Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042494

REVIEW

Action Recognition and Detection Based on Deep Learning: A Comprehensive Summary

Yong Li[1,4], Qiming Liang[2,*], Bo Gan[3] and Xiaolong Cui[4]

[1]College of Information Engineering, Engineering University of PAP, Xi'an, 710086, China

[2]PAP of Heilongjiang Province, Heihe Detachment, Heihe, 164300, China

[3]National Key Laboratory of Science and Technology on Electromagnetic Energy, Naval University of Engineering, Wuhan, 430033, China

[4]Joint Laboratory of Counter Terrorism Command and Information Engineering, Engineering University of PAP, Xi'an, 710086, China

*Corresponding Author: Qiming Liang. Email: liangqiming96@163.com

Received: 01 June 2023 Accepted: 30 August 2023 Published: 31 October 2023

ABSTRACT

Action recognition and detection is an important research topic in computer vision, which can be divided into action recognition and action detection. At present, the distinction between action recognition and action detection is not clear, and the relevant reviews are not comprehensive. Thus, this paper summarized the action recognition and detection methods and datasets based on deep learning to accurately present the research status in this field. Firstly, according to the way that temporal and spatial features are extracted from the model, the commonly used models of action recognition are divided into the two stream models, the temporal models, the spatiotemporal models and the transformer models according to the architecture. And this paper briefly analyzes the characteristics of the four models and introduces the accuracy of various algorithms in common data sets. Then, from the perspective of tasks to be completed, action detection is further divided into temporal action detection and spatiotemporal action detection, and commonly used datasets are introduced. From the perspectives of the two-stage method and one-stage method, various algorithms of temporal action detection are reviewed, and the various algorithms of spatiotemporal action detection are summarized in detail. Finally, the relationship between different parts of action recognition and detection is discussed, the difficulties faced by the current research are summarized in detail, and future development was prospected.

KEYWORDS

Action recognition; action detection; deep learning; convolutional neural networks; dataset

1 Introduction

Being widely used in fields like security, video content review, human-computer interaction, and others, action recognition and detection is among the important research directions in the field of computer vision [1]. The concept differs for the two components of action recognition and detection, which are action recognition and action detection [2]. Action recognition refers to judging the category of human action in a given video footage, while action detection determines the start and end time of

certain actions in the video, and locates the spatial position of figures in the picture, besides classifying the action. More specifically, action detection can be divided into temporal action detection and spatiotemporal action detection. Temporal action detection only determines the start and end time of certain actions, while spatiotemporal action detection needs to further determine the position of the figures in the picture.

Previous reviews of action recognition and detection focus more on the field of action recognition alone and fail to summarize the current research status comprehensively and accurately. A clear explanation of the difference and connection between the concept of action recognition and action detection is not given in some literature. For example, Hassner [3] reviewed the early development of action recognition, focusing on the commonly used datasets for action recognition. Luo et al. [4] reviewed various algorithms commonly used in action recognition from the perspective of descriptors. Zhao et al. [5] provided insight into traditional recognition methods and deep learning-based recognition methods from two aspects: input content and network depth. Chai et al. [6] focused on the comparison between action recognition methods based on descriptors and what is based on deep learning, before prospects the development direction of action recognition. Zhang [2] provided a comprehensive summary of the current research status of both action recognition and action detection, but the latest research results are not mentioned. Zhu et al. [7] reviewed the research status of action recognition and detection in detail, but the concept of action recognition isn't distinguished clearly from that of action detection. Sun et al. [8] summarized the current research on action recognition and detection in detail from the perspective of data mode.

In recent years, the transformer model, typified by Vision Transformer (VIT) [9], has made remarkable achievements, which reveals a new trend in the field of action recognition and detection. The existing literature rarely reviews action recognition and action detection side by side, and there are few introductions to transformer-based models. Therefore, based on dividing the action recognition and detection structure, this paper summarizes the action recognition and detection in detail from the perspective of the model structure, and much emphasis is put on the prominent transformer model. Besides, this paper summarizes the various algorithms of action recognition and detection, points out the difficulties faced by current research, and explores the subsequent trends.

2 Action Recognition

As shown in Fig. 1, action recognition can fit into either traditional frameworks or deep learning-based frameworks, which mainly consist of three steps: preprocessing, action expression, and classification [10]. Preprocessing includes serialization of video and extraction of optical flow features. In traditional frameworks, Action expression mainly includes feature extraction and coding, while deep learning frameworks use various deep neural networks to extract features. In traditional frameworks, algorithms such as Support Vector Machine (SVM) and random forests are mainly used for action classification, while the action classification of deep learning frameworks mainly uses Softmax and SVM.

2.1 Action Recognition Datasets

The earliest published dataset for action recognition was Kungliga Tekniska Högskolan (KTH) [11]. In recent years, UCF-101 [12] and HMDB51 [13] have been widely used in action recognition. The KTH dataset contains 6 types of actions completed by 25 people in 4 different scenarios, with a total of 2391 video samples, while UCF-101 contains 13,320 video samples in 101 categories, and HMDB51 includes 6,849 video samples in 51 action categories. With the deepening of research, the scenarios, action categories, and sample sizes covered by UCF-101 and HMDB51 are becoming difficult to meet

the needs of the study, thus they are now being phased out. Table 1 shows the comparison of action recognition algorithms in UCF-101 and HMDB51.

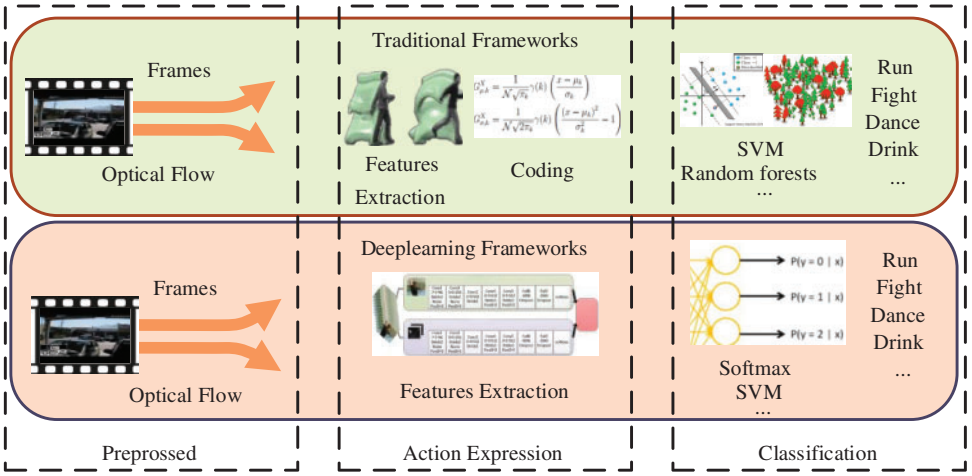

Figure 1: Action recognition flow chart

Table 1: Comparison of action recognition algorithms

Algorithms	UCF-101	HMDB51
C3D [14]	82.30%	51.60%
LRCN [15]	82.92%	–
IDT [16]	85.90%	57.20%
Two-stream [17]	88%	59.40%
P3D [18]	88.60%	–
VideoLSTM [19]	88.90%	56.40%
T3D [20]	90.30%	59.20%
T3D-transfer [20]	91.70%	61.10%
Two-stream fusion [21]	92.50%	65.40%
Hidden two-stream (TSN) [22]	93.20%	66.80%
T3D-TSN [20]	93.20%	63.50%
St-ResNet [23]	93.40%	66.40%
TSN [24]	94.20%	69.40%
St-ResNet + IDT [23]	94.60%	70.30%
Hidden two-stream (I3D) [22]	97.10%	78.70%
I3D [25]	97.90%	80.20%
LGD-3D two-stream [26]	98.20%	80.50%
HAF + BoW/FV halluc [27]	–	82.48%
R(2+1)D-BERT [28]	98.69%	85.1%
TO + MaxExp + IDT [29]	–	87.21%

Table 2: Common datasets for action recognition

Dataset	Year	Category	Highest accuracy
KTH [11]	2004	6	98.83% [30]
Weizmann [31]	2005	9	100% [32]
Hollywood2 [33]	2009	12	78.6% [34]
Olympics sports [35]	2010	16	96.6% [36]
HMDB51 [13]	2011	51	85.1% [28]
UCF-101 [12]	2012	101	98.69% [37]
Kinetics-400 [38]	2017	400	89.9% [37]
Something-Somethingv1 [39]	2017	174	60.9% [40]
Kinetics-600 [41]	2018	600	91.1% [42]
Something-Somethingv2 [39]	2018	174	75.4% [43]
Kinetics-600 [41]	2018	600	91.1% [42]
Kinetics-700 [44]	2019	700	83.4 [45]
RWF-2000 [46]	2020	2	89.3% [47]

In recent years, some research institutions have released datasets with larger sample sizes and richer scenarios. Take the datasets of the Kinetic [38] series and the Something-Something [39] series as an example. The Kinetics400 [38] dataset was released in 2017 and includes a total of 306,245 video samples in 400 categories. DeepMind then expanded the Kinetics400 dataset to release the larger Kinetics600 [41] and Kinetics700 [44]. The Something-Somethingv1 dataset was released at ICCV2017, with 2–6 seconds per video, divided into training sets, test sets, and validation sets according to the ratio of 8:1:1, containing more than 100,000 sample sizes. Something-Somethingv2 was released at CVPR2020, with a further expanded sample size of more than 200,000 and the image format was updated from the previous JPG to Webm.

At present, action recognition datasets devote to expanding around specific scenarios for fine-grained motion analysis to meet various task scenarios. Table 2 shows the basic information of commonly used datasets for action recognition.

2.2 Descriptor-Based Action Recognition

Before deep learning was widely used, action recognition mainly adopted descriptor-based methods. Action recognition methods that are based on descriptors can be divided into global feature-based and local feature-based methods. The global feature extraction method initially uses the direction gradient histogram (DGH), before developing into two methods: contour silhouette and human joint point method. For example, Bobick et al. [48] generated Motion History Images (MHI) through the construction of a two-dimensional motion energy map to achieve action classification. Yang et al. [49] constructed the coordinates of joint points and combines static posture, motion attributes, and overall dynamics for action recognition. Local feature extraction mainly includes two methods: spatiotemporal point of interest sampling and dense trajectory tracking. For example, Willems et al. [50] proposed an action recognition method based on 3D Harris corner point detection, and Wang et al. [16,51] proposed action recognition methods Dense Trajectories (DT) and Improved Dense Trajectories (IDT) based on dense trajectory tracking. Through multi-scale intensive sampling,

these feature points are closely tracked in the temporal dimension to form trajectories, via which the category of action is judged eventually.

2.3 Deep Learning-Based Action Recognition

When judging the category of action, humans usually need to distinguish both the static information of the actor and the dynamic information of the change of the actor's movement. Therefore, the implementation of action recognition relies on static spatial information and dynamic temporal information of the action in the video. According to the different network structures that obtain these two types of information, deep learning-based action recognition models can generally be divided into four categories: Two stream models, Temporal models, Spatiotemporal models, and the latest transformer models.

2.3.1 Two Stream Models

To obtain spatial features and temporal features, the two-stream model uses two parallel pathways to extract spatial and temporal features, respectively. It realizes the fusion of spatiotemporal feature information through appropriate feature fusion methods before finally realizing the classification of action. Such an idea was first proposed in 2014 by Simonyan et al. [17]. He inputs preprocessed video frames and optical flow maps to two parallel paths and then uses AlexNet [52] for feature extraction on both paths, where static spatial features are obtained from video frames, and dynamic temporal features are obtained from optical flow maps. The feature information is fused at the end of the channel to achieve action classification.

Feichtenhofer et al. [21] improved the feature fusion method on this basis. Feature fusion is performed in advance in the convolutional layer of either channel, and additional fusion is performed on the quasi-prediction layer to replace the previous end fusion method. The stated fusion method not only reduces the parameters of the model but also improves the accuracy of recognition. Then, Feichtenhofer et al. [53] introduced He et al.'s residual network (ResNet) [54] to the two stream models, introducing residual connections between the two-Stream architectures to enhance the spatiotemporal feature interaction and fusion between the two streams.

Similarly, focusing on the fusion of two stream features, Ng et al. [55] introduced the Long Short Term Memory network (LSTM) [56] based on the two stream neural networks. LSTM is used to fuse the output of two stream CNNs, effectively expressing the sequence of frames before and after through the memory unit of LSTM, strengthening the feature extraction ability of temporal information, and realizing the recognition of actions in long videos.

To achieve feature extraction of long-term video information, Wang et al. [24] used a sparse time sampling strategy to sample multiple video clips from the entire video at the input, giving the preliminary judgment result of the action category in each segment, and then combines the results of multiple fragments to carry out "consensus" to realize the classification of action, thereby realizing the recognition of long-range actions.

The two-stream model pre-processes video frames and optical flow maps at the input, which requires much time and computing power at the pre-processing stage of the data, making the model far from being able to achieve end-to-end recognition. For the above reasons, Zhu et al. [21] established a network structure called MotionNet based on the two-stream model, which can directly model the temporal features of video frames, replacing the role of the optical flow map.

Feichtenhofer et al. [57] made great improvements to the two-stream model and built a lightweight two-stream recognition network named SlowFast. SlowFast captures spatial semantics for slow channels operating at low frame rates, and fast channels operating at high frame rates, thereby capturing temporal information with fine temporal resolution. Finally, horizontal connections are used to fuse from the fast path to the slow path to achieve action classification.

Compared with Temporal models, Spatiotemporal models, and the latest transformer models., the Two stream models are more complex and the model training is cumbersome, hence is difficult to truly achieve end-to-end recognition. The idea of two-stream models, however, provides important inspiration for algorithm innovation in the field of action detection and promotes the development of action detection. The two-stream model is a compromise model made in the architecture at the beginning stage of the development of the action recognition algorithm, and some algorithms even need to train two pathways step by step during the model training process, which is very time-consuming. From the architecture design of the two-stream model, it can be clearly seen that the extraction of action features is difficult.

2.3.2 Temporal Models

To obtain spatial features and temporal features, the temporal model adopts a cascading method, in which the spatial semantic information is first extracted by the convolutional neural network (CNN), and the temporal feature information is then extracted by the recurrent neural network (RNN). Simple RNNs produce gradient divergence or vanishing gradients when processing long-term feature information, so the actual temporal model uses LSTM with a forgetting gate.

In the study of Donahue et al. [15], AlexNet and LSTM are used to cascade, and spatial and temporal features are modeled respectively before being classified by the fully connected layer at the end to construct Long-term recurrent convolutional networks (LRCN) for action recognition. To better represent the relationship in space and eliminate redundant information, Sudhakaran et al. [58] introduced ConvLSTM [59] to replace the traditional LSTM for violent scenes, which realized the fusion of spatiotemporal information and further improved the accuracy of recognition.

Li et al. [19] introduced the attention model into the LSTM network and constructed a new action recognition model VideoLSTM through the fusion of ConvLSTM and Attention LSTM. VideoLSTM introduces motion features and attention mechanisms to spatiotemporal positions, focusing on preserving spatial feature information between video frames. Wang et al. [60] combined I3D with LSTM based on the I3D network, and modeled the high-level temporal features obtained by the I3D model through LSTM.

When the CNN comes too complex, the constructed spatial feature map will be abstracted, resulting in the loss of temporal feature information, and the limitation of LSTM's ability to process temporal information. That accounts for the loss of the popularity of current temporal mode-based action recognition research. Since RNN networks and their variants cannot implement multi-GPU parallel computing, temporal models cannot achieve parallel training of models on multiple devices. Action recognition algorithms have high hardware requirements in the model training process, so the inability to build action recognition models by multi-device parallel training will bring great trouble to researchers, which is another important factor in the development bottleneck of current temporal models.

2.3.3 Spatiotemporal Models

Spatiotemporal models design an integrated structure and obtain space-time feature information at the same time. A spatiotemporal model usually uses 3D convolution, and the data with different time dimensions is performed on 3D convolution operation. In recent years, some scholars have proposed adopting a specially designed data processing method in the model to fuse the spatiotemporal feature information in advance, and then use 2D convolution to model it, which can also realize action recognition.

The use of 3D convolution for action recognition was first proposed by Ji et al. [61]. Du et al. [14] further extended 3D convolution to the pooling process to form 3D pooling and established the C3D (Convolutional 3D) model. Diba et al. [20] adopted the method of transfer learning in the model construction process and proposed a new temporal transition layer (TTL), which embeds TTL into DenseNet [62] that extends to a 3D structure, thereby constructing a new network Temporal 3D ConvNets (T3D).

3D convolution is very computationally intensive in the computation process, to alleviate this problem, Qiu et al. [18] formed a new convolutional block Pseudo-3D (P3D) by separating the convolution based on the ResNet. Based on the P3D structure, the new action recognition model P3D was successfully constructed. Final experiments show that P3D ResNet has significantly improved action recognition performance.

Transfer learning has a wide application in deep learning, which makes it easier to train new models. For example, in the field of object recognition, transfer learning via a trained model obtained by ImageNet can accelerate the convergence speed of a new model. A similar approach can be used in the field of action recognition to reduce the workload of training. To be able to use pre-trained models on a 3D convolutional network, Carreira et al. [25] expanded the two-dimensional convolution kernel and pooling kernel to 3D based on Inception-v1 [63]. They then pre-trained the three-dimensional model implicitly on ImageNet before obtaining the pre-trained model of 3D convolution in Kinetics. After pre-training, the Inflated 3D (I3D) model obtained by transfer learning gains a great improvement in the accuracy of action recognition, which also greatly reduces the difficulty of model training.

To make the 3D convolution model more lightweight, scholars have proposed many innovative methods, but it is more than difficult for 3D convolution is difficult to outperform 2D convolution. In 2019, Lin et al. [64] creatively performed the migration splicing of the feature map in the temporal dimension and proposed the Temporal Shift Module (TSM) by processing the temporal features before feature extraction. TSM fuses invisible temporal information into spatial features, and only 2D convolution can achieve the effect of 3D convolution, which alleviates computational overhead by sacrificing storage. Then, Shao et al. [65] proposed a new deformation displacement module, a temporal interlacing network (TIN), based on TSM, which further strengthened the fusion of spatiotemporal information. Fan et al. [66] proposed a learnable 3D shift network RubkisNet, which simultaneously migrates in both the spatial and temporal dimensions, and dynamically generates the proportion of the migration part. RubkisNet obtained a larger range of spatiotemporal information as well as higher accuracy.

In addition, Li et al. [67] extracted adjacent frame information and multi-frame global information by establishing a temporal excitation and aggregation block (TEA). Both short-time motion and long-term feature aggregation were considered, which effectively reduced the complexity of the network and also effectively avoided the drawbacks of 3D CNNs.

Before 2019, 3D convolution was mainly used in spatiotemporal models. Then through the ingenious design of data preprocessing, 2D convolution can also achieve accurate action classification. By sacrificing part of the storage, the computational overhead is greatly alleviated, and action recognition based on the spatiotemporal model has come to the prominent direction of current research.

2.3.4 Transformer Models

Transformer is an attention-based codec-like model that originated in the field of natural language processing (NLP) and has begun to achieve high accuracy in applications of computer vision after the release of the VIT model in 2021. The transformer model is a commonly used decoder architecture in the field of NLP, which has advantages in extracting "contextual" correlation information. It is now shining in the field of computer vision, including action recognition, and is becoming an important cross-modal architecture.

The transformer was first used in the field of action recognition after the release of Video Vision Transformer (VIVIT) [68] in 2021. Similar to VIVIT, Ullah et al. [69] completely abandoned CNNs based on VIT and adopt the attention structure to achieve action recognition. To alleviate the redundancy of the temporal information, Patrick et al. [70] introduced trajectory information based on the transformer, and obtain high accuracy in multiple datasets.

Truong et al. propose an end-to-end transformer structure, Direcformer [71]. The structure introduces ordinal time learning into the transformer, which helps to understand the chronological order of actions. To strengthen its ability to model different spatiotemporal spaces, Google proposed Multiview Transformers for Video Recognition (MTV) [72], which consists of multiple independent encoders to represent different dimensional views of the input video. MTV also fuses the information between the views through horizontal connections. Self-Supervised Video Transformer (SVT) [73] is a new self-supervised approach that trains a teacher-student model using similarity goals that are represented along spatiotemporal dimensions by spatiotemporal attention matching.

Although the transformer achieves very high accuracy in the field of computer vision, including action recognition, it incurs huge computational overhead, which places a burden on research institutions or researchers with average research conditions. Therefore, Recurrent Vision Transformer (RViT) [74] introduces a loop mechanism and integrates Attention Gate to establish a connection between the current frame and the previous hidden state, thereby extracting the global space-time features between frames, and alleviating the problem of insufficient computing power to a certain extent.

3 Action Detection

3.1 Action Detection Datasets

The datasets commonly used for temporal action detection are mainly THUMOS14 [75], MEX-action2 [76], and ActivityNet [77]. The THUMOS14 dataset includes an action recognition part and a temporal action detection part. The action recognition section includes all the categories covered by the UCF-101 dataset. The temporal action detection section includes 20 categories, divided into the training set, validation set, background fragment set, and test set. The MEXaction2 dataset includes two categories: horseback riding and bullfighting. The background fragment length of the MEXaction2 dataset is relatively long, while the proportion of labeled action fragments is low, which makes it more challenging for temporal action detection. ActivityNet is currently the largest database,

it also contains two tasks: action classification and temporal action detection. ActivityNet has a very large sample size of more than 20,000, covering 200 action categories. It can only be downloaded by writing a script based on the official YouTube link. The above dataset only coarsely labeled the temporal action information, which is easy to cause the problem of unclear temporal action boundaries during an experiment, so ECCV2022 released the carefully labeled FineAction [78] dataset. The FineAction dataset contains nearly 17,000 untrimmed videos and 103,000 fine motor temporal annotations. For all 106 action categories, the category definitions are clearer and the temporal annotations are more accurate.

J-HMDB-21 [79], UCF101-24 [80], and Atomic Visual Actions (AVA) [81] datasets are commonly used as spatiotemporal action detection datasets. J-HMDB-21 is a subset of the HMDB dataset, containing a total of 21 categories and 960 video samples. UCF101-24 is a subset of UCF101, including a total of 24 Action categories and 3207 video samples. Compared to either of the previous datasets, the labels of the AVA dataset are much sparser. The AVA dataset consists of 300 movies, each captured for 15 min and labeled second by second. The newly released MultiSports [82] dataset by ECCV2022 further increases the sample size and includes more complex scenes, which is a large-scale spatiotemporal action detection dataset mainly including basketball, football, gymnastics, and volleyball events.

3.2 Temporal Action Detection

Temporal action detection is different from action recognition, which not only needs to classify the action itself, but also needs to locate the temporal position of the action in the video, specifically to locate the start and end time of certain actions accurately from a long video containing background clips, and to determine the category of the action. Temporal action detection usually requires video data with a long time span during model training, such data is huge, and it takes a lot of time and computing resources in the process of data preprocessing and model training, so temporal action detection is very difficult for research institutions with poor research conditions or weak teams.

Tables 3 and 4 show respectively the accuracy of the commonly used algorithms for temporal action detection in the ActivityNet-1.3 dataset and the THUMOS'14 dataset, where mAP@k represents the Mean Average Precision of a certain algorithm when the intersection and union ratio is equal to k.

Table 3: Comparison of algorithms in ActivityNet-1.3

Algorithms	mAP@0.5	mAP
R-C3D [83]	26.8%	–
G-TAD [84]	50.36%	–
TAL-Net [85]	38.23%	20.22%
3C-Net [86]	37.2%	21.7%
BSN [87]	46.45%	30.03%
P-GCN [88]	48.26%	31.11%
SSN [89]	39.12%	32.26%
BMN [90]	50.07%	36.42%
Actionformer [91]	54.7%	36.6%

(Continued)

Table 3 (continued)

Algorithms	mAP@0.5	mAP
Internvideo [92]	–	39%
PRN + BMN [93]	59.7%	42%

Table 4: Comparison of algorithms in THUMOS'14

Algorithms	mAP@0.5	mAP@0.4
S-CNN [94]	19%	28.7%
CDC [95]	23.3%	29.4%
TURN-FL-16 + S-CNN [96]	25.6%	34.9%
3C-Net [97]	26.6%	34.1%
TAG [96]	28.25%	–
R-C3D [97]	28.9%	35.6%
SSN [89]	29.8%	41%
CBR-TS [98]	31%	41.3%
BMN [90]	32.2%	–
BSN [91]	36.9%	45%
G-TAD [84]	40.2%	–
MGG UNET [99]	37.4%	46.8%
TAL-Net [85]	42.8%	48.5%
P-GCN [88]	49.1%	57.8%
ReAct [100]	57.1%	65%
TadTR [101]	60.1%	69.1%
Actionformer [91]	71.0%	77.8%

3.2.1 Action Detection Based on Descriptors

Traditional temporal action detection methods use descriptors to generate target fragments, thereby achieving the detection of temporal action. For example, Richard et al. [102] identified action types by merging two models, one of which is a length model that combines action duration information and the other is a language model that combines contextual context. Yuan et al. [103] extracted a pyramid of score distribution features (PSDF) based on IDT features. They then used the LSTM network to process the PSDF feature sequence and obtained the prediction of the action fragment according to the output frame-level action category confidence score. By training video, Hou et al. [104] automatically determined the number as well as types of sub-movements in each action. To locate an action, the objective function, which combines the appearance, duration, and time structure of a certain sub-action, is optimized as the shortest path problem in the network flow formula, before the best combination is selected by considering both the sub-action score and the distance between the sub-action.

3.2.2 Deep Learning-Based Action Detection

Another primary approach to temporal action detection is to use deep neural networks. According to whether the target candidate region needs to be extracted independently, the relevant object detection algorithms can be divided into one-stage-based and two-stage-based algorithms. Similar to object detection, temporal action detection algorithms can also be divided into one-stage-based methods and two-stage-based methods according to either the process of feature extraction or whether the temporal candidate region, where independent extraction action occurs, is required.

Two-Stage Method

Inspired by the common object detection algorithm R-CNN, Shou et al. [93] proposed a temporal action detection method based on sliding window Segment-CNN (S-CNN) in 2016. The S-CNN cuts the original video into several clips with different lengths and then sends it to the C3D network through pooling operations to conduct action detection. The flexibility for S-CNN to judge the starting and ending times of certain actions is limited by the sliding window mechanism. In the following year, Shou et al. [94] drew on the ideas of Fully convolutional network (FCN) [105] and introduced convolutional devolutional-De-Convolutional filters based on the C3D algorithm. Frame-level fine-grained temporal action detection is then achieved upon the joint effect of upsampling in the temporal dimension and downsampling in the spatial dimension to achieve.

Also inspired by the object detection algorithm, Xu et al. [83] built Region Convolutional 3D Network (R-C3D) based on the Faster R-CNN [106], which encodes the video stream with a 3D fully convolutional network, generates a time range candidate region that may contain action, and then classifies and fine-tunes the candidate region. Unlike S-CNNs, R-C3D can perform end-to-end detection of actions out of video with arbitrary length.

Subject to similar influence, Chao et al. [85] used a Faster R-CNN-based multiscale framework to improve the calibration of receptive fields, enabling resilience to extreme changes in the duration of actions in certain videos. Then, by constructing a two-stream network, the characteristics of red-green-blue (RGB) and optical flow are fused, and action classification is conducted using the late fusion mechanism.

For temporal action detection in actual scenarios, there may be more than one action fragment contained in the video to be detected, so actions can be judged comprehensively by combining the action categories of multiple proposals. Hence Zeng et al. [88] used Graph Convolutional Networks (GCN) [107] to explore the connections among proposals and constructed a GCN-based temporal action detection framework Proposal-GCN (P-GCN).

Liu et al. [99] proposed to use both coarse-grained and fine-grained features to build an end-to-end network multi-granularity generator (MGG) through two modules, segment proposal producer (SPP) and frame actions producer (FAP), which is for finding action fragments. Gao et al. [108] proposed a relation-aware pyramid (RapNet) based on the pyramid network, which enhanced the global feature information representation and located the different lengths of Action fragments. Lin et al. [109] established a novel Dense Boundary Generator (DBG), which extracts spatiotemporal features like the two-stream model and establishes the action perception completeness regression branch and the time boundary classification branch to realize rapid detection of action.

The two-stage method can achieve high detection accuracy by obtaining proposals of the temporal dimension before classifying the action. However, apart from its low operation speed, the two-stage network model is too complex and requires a lot of computing resources.

One-Stage Method

The conventional approach for one-stage temporal action detection is to use convolutional layer generation proposals for recognition and boundary regression. The Action proposals obtained by this method are then assigned the same receptive fields. However, the temporal length varies for different actions. To solve such a problem, Long et al. [110] proposed Gaussian kernel learning, which expresses temporal information by learning a Gaussian kernel. Piergiovanni et al. [111] proposed a new convolutional layer Temporal Gaussian Mixture (TGM) layer, which also adopts a Gaussian model. It can capture long-distance dependencies in video effectively by using Gaussian kernels to calculate other time point features near the current temporal window. Inspired by the object detection related-algorithm, Lin et al. combined I3D with a borderless target detection algorithm, proposed a boundary learning consistency loss function, and constructed an anchor-free saliency-based action detection method Anchor-Free Saliency-Based Detector (AFSD) based on learning significance boundary features.

In recent years, with the rapid development of transformer, some scholars have also begun to solve the problem of temporal action detection with transformer. Liu et al. [100] embeded the video features and positions extracted by CNNs as input and decode a set of action predictions in parallel through the transformer. By focusing on certain segments in a video adaptively, it extracts the contextual information required to make motion predictions, which greatly simplifies the process of temporal action detection and increases the speed of detection. Shi et al. [112] also proposed a DETR-like temporal action detection method based on the transformer, which proposes attention, action classification enhancement loss, and fragment quality prediction related to IOU attenuation, and analyzes them from three aspects: attention mechanism, training loss, and network reasoning. Zhang et al. [113] used transformer as a basic module to design a minimalist temporal action detection scheme, where feature pyramid and local self-attention mechanism are used to model long-time temporal features, and classification and regression are realized without generating proposals or pre-defined bounding boxes. In addition, Liu et al. [114] also combined transformer to propose an end-to-end temporal action detection scheme, which obtains higher accuracy and a faster detection rate by constructing a medium-resolution benchmark detector.

3.3 Spatiotemporal Action Detection

So-called Spatiotemporal action detection is to determine the temporal and spatial position of certain actors in a video containing background clips and to realize action classification. In other words, spatiotemporal action detection needs to mark the position of the actor in the spatial picture based on temporal action detection. Spatiotemporal action detection can often be divided into multiple stages, including action recognition, target tracking, and object detection. Therefore, spatiotemporal action detection usually requires the construction of multiple network-level models in the process of model construction, which is difficult in model training. Table 5 shows the accuracy of the spatiotemporal action detection algorithms respectively over J-HMDB-21 and UCF101-24 datasets.

Puscas et al. [115] employed a selective search method to produce the initial segmentation of still image-based video frames. This initial recommendation set is pruned and temporarily extended using optical flow and transudative learning.

Table 5: Comparison of spatiotemporal action detection algorithms

Algorithms	J-HMDB-21	UCF101-24
Actionness [69]	39.3%	–
Peng w/o MR [70]	56.9%	64.8%
Peng w/o MR [70]	58.5%	65.7%
ACT [71]	65.7%	69.5%
Faster-RCNN + two-stream I3Dconv [72]	73.3%	76.3%
YOWO (16-frame) [73]	74.4%	87.2%
HIT [116]	83.8%	84.8%

Inspired by the object detection algorithm, Kalogeiton et al. [117] built an Action Tubelet detector (ACT) based on the SSD framework. ACT focuses on temporal features between successive frames, reduces the ambiguity of Action prediction, and improves the accuracy of spatiotemporal localization. Yet Gu et al. [118] used I3D for contextual temporal modeling and Faster R-CNN for end-to-end localization and action classification, which is also derived from the object detection algorithm.

Feichtenhofer et al. [57] used SlowFast for action recognition, Deepsort for object tracking, and YOLO for object detection, and realize action detection through a combination of three algorithms. Inspired by the human visual nervous system, Kpüklü et al. [119] proposed You Only Watch Once (YOWO), a unified architecture for spatiotemporal action detection. The network structure of YOWO is similar to the two-stream model, in which 3D-CNN branches and 2D-CNN branches are used in each parallel stream to extract spatiotemporal feature information, and feature fusion, as well as candidate region definition, is carried out in the end. YOWO uses 3D-ResNet-101 [120] to extract spatiotemporal features and solves classification problems with 3D-CNN branches. To solve the spatial localization problem, DarkNet-19 [121] is eventually used to extract the dual-dimensional features of keyframes.

Based on YOWO, Mo et al. [122] proposed to use of Linknet to introduce a connection between 2D and 3D-convolutional structures. They also use a custom bounding box similar to YOLOv2 to achieve the precise positioning of actors and update the YOWO network to the second version, which effectively reduces the complexity of the model and further improves the accuracy of the network.

The Holistic Interaction Transformer (HIT) [116] network is a comprehensive dual-mode framework based on the transformer, which includes RGB streams and pose streams. Each flow models human, target, and hand interactions. Within each subnetwork, an Intra-Modal Aggregation Module (IMA) is introduced, which merges individual interaction units selectively. The Attentive Fusion Mechanism (AFM) is then used to glue together the features produced by each pattern. Finally, HIT extracts clues from the time context using cache memory to better classify the possible actions.

4 Discussion

Action recognition and action detection have been widely used in practical scenarios. Action recognition can be applied to human-computer interaction, video content review, and other fields, while action detection can be applied to intelligent security, video content positioning, video search, and other fields. Action recognition is the prior work of action detection, and only when the relevant algorithms of action recognition tend to be mature, could action detection have good development.

Fig. 2 shows the important algorithms for action recognition and detection, above the arrow, are the action recognition algorithms, and below are the action detection algorithms. In the field of action recognition, mainstream models include the two-stream model, temporal model, spatiotemporal model, and transformer model.

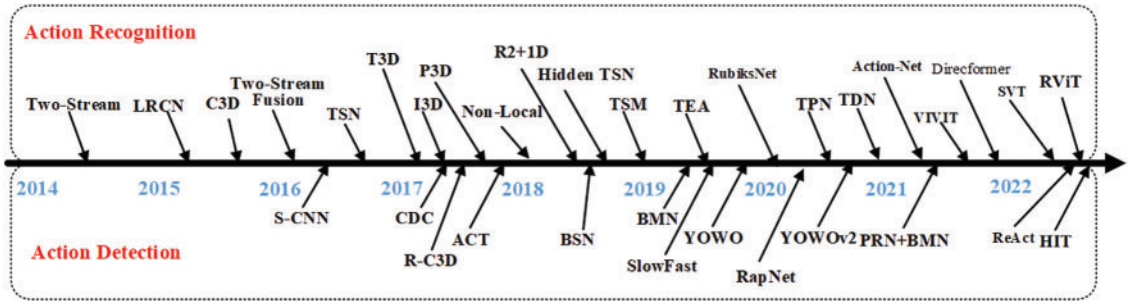

Figure 2: Important algorithms for action recognition and detection

Structurally, the two-stream model uses parallel CNNs to extract the spatiotemporal feature information separately, and it is difficult to train CNN models with two pathways separately during the model training process. The temporal model uses a cascading method to extract spatial information and temporal information respectively, which has a simple structure and less difficulty and can achieve end-to-end recognition. The traditional spatiotemporal model uses a 3D convolution network, and the model rises from two-dimensional to three-dimensional, which can extract spatiotemporal feature information at the same time, but the complexity of the model increases. In recent years, the spatiotemporal model uses the specially designed data preprocessing method to reduce the dimension of the convolution model, which greatly simplifies the structure of the model and reduces the difficulty in the training process. The recent popular transformer model comes from the NLP, which mainly adopts the attention mechanism model that is different from the other three models, and the model complexity is high and the training is difficult.

In terms of development trends, transformer is currently the most popular model, which has high accuracy but is limited by the complexity of the model, and there is a big gap from the actual deployment application. After 2019, spatiotemporal models have emerged in a new direction of using data preprocessing to achieve spatiotemporal feature data fusion, which has provided new ideas for more researchers for some time. The temporal model relies on the LSTM network and extreme variants to obtain temporal information, but it has fallen into a bottleneck, and there are no good innovation points recently. As one of the earliest deep learning models in the field of action recognition, the two-stream model has structural drawbacks, but it is still an important research direction. Table 6 shows the comparison of the characteristics of the four models.

Action detection emerged later than action recognition, but it has undergone faster development under the influence of action recognition algorithms. Temporal action detection can be divided into the one-stage method and the two-stage method according to whether the candidate region is obtained step by step. The latter can obtain higher accuracy, but the corresponding model is more complex, while the former has a more concise structure yet with lower accuracy. The role that action detection plays in the video field is just like the role object detection plays in the image field, as so many algorithms of action detection have been affected by object detection.

Table 6: Comparison of action recognition models

Models	Strengths	Weaknesses
Two stream models	Early start, high accuracy rate, wide use	Model training takes time
Temporal models	Simple structure for end-to-end applications	The feature extraction ability is limited and the development is stagnant
Spatiotemporal models	Fast recognition speed and flexible structure	3D convolution consumes a lot of computing resources
Transformer models	High accuracy	The model is complex and consumes a lot of computing resources

In terms of the tasks that need to be completed, action recognition is the premise of action detection and the most important step in action detection. From the perspective of algorithms, the research of action recognition algorithms is the basis of action detection research, and the research of action detection must solve a series of problems faced by action recognition. Action detection not only needs to pay attention to the link of action recognition but also needs to solve the problem of the time and space position of the actor. In general, many algorithms for action detection are just starting, and room for improvement in efficiency and accuracy remains large.

4.1 Difficulties and Challenges

4.1.1 Difficulty in Data Collection

At present, action recognition and action detection based on deep learning is mainly based on supervised learning, which has a strong dependence on data [123]. Hence the requirements for sample size as well as scenarios covered by datasets are increasing. Action recognition and action detection datasets are usually video data, which is larger than images and more cumbersome in the process of data collection, pruning, and labeling.

Action detection datasets need to label not only the categories of actions but also the spatial and temporal locations of actions, which is an arduous task. Temporal action detection requires the starting and ending time of the action to be annotated at a frame level. Spatiotemporal action detection also requires accurate labeling of the spatial location of the actor. This has led to a significant increase in the workload of dataset calibration, so some datasets have to adopt compromise methods such as sparse calibration to ease the pressure of data calibration.

4.1.2 High Hardware Requirements

At present, action recognition and action detection are facing mass computing power costs. With the gradual complication of deep learning models, especially since the transformer has penetrated the field of computer vision, it has become more difficult to train models, for the requirements for computer GPU computing power have gradually increased. The computing power overhead cannot be provided by ordinary hardware, which prevents the large-scale application of current action recognition and detection algorithms [124].

With the development of action recognition and action detection, to obtain features from videos effectively, the scale of corresponding datasets has been expanding continuously. Thus, some open-source datasets come to hundreds of GB or even several TB, which is a huge burden on the storage as well as read-and-write capacity for computers. At present, most datasets need to undergo preprocessing processes such as serialization or optical fluidization before training, bringing heavy computational and read-write burden to the computer.

4.1.3 Difficulty Judging Action Features

Action recognition faces the following difficulties:

The first is the complexity of fine-grained recognition. Just as there are a thousand Hamlets in the eyes of a thousand viewers, human action is complex and diverse, which may have different meanings from different perspectives. Therefore, it is difficult to strictly divide the categories of action, just as flapping actions, the speed of which directly determines whether the action itself has violent attributes.

The second is the complexity of spatial information. Lighting variations, occlusion, and noise issues caused by video background information can affect feature extraction adversely. Different observing angles towards the actor will also cause problems related to scale transformation, which will also bring trouble to the judgment of Action characteristics.

The third is the complexity of temporal information. The modeling problem of temporal dimension is the core problem of action recognition, yet also a key difficulty. From the viewpoint of current development status, the extraction of temporal information remains still very difficult [125].

Action detection also faces several difficulties:

The first is that action detection is limited by the performance of action recognition. Action recognition is the basis of action detection, but there are still many problems to be solved in the current action recognition tasks, which brings basic difficulties to subsequent action detection.

The second is the ambiguity of Action space-time dimension positioning. From the perspective of the temporal dimension, the definition of starting and ending points of certain actions is vague, and the length of actions also varies. From the perspective of spatial dimension, the motion problem needs to be considered when positioning the actor in the spatial dimension, and the combination of multiple frames avoids the problem of jittering.

4.2 Future Research Trends

4.2.1 Enrich the Dataset

Action detection and recognition mainly adopt supervised learning, and sufficient data support must be ensured. Due to the complexity of human action and the fine-grained requirements of practical applications, it is necessary to further expand on the existing datasets. At present, there are many datasets in the field of action recognition, but for specific task scenarios, the data covering related types of actions still need to be expanded. Research for action detection started late, hence the datasets of state-of-art research are few, and the coverage for scenarios remains incomplete. Therefore, it is now necessary to focus on enriching relevant datasets to rid them of data scarcity. For specific task scenarios, it can also be extended based on the existing datasets using data augmentation such as Rotation, Mixup, adding noise, and so on, which can alleviate the problem of insufficient data [126].

4.2.2 Few-Shot Learning

Given the above problems, in specific scenarios of action recognition and detection, such as the recognition and detection of violent acts in security scenarios, the recognition and detection of illegal operations in industrial production scenarios, etc., the few-shot learning method can be used to relieve pressure. The basic idea of few-shot learning is to train the network to learn metaknowledge from a large number of prior tasks, and then use the existing prior knowledge to guide the model to learn faster in the new task [127]. Few-shot learning can obtain data features from a small number of samples, reducing the intensive dependence on data in behavior recognition and behavior detection.

4.2.3 Model Lightweight

Because the existing algorithms have huge computing power overhead and are difficult to promote and deploy on a large scale, lightweight operations such as pruning the model are an important direction for subsequent research. Besides, it is also necessary to consider reducing the complexity of the model when designing the network structure [128]. For example, algorithms such as the SlowFast network and TSM network can effectively promote the progress of research by reducing computing overhead. From the perspective of practical application, the action recognition algorithm or action detection algorithm cannot be deployed in the actual task scenario. The model architecture of the algorithm is too complex, which puts forward high requirements for hardware. Therefore, from the perspective of real needs, model lightweight is a task that must be completed.

4.2.4 Transformer Model

From the above content, the current transformer model has extensive participation in action recognition and action detection. As a popular model across the fields of NLP and computer vision, the transformer model will play an important role in the future development of action recognition and detection. At present, Transformer has higher accuracy than CNN in action recognition and action detection tasks, and can better connect or collaborate with NLP in future large model research. The transformer model also needs to be optimized in terms of model parameters to reduce the hardware requirements, so it could be optimized on this basis.

5 Conclusion

This paper systematically reviews the current research status of action recognition and detection, focuses on four commonly used models for action recognition, divides action detection into temporal action detection and spatiotemporal action detection, and elaborates on the context of algorithm development in each scenario. Finally, this paper summarizes the action recognition and detection, sorts out the differences and connections between various algorithms, and expounds on the prominent problems faced by current research and the general direction of the next development.

Acknowledgement: None.

Funding Statement: This work was supported by the National Educational Science 13th Five-Year Plan Project (JYKYB2019012), the Basic Research Fund for the Engineering University of PAP (WJY201907) and the Basic Research Fund of the Engineering University of PAP (WJY202120).

Author Contributions: Study conception and design: Y. Li, X. Cui; data collection: Q. Liang; analysis and interpretation of results: B. Gan; draft manuscript preparation: Q. Liang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data in this paper can be found in Google Scholar.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Gong, M. Deng, S. Li, T. Hu, Y. Sun *et al.,* "Sika deer behavior recognition based on machine vision," *Computers, Materials & Continua*, vol. 73, no. 3, pp. 4953–4969, 2022.

[2] S. Zhang, "Research on human action detection and recognition in videos," Ph.D. dissertation, Huazhong University of Science and Technology, China, 2019.

[3] T. Hassner, "A critical review of action recognition benchmarks," in *Proc. of CVPR*, Paris, France, pp. 245–250, 2013.

[4] H. Luo, C. Wang and F. Lu, "Review of video action recognition," *Journal on Communications*, vol. 39, no. 6, pp. 169–180, 2018.

[5] D. Zhao, J. Zhang, C. Guo, D. Zhao and M. A. S. Hakimi, "Review of video action recognition method based on the depth of learning," *Telecommunication Science*, vol. 5, no. 12, pp. 99–111, 2019.

[6] Q. Chai, Y. Deng, H. Li, Y. Yu and S. Ming, "Survey on human action recognition based on deep learning," *Computer Science*, vol. 47, no. 4, pp. 85–93, 2020.

[7] Y. Zhu, X. Li, C. Liu, M. Zolfaghari and M. Li, "A comprehensive study of deep video action recognition," arXiv preprint arXiv:2012.06567, 2020.

[8] Z. Sun, J. Liu, Q. Ke, H. Rahmani, M. Bennamoun *et al.,* "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 67–77, 2022.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[10] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *Computer Vision and Image Understanding*, vol. 182, no. 3, pp. 50–63, 2019.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. of ICCV*, Seoul, Korea, pp. 4489–4497, 2015.

[12] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan *et al.,* "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. of CVPR*, Seattle, WA, USA, pp. 2625–2634, 2015.

[13] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. of ICCV*, Paris, France, pp. 523–534, 2013.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 15, no. 2, pp. 15–24, 2014.

[15] Z. Qiu, T. Yao and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proc. of ICCV*, Venice, Italy, pp. 5533–5541, 2017.

[16] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain and C. G. M. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.

[17] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami and R. Yousefzadeh, "Temporal 3D convents: New architecture and transfer learning for video classification," arXiv preprint arXiv:1711.08200, 2017.

[18] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. of 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition*, Xi'an, China, pp. 17892543, 2016.

[19] Y. Zhu, Z. Lan, S. Newsam and A. G. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Proc. of ACCV*, Perth, Australia, 2019.
[20] R. Christoph and F. A. Pinz, "Spatiotemporal residual networks for video action recognition," in *Proc. of Advances in Neural Information Processing Systems*, Honolulu, USA, pp. 3468–3476, 2016.
[21] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin *et al.,* "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. of ECCV*, Berlin, France, pp. 20–36, 2016.
[22] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. of CVPR*, Honolulu, USA, pp. 6299–6308, 2017.
[23] Z. Qiu, T. Yao, C. W. Ngo, X. Tian and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proc. of CVPR*, Long Beach, USA, pp. 1521–1529, 2019.
[24] W. Lei, P. Koniusz and D. Q. Huynh, "Hallucinating IDT descriptors and I3D optical flow features for action recognition with CNNs," in *Proc. of ICCV*, Soule, Korea, pp. 521–543, 2019.
[25] M. Kalfaoglu, S. Kalkan and A. Alantan, "Late temporal modeling in 3D CNN architectures with BERT for action recognition," in *Proc. of ECCV*, Glasgow, UK, 2020.
[26] P. Koniusz, L. Wang and K. Sun, "High-order tensor pooling with attention for action recognition," arXiv preprint arXiv:2110.05216, 2021.
[27] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. of ICPR*, Cambridge, UK, Piscataway, pp. 32–36, 2004.
[28] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *Computer Science*, vol. 51, no. 4, pp. 1–7, 2012.
[29] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. of ICCV*, Barcelona, Spain, vol. 4, no. 5, pp. 1–6, 2011.
[30] W. R. Xu, Z. J. Miao and Y. Tian, "A novel mid-level distinctive feature learning for action recognition via diffusion map," *Neurocomputing*, vol. 21, no. 8, pp. 185–196, 2016.
[31] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes," in *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, London, UK, pp. 2247–2253, 2005.
[32] M. Tong, H. Y. Wang and W. J. Tian, "Action recognition new framework with robust 3D-TCCHOGAC and 3D-HOOFGAC," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 3011–3030, 2017.
[33] M. Marszalek, I. Laptev and C. Schmid, "Actions in context," in *Proc. of CVPR*, Miami, USA, pp. 4567–4589, 2009.
[34] Y. Wang, V. Tran and M. Hoai, "Evolution-preserving dense trajectory descriptors," arXiv preprint arXiv:1702.04037, 2017.
[35] J. C. Niebles, C. W. Chen and F. F. Li, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. of ECCV*, Heraklion, Crete, Greece, Berlin, pp. 392–405, 2010.
[36] Y. W. Li, W. X. Li and V. Mahadevan, "VLAD3: Encoding dynamics of deep features for action recognition," in *Proc. of CVPR*, Las Vegas, The USA, pp. 1951–1960, 2016.
[37] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang *et al.,* "Multiview transformers for video recognition," in *Proc. of CVPR*, New Orleans, USA, pp. 271–283, 2022.
[38] W. Kay, J. Carreira, K. Simonyan, B. Zhang and A. Zisserman, "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.
[39] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska and R. Memisevic, "The "Something Something" video database for learning and evaluating visual common sense," in *Proc. of ICCV*, Venice, Italy, pp. 1–15, 2017.
[40] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu *et al.,* "UniFormer: Unified transformer for efficient spatial-temporal representation learning," in *Proc. of ICLR*, Macao, China, pp. 541-471, 2022.
[41] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier and A. Zisserman, "A short note about Kinetics-600," arXiv preprint arXiv:1808.01340, 2018.
[42] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao *et al.,* "Merlot reserve: Neural script knowledge through vision and language and sound," in *Proc. of CVPR*, New Orleans, USA, pp. 124–131, 2022.

[43] Z. Tong, Y. Song, J. Wang and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," arXiv preprint arXiv:2203.12602, 2022.
[44] J. Carreira, E. Noland, C. Hillier and A. Zisserman, "A short note on the kinetics-700 human action dataset," arXiv preprint arXiv:1907.06987, 2019.
[45] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang *et al.,* "Multiview transformers for video recognition," in *Proc. of CVPR*, New Orleans, USA, pp. 413–425, 2022.
[46] M. Cheng, K. Cai and M. Li, "RWF-2000: An open large scale video database for violence detection," arXiv preprint arXiv:1911.05913, 2019.
[47] Q. Liang, Y. Li, B. Chen and K. Yang, "Violence behavior recognition of two-cascade temporal shift module with attention mechanism," *Journal of Electronic Imaging*, vol. 30, no. 4, pp. 43009, 2021.
[48] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.
[49] X. D. Yang and Y. L. Tian, "Effective 3D action recognition using EigenJoints," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2–11, 2014.
[50] G. Willems, T. Tuytelaars and L. J. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. of ECCV*, Marseille, France, pp. 6550–6579, 2008.
[51] H. Wang, A. Klser, C. Schmid and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, pp. 60–79, 2013.
[52] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 24–31, 2012.
[53] C. Feichtenhofer, A. Pinz and R. P. Wildes, "Spatiotemporal residual networks for video action recognition," *Advances in Neural Information Processing Systems*, vol. 24, no. 2, pp. 51–62, 2016.
[54] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, Las Vegas, USA, pp. 770–778, 2016.
[55] Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. of CVPR*, Boston, USA, pp. 841–856, 2015.
[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[57] C. Feichtenhofer, H. Fan, J. Malik and K. He, "Slowfast networks for video recognition," in *Proc. of ICCV*, Seoul, Korea, pp. 529–541, 2019.
[58] S. Swathikiran, S. Escalera and O. Lanz, "Gate-shift networks for video action recognition," in *Proc. of CVPR*, Seattle, USA, pp. 641–652, 2020.
[59] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong *et al.,* "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, pp. 802–810, 2015.
[60] X. Wang, Z. Miao, R. Zhang and S. Hao, "I3D-LSTM: A new model for human action recognition," *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 3, pp. 84–96, 2019.
[61] S. Ji, W. Xu, M. Yang and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.
[62] G. Huang, Z. Liu, V. D. M. Laurens and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of CVPR*, Honolulu, USA, pp. 614–625, 2017.
[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet and A. Rabinovich, "Going deeper with convolutions," in *Proc. of CVPR*, Boston, USA, pp. 1–9, 2015.
[64] L. Ji, C. Gan and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. of ICCV*, Seoul, Korea, pp. 7083–7093, 2019.
[65] S. Hao, S. Qian and Y. Liu, "Temporal interlacing network," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 34, no. 7, pp. 154–168, 2020.
[66] L. Fan, S. Buch, G. Wang, R. Cao, Y. Zhu *et al.,* "RubiksNet: Learnable 3D-shift for efficient video action recognition," in *Proc. of ECCV*, Glasgow, UK, pp. 460–489, 2020.

[67] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang *et al.,* "TEA: Temporal excitation and aggregation for action recognition," in *Proc. of CVPR*, Seattle, USA, pp. 909–918, 2020.
[68] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lui *et al.,* "VIVIT: A video vision transformer," in *Proc. of ICCV*, pp. 5146–5158, 2021.
[69] U. Mohib, M. Y. Muhammad, M. Ahmed, D. K. Sultan, U. Habib *et al.,* "Attention-based LSTM network for action recognition in sports," *Electronic Imaging*, vol. 6, pp. 302-1, 2021.
[70] M. Patrick, D. Campbell, Y. M. Asano, I. Misra, F. Metze *et al.,* "Keeping your eye on the ball: Trajectory attention in video transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12493–12506, 2021.
[71] T. Truong, Q. Bui, C. N. Duong, H. Seo, S. Lam Phung *et al.,* "Direcformer: A directed attention in transformer approach to robust action recognition," in *Proc. of CVPR*, New Orleans, USA, pp. 5124–5139, 2022.
[72] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang *et al.,* "Multiview transformers for video recognition," in *Proc. of CVPR*, New Orleans, USA, pp. 2587–2599, 2022.
[73] K. Ranasinghe, M. Naseer, S. Khan, F. S. Khan and M. Ryoo, "Self-supervised video transformer," in *Proc. of CVPR*, New Orleans, USA, pp. 3514–3526, 2022.
[74] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen *et al.,* "Recurring the transformer for video action recognition," in *Proc. of CVPR*, New Orleans, USA, pp. 2538–2549, 2022.
[75] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo *et al.,* "Learning to localize actions from moments," in *Proc. of European Conf. on Computer Vision*, pp. 52413, 2020.
[76] M. Crucianu, *MEXaction2: Action Detection and Localization Dataset*. Henderson, NV, USA: Tech Science Press, 2015. [Online]. Available: http://mexculture.cnam.fr/xwiki/bin/view/Datasets/Mex+action+dataset
[77] D. Zhang, X. Dai and Y. Wang, "Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection," in *Proc. of Asian Conf. on Computer Vision*, Perth, Australia, pp. 54961, 2019.
[78] Y. Liu, L. Wang and Y. Wang, "FineAction: A fine-grained video dataset for temporal action localization," *IEEE Transactions on Image Processing*, vol. 31, pp. 6937–6950, 2022.
[79] U. K. Dutta, M. Harandi and C. C. Sekhar, "Unsupervised deep metric learning via orthogonality based probabilistic loss," in *Proc. of Computer Vision and Pattern Recognition*, Seattle, USA, pp. 57891, 2020.
[80] Y. Peng, Y. Zhao and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," in *Proc. of Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 85461, 2017.
[81] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru *et al.,* "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. of Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 68415, 2018.
[82] Y. Li, L. Chen, R. He, Z. Wang and L. Wang, "Multisports: A multi-person video dataset of spatio-temporally localized sports actions," in *Proc. of CVPR*, Boston, USA, 2021.
[83] H. Xu, A. Das and K. Saenko, "R-C3D: Region convolutional 3d network for temporal activity detection," in *Proc. of ICCV*, Venice, Italy, pp. 5341–5358, 2017.
[84] M. Xu, C. Zhao, D. S. Rojas, A. Thabet and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proc. of CVPR*, Seattle, USA, pp. 10156–10165, 2020.
[85] Y. W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng *et al.,* "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. of CVPR*, Salt Lake City, USA, pp. 1130–1139, 2018.
[86] S. Narayan, H. Cholakkal, F. S. Khan and L. Shao, "3C-NET: Category count and center loss for weakly-supervised action localization," in *Proc. of ICCV*, Seoul, Korea, pp. 8679–8687, 2019.
[87] T. Lin, X. Zhao, H. Su, C. Wang and M. Wang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. of ECCV*, Munich, Germany, pp. 3–19, 2018.

[88] R. Zeng, W. Huang, C. Gan, M. Tan and J. Huang, "Graph convolutional networks for temporal action localization," in *Proc. of ICCV*, Seoul, Korea, pp. 7094–7103, 2019.
[89] Y. Zhao, Y. Xiong, L. Wang, Z. Wang, X. Tang *et al.,* "Temporal action detection with structured segment networks," in *Proc. of ICCV*, Venice, Italy, pp. 2914–2923, 2017.
[90] T. Lin, X. Liu, X. Li, E. Ding and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. of ICCV*, Seoul, Korea, pp. 3889–3898, 2019.
[91] C. Zhang and J. Wu, "Actionformer: Localizing moments of actions with transformers," in *Proc. of ECCV*, Tel Aviv, Israel, 2022.
[92] X. Wang, Z. Qing, Z. Huang, Y. Feng and N. Sang, "Proposal relation network for temporal action detection," arXiv preprint arXiv:2106.11812, 2021.
[93] Z. Shou, D. Wang and S. F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. of CVPR*, Las Vegas, USA, pp. 1049–1058, 2016.
[94] Z. Shou, J. Chan, A. Zareian, K. Miyazawa and S. F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. of CVPR*, Venice, Italy, pp. 5734–5743, 2017.
[95] J. Gao, Z. Yang, K. Chen and R. Nevatia, "Turn tap: Temporal unit regression network for temporal action proposals," in *Proc. of CVPR*, Venice, Italy, pp. 3628–3636, 2017.
[96] M. Ullah, H. Ullah, S. D. Khan and F. A. Cheikh, "Stacked LSTM network for human activity recognition using smartphone data," in *Proc. of 2019 8th European Workshop on Visual Information Processing*, Rome, Italy, pp. 175–180, 2019.
[97] H. Xu, A. Das and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. of ICCV*, Venice, Italy, pp. 5783–5792, 2017.
[98] J. Gao, Z. Yang and R. Nevatia, "Cascaded boundary regression for temporal action detection," arXiv preprint arXiv:1705.01180, 2017.
[99] Y. Liu, L. Ma, Y. Zhang, W. Liu and S. F. Chang, "Multi-granularity generator for temporal action proposal," in *Proc. of CVPR*, Long Beach, USA, pp. 3604–3613, 2019.
[100] D. Shi, Y. Zhong and Q. Cao, "React: Temporal action detection with relational queries," in *Proc. of ECCV*, Tel Aviv, Israel, 2022.
[101] X. Liu, Q. Wang, Y. Hu, X. Tang and X. Bai, "End-to-end temporal action detection with transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5427–5441, 2022.
[102] A. Richard and J. Gall, "Temporal action detection using a statistical language model," in *Proc. of CVPR*, Las Vegas, USA, pp. 3131–3140, 2016.
[103] J. Yuan, B. Ni, X. Yang and A. A. Kassim, "Temporal action localization with pyramid of score distribution features," in *Proc. of CVPR*, Las Vegas, USA, pp. 3093–3102, 2016.
[104] R. Hou, R. Sukthankar and M. Shah, "Real-time temporal action localization in untrimmed videos by sub-action discovery," in *Proc. of British Machine Vision Conf.*, London, UK, pp. 1–7, 2017.
[105] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of CVPR*, Boston, USA, pp. 3431–3440, 2015.
[106] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 36, no. 6, pp. 91–99, 2015.
[107] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.
[108] J. Gao, Z. Shi, G. Wang, J. Li and X. Zhou, "Accurate temporal action proposal generation with relation-aware pyramid network," in *Proc. of AAAI*, New York, USA, pp. 10810–10817, 2020.
[109] C. Lin, J. Li, Y. Wang, Y. Tai, D. Luo *et al.,* "Fast learning of temporal action proposal via dense boundary generator," in *Proc. of AAAI*, New York, USA, pp. 11499–11506, 2020.
[110] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo *et al.,* "Gaussian temporal awareness networks for action localization," in *Proc. of CVPR*, Long Beach, USA, pp. 1587–1594, 2019.
[111] A. J. Piergiovanni and M. Ryoo, "Temporal gaussian mixture layer for videos," in *Proc. of ICML*, Long Beach, USA, pp. 1254–1268, 2019.

[112] D. Shi, Y. Zhong and Q. Cao, "React: Temporal action detection with relational queries," arXiv preprint arXiv:2207.07097, 2022.
[113] C. Zhang, J. Wu and Y. Li, "Actionformer: Localizing moments of actions with transformers," arXiv preprint arXiv:2202.07925, 2022.
[114] X. Liu, B. Song and X. Bai, "An empirical study of end-to-end temporal action detection," in *Proc. of CVPR*, New Orleans, USA, pp. 1286–1299, 2022.
[115] M. M. Puscas, E. Sangineto, D. Culibrk and N. Sebe, "Unsupervised tube extraction using transductive learning and dense trajectories," *Computers & Graphics*, vol. 38, no. 1, pp. 300–309, 2015.
[116] G. J. Faure, M. H. Chen and S. H. Lai, "Holistic interaction transformer network for action detection," arXiv preprint arXiv:2210.12686, 2022.
[117] V. Kalogeiton, P. Weinzaepfel, V. Ferrari and C. Schmid, "Action tubelet detector for spatio-temporal action localization," in *Proc. of ICCV*, Venice, Italy, pp. 4405–4413, 2017.
[118] C. Gu, C. Sun, D. A. Ross and C. Vondrick, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. of CVPR*, New York, USA, pp. 6047–6056, 2018.
[119] O. Kpüklü, X. Wei and G. Rigoll, "You only watch once: A unified CNN architecture for real-time spatiotemporal action localization," arXiv preprint arXiv:1911.06644, 2019.
[120] K. Hara, H. Kataoka and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?" in *Proc. of CVPR*, New York, USA, pp. 6546–6555, 2018.
[121] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of CVPR*, Seattle, USA, pp. 779–788, 2016.
[122] J. Zhao, S. Chong, L. Huang, X. Li, C. He *et al.,* "Action recognition based on CSI signal using improved deep residual network model," *Computer Modeling in Engineering & Sciences*, vol. 130, no. 3, pp. 1827–1851, 2022.
[123] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang *et al.,* "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, pp. 1005, 2019.
[124] X. Jiang, Z. Hu, S. Wang and Y. Zhang, "A survey on artificial intelligence in posture recognition," *Computer Modeling in Engineering & Sciences*, vol. 137, no. 1, pp. 35–82, 2023.
[125] S. Mo, X. Tan, J. Xia and P. Ren, "Towards improving spatiotemporal action recognition in videos," arXiv preprint arXiv:2012.08097, 2020.
[126] K. Chen, Z. Zhu, X. Deng, C. Ma and H. Wang, "A review of deep learning research on multi-scale object detection," *Journal of Software*, vol. 4, pp. 1201–1227, 2021.
[127] Z. Bao, D. Liu and J. Mi, "Review of video action recognition under weak supervision and few-shot learning," *Application Research of Computers*, vol. 6, pp. 1629–1635, 2023.
[128] J. Wang, S. Feng and Y. Cheng, "A review of lightweight neural network structure research on deep learning," *Computer Engineering*, vol. 47, pp. 1–13, 2021.