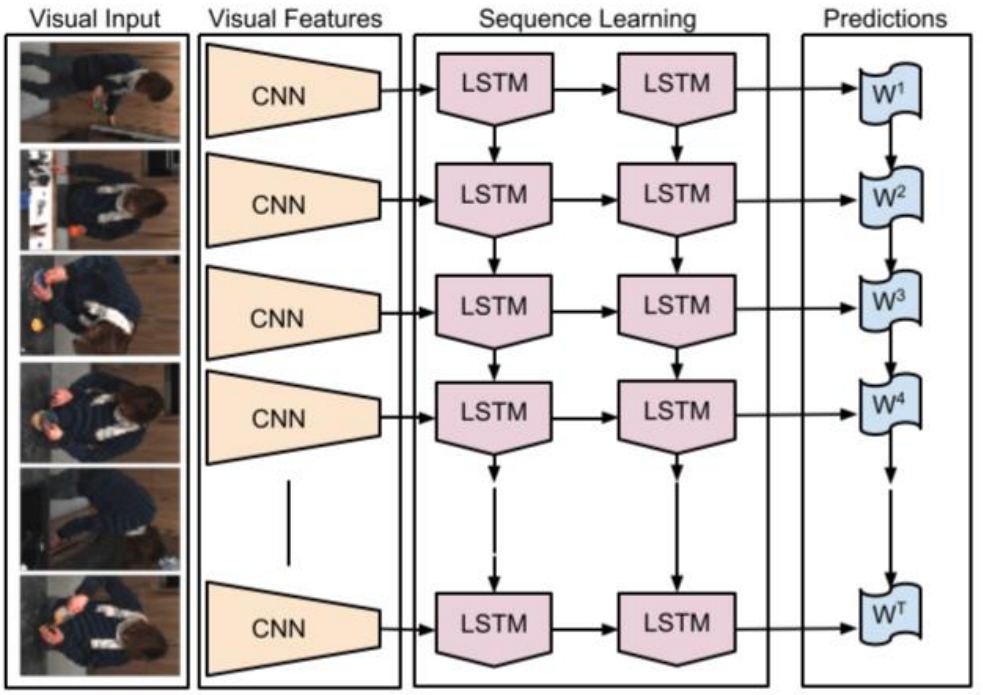

Figure 5. Architecture of LRCNs. [23]

DOI: 10.54254/2755-2721/6/20230806

Approach in [24] is similar to that in [23], which is connecting LSTM cells to the output of CNN. Yue-Hei et al. [24] proposed two video-classification methods that can combine CNN outputs at the frame level to get video-level predictions, that is, Feature Pooling methods and LSTM. LSTM's hidden state evolves with each subsequent frame.

3.2.3. Combination of 2D-CNN and 3D-CNN. In recent years, in addition to the separate 2D convolution and 3D convolution neural networks, the 2D+3D fusion structure has also been widely proposed.

Since activity understanding involves long-term activities that may last for several seconds, to attain full accuracy, action detection needs to integrate the long-term context. To capture temporal relationships between frames, several 3D CNN designs have been developed, but due to the high computing cost, it cannot be used to cover the entire video. Consequently, in order to address the above issues, Zolfaghari et al. [25] introduced an Efficient Convolutional Network for Online Video Understanding (ECO). ECO is an end-to-end trainable architecture. It uses a 2D convolutional architecture to process a single frame from a temporal neighborhood effectively so that the appearance features of that frame can be captured. To considerably outperform the belief derived from a single frame, especially for complicated long-term operations, a 3D network is given feature representations of distant frames and learns the temporal context between these frames.

In [25], input video will go through a 2D Net first, and then a 3D Net. However, Sudhakaran et al. [26] proposed a different approach in 2020, that is, spatial gating in spatial-temporal decomposition of 3D kernels using Gate-Shift Module (GSM). The network contains several branches, and the Gate-Shift Module can be inserted into these branches. A 2D-CNN may train to mix and route features across time adaptively with essentially no additional parameters or processing overhead when GSM is connected.

3.2.4. More recent 3D-CNN approaches. In 2019, Feichtenhofer et al. [27] proposed SlowFast networks for video recognition. This model contains a Slow pathway and a Fast pathway. Aiming at capturing spatial semantics, the slow pathway runs at a low frame rate. High frame rates are used by fast pathways to capture motion at fine temporal resolution.

In 2020, Li et al. [28] proposed a SmallBig network. In the SmallBig network, small view and big view work together to capture contextual semantics while learning the core semantics using the small view branch.

Feichtenhofer C [29] proposed X3D, a family of efficient video networks, which takes the residual network as the backbone, and expands along six dimensions, namely, frequency, time dimension, space dimension, number of residual module layers, number of output channels of each convolution layer, and number of output channels of each residual module, to explore the impact of different dimensions on recognition accuracy. The outcomes demonstrate that each expansion helps to enhance performance, and the extended residual module's number of output channels is the best.

3.3. Transfomer

3.3.1. Transfomer. In 2017, Vaswani et al. [30] suggested a straightforward network architecture called the Transformer that completely does away with recurrence and convolutions in favor of attention mechanisms. As depicted in Figure 6 [30], the Transformer uses stacked self-attention and point-wise, fully connected layers for both the encoder and decoder.

DOI: 10.54254/2755-2721/6/20230806

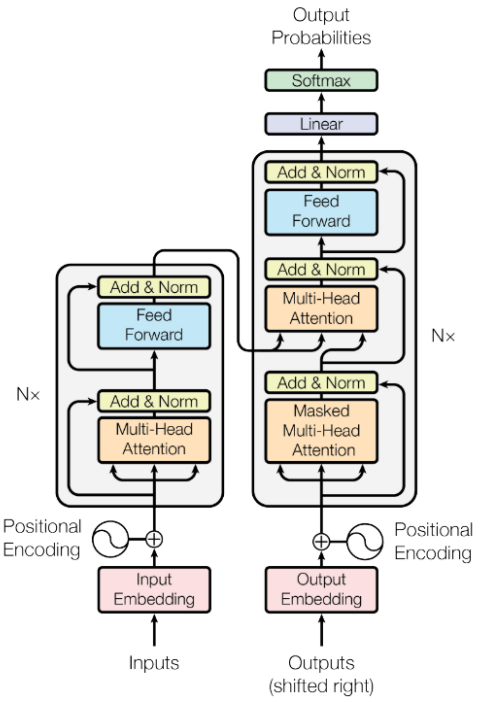

Figure 6. Architecture of the Transformer. [30]

3.3.2. Transformer + CNN. CNN convolutional neural network can effectively reduce local redundancy because of its excellent local receptive field. At the same time, the problem is that it is difficult to capture global dependencies. The Transformer in natural language processing has been widely concerned by researchers with its self-attention mechanism, and can well capture the global dependency information. Therefore, some researchers proposed to combine Transformer [30] and CNN in image recognition, video understanding and other fields, and achieved good results.

Girdhar et al. [31] presented Action Transformer, a novel video action recognition network that classifies a person of interest's actions using a Transformer architecture that's been altered as its "head". By this means, Action Transformer can also determine and utilize the contextual information (other persons, other objects). The model is made up of separate base and head networks (as shown in Figure 7 [31]).

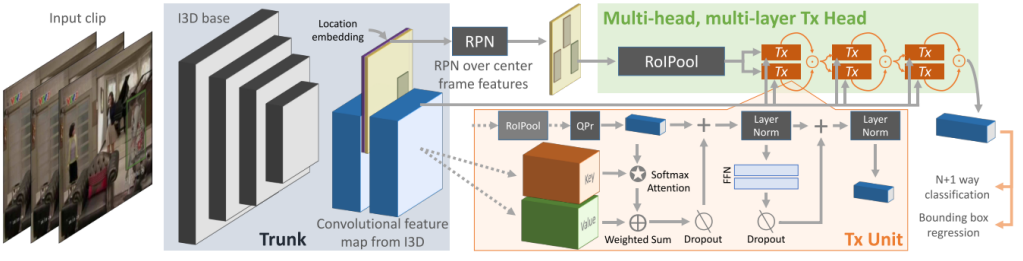

Figure 7. Base network architecture of Action Transformer [31].

In 2022, Liu Z et al. [32] proposed a model called Video Swin Transformer. Video Swin Transformer is implemented using a general-purpose vision backbone for image understanding, a spatiotemporal adaptation of Swin Transformer [33].

In conclusion, we put the accuracy of some important video action recognition methods in Table1 and Table 2 (based on different datasets).

DOI: 10.54254/2755-2721/6/20230806

Table 1. Summary of accuracy of main video action recognition methods mentioned in the paper.

		Accuracy			
Method	Year	UCF-101	HMDB-51	Hollywood2	YouTube
Two-stream model (fusion by averaging) [12]	2014	86.9%	58.0%	-	-
Two-stream model (fusion by SVM) [12]	2014	88.0%	59.4%	-	-
Very deep Two-Stream ConvNets [13]	2015	91.4%	-	-	-
Temporal segment network (TSN) (3 modalities) [14]	2016	94.2%	69.4%	-	-
Spatiotemporal fusion Two-stream networks (S:VGG-16, T:VGG-16) [15]	2016	92.5%	65.4%	-	-
Spatiotemporal fusion Two-stream networks + IDT (S:VGG-16, T:VGG-16) [15]	2016	93.5%	69.2%	-	-
R(2+1)D-TwoStream (pretrained on Kinetics) [16]	2018	97.3%	78.7%	-	-
TSM [17]	2019	95.9%	73.5%	-	-
TEINet [18]	2020	96.7%	72.1%	-	-
C3D [20]	2015	85.2%	-	-	-
C3D + IDT [20]	2015	90.4%	-	-	-
I3D (RGB + flow) [21]	2017	93.4%	66.4%	-	-
Long-term recurrent ConvNet (LRCNs) [23]	2015	82.9%	-	-	-
ECO [25]	2018	94.8%	72.4%	-	-
Transformations [1]	2016	92.4%	62.0%	-	-

Table 2. Summary of accuracy of main video action recognition methods mentioned in the paper.

		Accuracy			
Method	Year	Something-Something V1	Something-Something V2	Kinetics	Mini-Kinetics
Non-local I3D [22]	2018	-	-	83.8%	-
GSM [26]	2020	55.16%	-		-
SlowFast [27]	2019	-	-	79.8%	-
SmallBig [28]	2020	-	-	78.7%	-
X3D [29]	2020	-	-	79.1%	-

DOI: 10.54254/2755-2721/6/20230806

4. Video motion recognition datasets

4.1. The UCF101 dataset

The University of Central Florida created the massive UCF-101 database [34], which contains videos in 101 different categories. 13,320 distinct videos from the YouTube site are included. It mainly includes five categories of actions: interaction between people and objects, simple body actions, interaction between people, playing musical instruments, and sports.

4.2. The HMDB dataset

HMDB [35] dataset is an important dataset for action recognition. Most of the videos in this dataset are from movies, and some are from Google videos, YouTube etc. The HMDB dataset contains a total of 6,766 video clips, divided into 51 distinct action categories, each containing at least 101 clips. The actions are divided into 5 categories: general body movements, body movements with object interaction, general facial actions, facial actions with object manipulation and body movements for human interaction.

4.3. The Hollywood2 dataset

The IRISA institute created two datasets: Hollywood1 and Hollywood2 [36], to handle real world issues. Hollywood2 contains 3669 samples from 69 Hollywood films, divided into 12 action categories and 10 scenes. It contains 12 different types of action: picking up the phone, operating a vehicle, eating, engaging in conflict, exiting a vehicle, shaking hands, hugging, and kissing, as well as jogging, sitting down, sitting up, and standing up.

4.4. The YouTube dataset

YouTube [37] includes 11 actions with 1168 videos. The 11 action disciplines are walking a dog, volleyball spiking, trampoline leaping, tennis swinging, swinging, soccer juggling, horseback riding, golf swinging, diving, biking/cycling and basketball shooting.

4.5. The Kinetics dataset

The three distinct datasets that make up the Kinetics dataset [38] are Kinetics400, Kinetics600, Kinetics700. These three datasets can be distinguished from one another by the number of videos they include. The dataset contains 400 human action classes, each action has at least 400 video clips. Each 10-second clip comes from a separate YouTube video and lasts about that long. The list of action classes includes: Person Actions, Person-Person Actions and Person-Object Actions. Some actions require temporal reasoning to distinguish since they are fine grained, such as different types of swimming. Other actions, such as playing various wind instrument kinds, call for greater focus on the object to be distinguished.

4.6. The Something-Something dataset

Something-Something [39] contains 174 different types of videos, which are predefined human-object interactions with everyday objects. It contains 108, 499 videos (version 2 contains 220, 847) across 174 labels, with duration ranging from 2 to 6 seconds.

4.7. The Charades dataset

The Charades dataset [40] is collected through Amazon Mechanical Turk, which collects the daily leisure activities of hundreds of people. The dataset includes 9848 annotated videos, each lasting 30 seconds on average. Charades provides a total of 27847 video descriptions, 66500 temporal localization intervals for 157 action classes, and 41104 tags for 46 object classes.

In conclusion, we compared the common datasets for video action recognition in Table 3.

DOI: 10.54254/2755-2721/6/20230806

Table 3. Summary of common datasets.

Datasets	Number of videos	Number of categories
UCF-101 [34]	13320	101
HMDB-51 [35]	7000	51
Hollywood2 [36]	3669	12
YouTube [37]	1168	11
Kinetics [38]	650000	400/600/700
Something-Something V2 [39]	220847	174
Charades [40]	9848	157

5. Conclusion

In this paper, we make a comprehensive arrangement of video action recognition algorithms based on RGB data, including traditional methods and deep learning based methods. This paper focuses on the collation of RGB data based algorithms, and in the future, we plan to focus on the comparative analysis of Kinect-Based methods, including Skeleton-Based Action Recognition, Depth-Based Action Recognition and Action Recognition via a Combination of Skeleton and Depth Features.

References

[1] Wang X, Farhadi A, Gupta A. Actions~ transformations[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 2658-2667.

[2] Yao G, Lei T, Zhong J. A review of convolutional-neural-network-based action recognition[J]. Pattern Recognition Letters, 2019, 118: 14-22.

[3] Poppe R. A survey on vision-based human action recognition[J]. Image and vision computing, 2010, 28(6): 976-990.

[4] Bobick A F, Davis J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23(3): 257-267.

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[6] Klaser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3d-gradients[C]//BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008: 275: 1-10.

[7] Yilmaz A, Shah M. Actions sketch: A novel action representation[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 984-989.

[8] Laptev I. On space-time interest points[J]. International journal of computer vision, 2005, 64(2): 107-123.

[9] Peng X, Wang L, Wang X, et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice[J]. Computer Vision and Image Understanding, 2016, 150: 109-125.

[10] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.

[11] Wang H, Schmid C. Action recognition with improved trajectories[C]//Proceedings of the IEEE international conference on computer vision. 2013: 3551-3558.

[12] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27.

[13] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets[J]. arXiv preprint arXiv:1507.02159, 2015.

[14] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, Cham, 2016: 20-36.

DOI: 10.54254/2755-2721/6/20230806

[15] 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. 2016: 1933-1941.
[16] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.
[17] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7083-7093.
[18] Liu Z, Luo D, Wang Y, et al. Teinet: Towards an efficient architecture for video recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11669-11676.
[19] 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, 2012, 35(1): 221-231.
[20] 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. 2015: 4489-4497.
[21] 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. 2017: 6299-6308.
[22] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.
[23] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.
[24] 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. 2015: 4694-4702.
[25] Zolfaghari M, Singh K, Brox T. Eco: Efficient convolutional network for online video understanding[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 695-712.
[26] Sudhakaran S, Escalera S, Lanz O. Gate-shift networks for video action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1102-1111.
[27] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.
[28] Li X, Wang Y, Zhou Z, et al. Smallbignet: Integrating core and contextual views for video classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1092-1101.
[29] Feichtenhofer C. X3d: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 203-213.
[30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[31] Girdhar R, Carreira J, Doersch C, et al. Video action transformer network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 244-253.
[32] Liu Z, Ning J, Cao Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3202-3211.
[33] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.
[34] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos

DOI: 10.54254/2755-2721/6/20230806

in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.
[35] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[C]//2011 International conference on computer vision. IEEE, 2011: 2556-2563.
[36] Marszalek M, Laptev I, Schmid C. Actions in context[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 2929-2936.
[37] Liu J, Luo J, Shah M. Recognizing realistic actions from videos “in the wild”[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 1996-2003.
[38] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.
[39] Goyal R, Ebrahimi Kahou S, Michalski V, et al. The" something something" video database for learning and evaluating visual common sense[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5842-5850.
[40] Liu C, Hu Y, Li Y, et al. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding[J]. arXiv preprint arXiv:1703.07475, 2017.