ITC 2/53

Information Technology and Control
Vol. 53 / No. 2 / 2024
pp. 619-652
DOI 10.5755/j01.itc.53.2.35769

An Overview of Behavioral Recognition

Received 2023/11/30 | Accepted after revision 2024/04/29

HOW TO CITE: Xie, Y., Xiang, J., Li, X., Duan, J., Li, Z. (2024). An Overview of Behavioral Recognition. *Information Technology and Control,* 53(2), 619-652. https://doi.org/10.5755/j01.itc.53.2.35769

An Overview of Behavioral Recognition

Yunjie Xie, Jian Xiang, Xiaoyong Li, Jiawen Duan, Zhiqiang Li
School of Information and Electronic Engineering, Zhejiang University of Science and Technology,
Hangzhou 310023, China

Corresponding author: freexiang@gmail.com

Human behavior recognition has become a popular research topic in the field of computer vision. With the introduction of deep learning and attention mechanisms, this field has been further promoted. However, issues such as dataset acquisition and preprocessing operations on multimodal datasets, modeling of long time information in videos, and fusion of temporal and spatial information still exist. In this paper, we first outline some video action recognition datasets and related preprocessing techniques, including frame extraction, optical flow extraction, and skeletal feature acquisition. Then, the relevant models are classified and parsed according to their characteristics and the types of input data modalities. In addition, we evaluate the performance of the models on several benchmark datasets to gain a deeper understanding of the model development process. Finally, we summarized the current challenges faced in the field of video behavior recognition, including model timeliness, data set subjectivity and effective fusion of multi-modal features, and proposed possible future improvement directions in order to provide more ideas and methods for subsequent research.

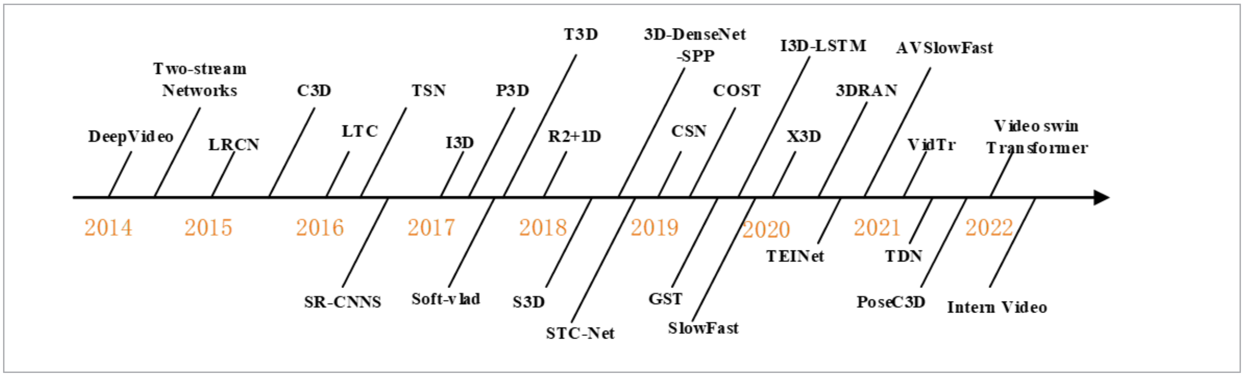
KEYWORDS: Human action recognition; Deep learning; Video; Dataset.

1. Introduction

In recent years, with the rapid updating of technological products such as smart phones and the wide dissemination of self-media, the number of videos on the Internet has shown massive growth. Video information gradually replaces pictures and text as the main information dissemination medium in people's daily life. Due to the rich multimodal information contained in video data, such as RGB data streams, sound, etc., as well as the constant generation of new data, video behavior recognition has received more and more attention in the field of computer vision. This research direction has practical application value and significance, such as in the fields of unmanned driving, virtual reality, intelligent medical care and intelligent security, which not only brings great value to human beings, but also greatly improves people's quality of life and sense of well-being.

Figure 1

Representative work in chronological video action recognition



The main goal of video behavior recognition is to establish a relational mapping between video information and human actions through various methods, so that computers can automatically recognize human actions in videos and make the technology better serve human beings.

In the past decades, with the emergence of a large number of video action recognition methods and high-quality

large-scale action recognition datasets, as well as the successful application of neural networks in the image domain (e.g., image classification, target detection, segmentation) and the rapid development of attention mechanisms [81], there has been an increasing amount of research on video action recognition, and the research on how to use deep learning for video action recognition has begun to receive attention . In Figure 1, we provide a chronological overview of some representative works in recent years.

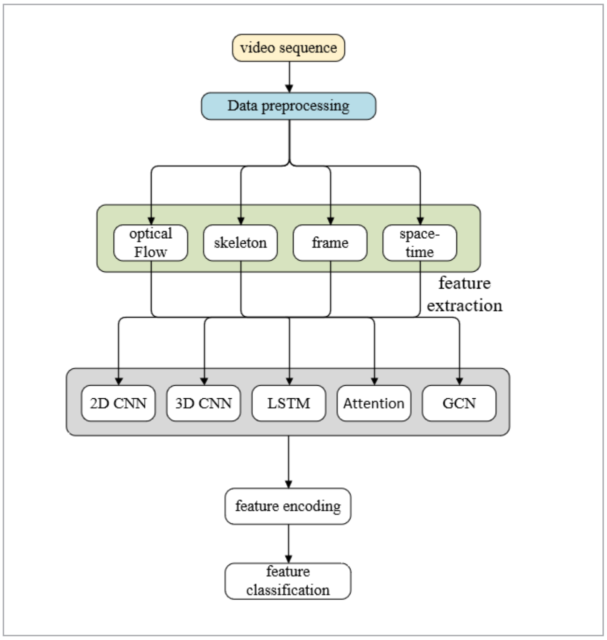
The implementation of video action recognition usually requires three components: feature extraction, feature coding and feature classification. Feature extraction is the extraction of the most useful features in the data by some specific methods, while feature coding is the numerical processing of the features (e.g., normalization or vectorization) so that different types of features can be compared in the same numerical space. Feature classification is to input the data after feature extraction and feature coding into the model and classify the data according to specific needs.

In this paper, we first introduce the video behavior recognition dataset, and then discuss and evaluate the model performance according to different model structures, from traditional manual feature extraction to the now widely used deep learning feature extraction methods. Finally, we summarize the existing models and look ahead to possible future work. The flow of video behavior recognition is shown in Figure 2.

The other chapters of this article are introduced as follows: First, in Section 2, we mainly analyze and summarize the modalities of the data. Secondly, in Section 3, we mainly review the video action recognition methods based on deep learning in re-

Figure 2

Video Behavior Recognition Process



cent years and introduce some related work. Subsequently, in Section 4, we compare mainstream methods on benchmark data sets, and display and analyze the experimental results. Then, in Section 5, we summarize and provide possible future research directions as well as challenges. Finally, we conclude the entire article.

2. Materials and Method

Unlike image tasks, the underlying data for video action recognition comes from video data from various sources. As with image tasks, deep learning methods usually improve accuracy when the amount of training data increases. In the case of the video action recognition task, this means that we need a large dataset with annotations for more effective learning.

In addition, different models also need a common dataset to measure their performance. For the video action recognition dataset, it mainly comes from the video data of various platforms on the web and the videos shot by relevant people, the former mainly obtains the relevant subtitles of the videos and matches them with the corresponding actions, and the latter mainly by manually annotating the relevant information of the actions. We briefly summarize the datasets related to video action recognition, as shown in Table 1.

Nowadays, the source of data is no longer a problem, the data size is getting bigger and bigger, and the data modality is getting more and more common, such as video, RGB image, depth map, infrared map, optical flow, skeletal data, audio, and so on. Processing these data as well as training the model requires more human and material resources, and consumes greater resources. Therefore, how to preprocess the acquired data as well as extract the features of different modalities becomes an urgent problem. Video-based action recognition is to recognize human actions from action video sequences, and there are three ideas to solve this video-based action recognition problem: first, the method of directly extracting and classifying the spatio-temporal features of the sequences; second, the method of extracting the skeletal information (2D or 3D skeletal information) for training; and third, the method of using data fusion of multiple modalities to combine the features of data from different modalities for training.

Table 1

List of popular datasets for video action recognition

Datasets	Vintages	Data modality	Actions	Samples	Average time
HMDB51 [38]	2011	RGB	51	7K	~5s
UCF101 [67]	2012	RGB	101	13.3K	~6s
Sports1M [37]	2014	RGB	487	1.1M	~5.5m
ActivityNet [28]	2015	RGB	200	28K	~7m
YouTube8M [1]	2016	RGB	3862	8M	230s
Kinetics400 [35]	2017	RGB	400	306K	10s
Kinetics600 [7]	2018	RGB	600	482K	10s
Kinetics700 [8]	2019	RGB	700	650K	10s
AVA [24]	2017	RGB	80	385K	15m
AVA-kinetics [40]	2020	RGB	80	624K	15m
NTU RGB+D [63]	2016	RGB,S,D,IR	60	56K	~8s
NTU RGB+D 120 [47]	2019	RGB,S,D,IR	120	115K	~8s
Sth-Sth V1 [23]	2017	RGB	174	108K	5s
Sth-Sth V2 [23]	2017	RGB	174	220k	5s

Video-based action recognition and image classification are very similar in some aspects. Compared to image classification, video data introduces a temporal dimension. Relevant researchers usually further process the acquired video data by extracting consecutive video segments into frame-by-frame RGB images according to certain time intervals, and then the extracted video frames are used for target detection and classification at the image level according to the temporal order, but since a single data pattern does not recognize the corresponding actions well, in subsequent studies, the RGB images are mostly used in combination with other modalities.

Skeletal data describes the human body in graphical form. Specifically, it exists in the form of joints, and there is usually a limit on the number of joints to ensure that the required skeletal information can be comprehensively covered while reducing data redundancy. Compared to RGB video, the skeletal representation is more robust to changes in viewpoint and appearance. In addition, the pelvic bone of the model is usually chosen as the root bone of the model, and the transformation matrix of each bone relative to the root bone is recursively derived based on the root bone. The feature dimensions obtained from human skeletal data are much lower compared to RGB image frames, and therefore more computationally efficient.

The 3D skeletal data can more directly represent body part motion-related features such as joint angles and velocities, thus allowing for easier and more accurate action recognition. The current deep learning methods for human behavior recognition based on skeletal data can be broadly classified into three types according to the expression of skeletal keypoints: LSTM-based recurrent neural network, convolutional neural network (CNN)-based and graph neural network (GCN)-based.

Depth image data is essentially the effect of combining a normal RGB three-channel color image with a depth map. It contains image channels with information about the distance to the surface of the object in the viewpoint scene, the channels themselves are similar to grayscale images, and each pixel value is the actual distance to the object as measured by the sensor. Compared to RGB image data, depth image data is more resistant to the effects of external factors unrelated to behavior, such as the lighting of the shooting environment and the texture of the actor's clothing. In depth video images, the actor and the surrounding shooting scene are usually highly recognizable, and the depth data obtained is less susceptible to interference from external factors, so subsequent related studies often use depth image data.

Optical flow feature is an important feature in human motion recognition. It is a trajectory feature produced by the movement of the foreground target itself in the scene, the shift of the camera's shooting viewpoint, or the joint movement of the two that makes the pixel points in the video sequence change over time. The computational basis is based on the assumption that the luminance change of an image only originates from the movement of an object, and reflects the motion of a human body by utilizing the luminance change of pixel points on adjacent frames in the time domain. In order to track the position of the human body during motion, FlowNet2 proposed by Nvidia, by changing the training strategy, introduces a branching network to deal specifically with small movements of the object, uses a stacked architecture to optimize the optical flow effect, and distorts the input image to achieve higher resolution optical flow results. PWC-Net proposed by Berlin et al. feeds each consecutive frame into the PWC-Net to compute the optical flow features in PWC-Net, which utilizes multi-scale features to replace the sub-network cascade, and its design follows the three principles of Pyramid Feature Extraction (Pyramid), Optical Flow Mapping (Warping), and Matching Relevance Cost Volume Measurement (Cost Volume), whereas, both Warping and Cost Volume do not contain any learning parameter, which can be used to optimize the optical flow results while ensure that the network is effective, greatly reducing the size of the network model.

The variety of data types and modalities also makes the selection of a particular modality or multiple modalities for the video behavior recognition problem a critical step for researchers to weigh. The characteristics of different modalities and the applicable scenarios are described in Table 2.

Table 2

Advantages and disadvantages of each mode and applicable scenarios

Modalities	Vantages	Disadvantages	Applicable Scenarios
RGB	Low cost, easy to collect; rich appearance information; wide range of applications	Very sensitive to changes in light, background and perspective	Less irrelevant content, single scene
Depth chart	Provides structural and shape information in three dimensions; robust to illumination	Lack of color and texture information; distance limitations exist; susceptibility to occlusions	Dim or bright, close up, unobstructed scenes
Audio frequency	Easily localize actions by combining video information in time series	Lack of self-appearance information	Suitable for multimodal learning, as a form of auxiliary information
Optical flow	Provides more timing information, works well	High requirements for the movement changes of the joints, more cumbersome acquisition methods, high computational costs	Behavioral scenes with large, unobstructed movements
Skeletal graph	Provides three-dimensional information about the human body's posture; representation is simple and effective; insensitive to viewpoint and context	Lacks appearance and shape information; indicates sparse content; noisy information	Scenes that can effectively extract skeletal keypoints

3. Deep Learning Video Action Recognition

In recent years, with the tremendous progress of deep learning technology, various deep learning models have been proposed. Due to the powerful representation and superior performance of deep learning-based methods, the current mainstream research in this field focuses on designing different types of deep learning frameworks. Therefore, in this section, we review the deep learning-based approaches for video action recognition in recent years and preset some related works. In the following, they are broadly categorized according to the different characteristics of the models. The advantages and disadvantages are shown in Table 3.

3.1. Handcrafted to 2D CNN Based

As early as before 2015, when convolutional neural networks have not been applied on a large scale, hand-crafted features represented by IDT [83], including spatio-temporal volume-based [5], spatio-temporal points of interest (STIP)-based [39], and trajectory-based [88], have dominated the field of video behavior recognition because of their relatively high accuracy and good robustness. With the popularity of deep convolutional neural networks in visual tasks, the problem of hand-crafted features has been gradually exposed, which not only requires high production cost but is also difficult to be scaled up, therefore, it is a natural step to apply deep learning to video problems.

DeepVideo [37] is a pioneering work in applying deep learning to video, and its main research is how to migrate CNNs from the field of image recognition to the field of video behavior recognition, the authors proposed to use a 2D CNN model on each video frame individually, and tested several fusion methods (late fusion, early fusion, and slow fusion) to learn spatio-temporal features for video action recognition. However, its migration learning performance on UCF101 [67] is about 20% lower than the hand-crafted IDT features. Nevertheless, DeepVideo's several fusion methods have contributed to the subsequent development in the video field. The advantages and disadvantages of the manual and deep learning methods are shown in Table 4.

Table 3
Advantages and disadvantages of deep learning based correlation methods

Deep Learning Specific Methods		Vantages	Disadvantages
Introduction of CNN		Replaced manual production and increased efficiency	Recognition accuracy is not high enough and is less effective
Two-stream convolutional networks		Integration of temporal information mitigates the lack of dynamic features	Separation of temporal and spatial streams, computationally overloaded
Multi-stream CNN		More comprehensive compensation for overall information	Complexity is greatly increased and the lift is not significant
CNN Series	3D CNN	Ability to extract both temporal and spatial information	The amount of parameters is too large to optimize
	Variants of 3D CNN	Reduces the associated computational effort to some extent	Not as effective for long videos
LSTM Series	Conventional LSTM	Acquisition of movement information over longer time spans	Large number of participants, easy to lose information and difficult to optimize
	Variants of LSTM	Provides more comprehensive information on spatial and temporal representations	Reduced training difficulty compared to conventional LSTMs
Introduction of the attention mechanism		Easier access to appropriate spatio-temporal information	Significantly improved accuracy without increasing the amount of computation
Introduction of Skeletal + Graph Convolution		Easier access to relevant campaign information, greatly reducing redundancy	Skeletal joint points are difficult to obtain

Table 4
Comparison of the advantages and disadvantages of crafting and deep learning method

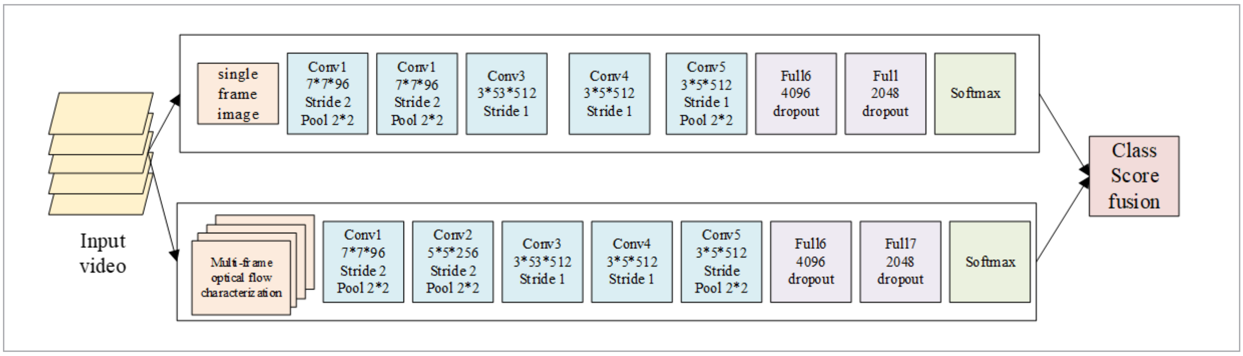
Feature Extraction Methods	Vantages	Disadvantages
handicraft	Suitable for small datasets, relatively fast	Requires some prior knowledge of design features
deep learning	Suitable for large datasets, features learned through web learning	Slower speeds, higher hardware requirements

3.2. Based on Dual/Multistream Networks and Their Variants

In order to solve the problem that the recognition performance of the RGB unimodal input model is limited by the lack of dynamic feature information, Simonyan et al. [66] proposed a dual-stream network with spatial and temporal streams, as shown in Figure 3, where the spatial stream utilizes the original video frames as the input to capture the visual appearance information, and the temporal stream takes a set of optical flow images as the input to capture the motion information between the video frames. The dual-stream CNN recognition model is trained separately, and the prediction results of the two networks are finally fused. This method is also the first CNN-based method to achieve similar performance to the previous best hand characterized IDT on UCF101 and HMDB51. Many network structures have been subsequently explored on this basis. The advantage of the two-stream structure is its high accuracy, but it is slow. The model proposed by Wang et al. [88] feeds multi-scale video frames and optical streams

Figure 3

Structure of the two-stream network model


single frame image
Conv1 7*7*96 Stride 2 Pool 2*2
Conv1 7*7*96 Stride 2 Pool 2*2
Conv3 3*53*512 Stride 1
Conv4 3*5*512 Stride 1
Conv5 3*5*512 Stride 1 Pool 2*2
Full6 4096 dropout
Full 2048 dropout
Softmax
Input video
Multi-frame optical flow characterization
Conv1 7*7*96 Stride 2 Pool 2*2
Conv2 5*5*256 Stride 2 Pool 2*2
Conv3 3*53*512 Stride 1
Conv4 3*5*512 Stride 1
Conv5 3*5*512 Stride Pool 2*2
Full6 4096 dropout
Full7 2048 dropout
Softmax
Class Score fusion

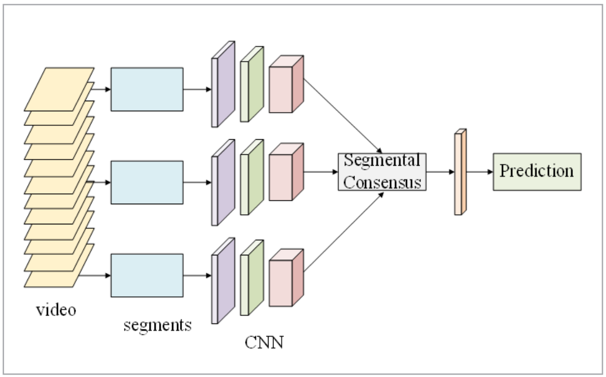
into a dual-stream CNN, which transforms the convolutional feature maps by spatio-temporal normalization and channel normalization. In order to better fuse the optical flow with RGB images, Wan et al. [82] proposed a dual-stream convolutional network (LSF CNN) with long- and short-term spatiotemporal features, which consists of a long-term spatiotemporal feature extraction network (LT-Net), which takes the stacked RGB images as inputs, and a short-term spatiotemporal feature extraction network (ST-Net), which takes optical flow as inputs and extracts the optical flow from the two neighboring frames are estimated. The dual-scale spatio-temporal features are then fused in a fully connected layer and fed into a linear support vector machine (SVM) to fully learn the deep features in both spatial and temporal domains.

After that, many papers on dual-stream networks have been released, and several other works [13, 22, 91], extended the dual-stream model to extract long-term video-level information for video behavior recognition. However, the dual-stream network uses a relatively shallow network, so many works naturally thought of using a deeper network to further improve the accuracy. Wang et al. [90] found that the effect of simply replacing the network with a deeper one is not very obvious, and the authors believed that the training dataset for action recognition is very small compared with ImageNet, which is easy to overfitting on the training dataset. In order to solve this problem, they designed several improved experiments, such as: using smaller learning rate, using more data enhancement techniques, using higher Dropout, etc. to prevent the overfitting of the deep network. Based on this, a two-stream network is trained using the VGG16 model, and the performance of the model is greatly improved. Since then, research on deeper networks has gradually achieved better results, e.g., ResNet [27], Inception [71], and proved that deeper networks can usually achieve higher video action recognition accuracy [117].

Wang et al. [86] proposed a new deep architecture (SR-CNNS), which not only shares great modeling capability with the original dual-stream CNN, but also shows support for dual-stream semantic region CNN, and uses an end-to-end network Faster-RCNN [62] network instead of the standard spatial streams, which can then extract semantic information about objects, people and scenes. Meanwhile, Feichtenhofer et al. [19] discussed the dual-stream network, mainly discussed and verified the fusion of the network, and concluded that the optimal fusion position is at the last convolutional layer, and used 3D pooling instead of 2D pooling after fusion to further improve the performance. The residual connection is also proposed to establish information connection between spatio-temporal convolutions to facilitate the information interaction and better learning of spatio-temporal features. Wang et al. [85] used a spatio-temporal pyramid approach to fuse the spatial and temporal features in the pyramid structure so that they can reinforce each other. This model also adopts a hierarchical fusion strategy and uses a uniform spatio-temporal loss for overall training, which also achieves good results. Jain et al. [32] proposed a Hybrid Random Forest Bi-Convolutional Recurrent Neural Network (HRF Bi-CRNN), in which a model based on a Bi-Convolutional Recurrent Neural Network (CRNN) is fused with Random Forest classification for accurate HAR. In addition, a Bi-CRNN based autoencoder is used to learn the spatio-temporal data from motion and sensors, which is further utilized to integrate the Random Forest for human gestures and movements for final classification and detection.

Similarly, more and more attention has been paid to another problem of optical flows, i.e., they still cannot capture temporal information at a distance. To solve this problem, Wang et al. [91] constructed a Temporal Segment Network (TSN) based on the idea of remote temporal structure modeling for video-level action recognition, and its model structure is shown in Figure 4. Specifically, TSN first divides the whole video into several segments, which are uniformly distributed along the time dimension. Then TSN randomly selects a video frame in each segment according to a certain weight. However, unlike the original dual-stream network that uses ClarifaiNet [110] as the base model, this model adopts the BN-Inception [31] network to extract features along the temporal and spatial streams separately. TSN is able to model long time video because the model is based on the whole video, and this method also reduces the training cost of long video sequences, and many subsequent works have been improved on this basis. In addition, the Siamese network designed by Wang et al. [94] extends the two-stream network to a two-stream concatenated network by extracting features from the frames before the action occurs (prerequisite) and the frames after the action occurs (resultant), and modeling the action as a transformation on the high-level feature space. Considering that many frames in a video sequence may not work well for video behavior recognition, the method proposed by Kar et al. [36] learns to pool these frames with discriminative and informative properties while discarding most of the non-informative frames in a single time-scan of the video, and then embeds the put method into deep learning to further improve the efficiency of the model.

Figure 4
TSN model structure



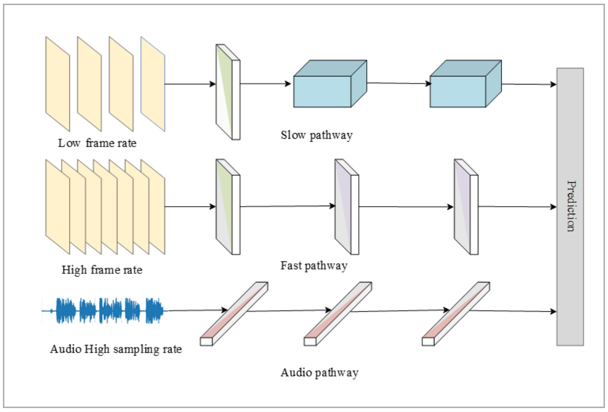
After the wide application of dual-stream networks, more and more researchers are trying to merge different types of data, or further extend the dual-stream networks into multistream networks, the core of which is multimodal applications. The most commonly used data types are RGB maps, depth maps, optical flow, skeletal keypoints, audio and so on. Object information data is another important data source, because most human behaviors involve human-object interactions, such as playing basketball, swimming, riding a bicycle, and so on. Wu et al. [99] proposed a new object- and scene-based semantic fusion network. The network uses a three-layer neural network that combines low-level CNN features containing frame-based features, object features from a large-scale CNN object detector, and scene features from a CNN scene detector, and discovers semantic information between video classes and objects and scenes by fusing the information from the network checking and backpropagation.

Audio data usually appears together with video data, which is complementary to visual information. Wu et al. [100] introduced a multistream framework to fully utilize the rich multimodal information in video. The model first trains three convolutional neural networks to model spatial, short-term motion and audio, and then employs a long-short-term memory network to explore the dynamic information over long periods of time. The optimal fusion weights are adaptively determined during the fusion process to generate the final scores for each class, which outperforms the state-of-the-art models at that time. Later, Xiao et al. [103] designed AVSlowFast based on SlowFast [17], and its model structure is shown in Figure 5. Similar to the native SlowFast, it not only has slow and fast visual pathways, but also has faster audio pathways which are deeply fused with it to unify the representations, so that the audio contributes to the hierarchical audio-visual concept.

Wang et al. [87] proposed a three-stream convolutional network for motion feature extraction. The model consists of RGB data frames, optical streams and globally stacked motion difference images

Figure 5

AVSlowFast model structure



(MSDI) to generate the corresponding spatial, local temporal and global temporal streams, respectively. In addition, the model also combines the advantages of Gaussian Mixture Model (GMM) and VLAD to encode the data according to the overall profile distribution and the corresponding differences with respect to the clustering centers, and a soft vector called Soft-VLAD is developed to further represent the extracted features. Bilen et al. [4] proposed a

four-stream network structure, which uses a dynamic image method based on RGB images and optical streams, i.e., the temporal data such as RGB images and optical streams are encoded using "sorting pools", and then the resulting RGB dynamic image network and dynamic optical stream network are combined with the original RGB network and optical stream network to form a four-stream network. After that, the resulting RGB dynamic image network and dynamic optical flow network are combined with the original RGB network and optical flow network to form a four-stream network structure, which is finally fused to further predict the class of actions.

The dual/multistream 2D CNN architecture learns different types of information (e.g., spatial and temporal) from the input video through separate networks, and then fuses them to obtain the final result, which enables traditional 2D CNNs to process video data efficiently and achieve high video behavior recognition accuracy. However, multiple streams of inputs means that the number of parameters to be trained for the deep model will be larger, and how and where to fuse the different data types is also a key issue, which requires higher design requirements for the feature fusion module. At the same time, this type of architecture is still not powerful enough for long-term dependency modeling, i.e., it has limitations in effectively modeling video-level temporal information, which can be compensated by time-series modeling networks (e.g., LSTM).

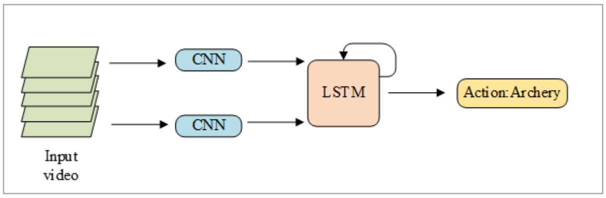
3.3. Based on CNN and LSTM

For problems such as long time span action information in video cannot be effectively extracted, researchers believe that video is a time sequence in nature, and have explored recurrent neural networks (RNN) for temporal modeling in video as well as contextual inference networks, which effectively extracts the global contextual information,

especially using the Long Short-Term Memory Network (LSTM), whose model structure is shown in Figure 6.

Figure 6

LSTM-based Behavior Recognition Model

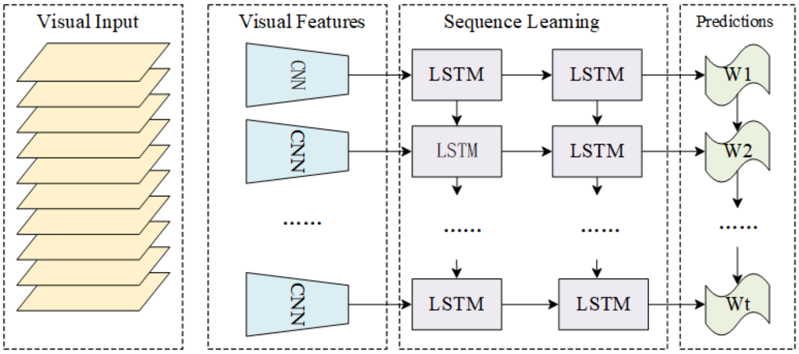


The LRCN proposed by Donahue et al. [14] in 2015 is considered to be a pioneer in the field, and its model structure is shown in Figure 7. Compared with the previous models, the circular convolution models have double depth because they can combine learning at the spatial and temporal levels, respectively, use the feature maps of the CNNs as inputs to the deep LSTM network, and aggregate the frame-level CNN features into the video-level prediction. This model has obvious advantages when the target concept is complex or the training data is limited, and can model complex temporal information dynamically, but it is still a bit inferior to the hot dual-stream network at that time. Since then, however, many variations on the CNN-LSTM framework have emerged.

Ullah et al. [79] proposed an action recognition method using CNN and Deep Bidirectional LSTM (DB-LSTM) networks to process video data. Specifically,

Figure 7

LRCN model structure



the model, in order to reduce redundancy and complexity, extracts deep features from every six frames of the video by introducing a data augmentation module that weights different categories of actions, and then learns the sequential information between the frame features using the DB-LSTM network. Meanwhile, the model also adopts a bidirectional LSTM structure, which can better capture the long-term dependencies in the video sequences, and thus improve the performance of the model. Gammulle et al. [21] focused their attention on learning salient spatial features through CNNs and then mapping their temporal relationships with the help of a Long Short-Term Memory (LSTM) network, which better integrates spatial features from CNNs and temporal features from LSTM models. Srivastava et al. [68] used an encoder LSTM to map the input video into a fixed-length representation, which is then decoded by a decoder LSTM to perform the tasks of video reconstruction and prediction in an unsupervised manner. The LiteEval model designed by Wu et al. [101] utilizes a lightweight CNN along with the co-operation of a coarse LSTM and a fine LSTM to dynamically decide whether or not to compute more robust features for the incoming video frames at a finer scale for efficient behavior recognition. Some other works [26, 114], employ a bidirectional LSTM, which consists of two independent LSTMs for learning forward and backward temporal information for video action recognition.

Li et al. [45] proposed a model consisting of a multi-stream 2D network for learning spatial and temporal representations. The model enhances action recognition by learning a hierarchical multi-granular deep spatio-temporal video representation. Specifically, it models each granularity as a single stream via a 2D CNN (for both frame and motion streams). In addition, the model uses a Long Short-Term Memory Network (LSTM) on the frame and motion streams to record long-term temporal dynamics. With a softmax layer on top of each stream, classification scores can be predicted from all streams and then learned in an end-to-end manner based on a novel fusion scheme with multi-granularity score distribution.

Liu et al. [48] proposed a global context-aware attention model (GCA-LSTM) based on gradient-tailored adaptive long-term memory network, which can selectively focus on the information joints in the action sequence with the help of global context information and adjust their attention weights accordingly, which effectively mitigates the problems of gradient vanishing and gradient exploding, and thus improves the model's robustness and generalization ability. In order to realize a reliable attentional representation of the action sequences, a cyclic attention mechanism is proposed for the GCA-LSTM network, in which the attentional performance is iteratively improved to further enhance the performance of the network. Yu et al. [109] proposed a pseudo-recursive residual neural network (P-R RNN), which utilizes a cyclic recursive architecture to correlate the features of the previous and current frames through a P-R unit, with different connections between the units to form each neural network. A two-stream CNNS model (GoogLeNet) is used to extract local temporal and spatial features, respectively. Then, the local spatial and temporal features are integrated into the global long-term temporal features using P-R RNN. Finally, the Softmax layer fuses the outputs of the dual-stream P-RRNN, which also shows an excellent result, and its model structure is shown in Figure 12.

Wang et al. [92] proposed a lightweight action recognition architecture composed of CNN, LSTM, attention model and joint optimization, which only uses RGB images as input data, and firstly uses CNN convolutional layer and fully connected layer to extract local spatial features and semantic features to separate the object from the background. Accordingly, for temporal feature extraction, two LSTM networks named ConvLSTM and FC-LSTM are constructed after the CNN convolutional layer and the fully connected layer, respectively, to model the temporal information in different visual perception layers, and at the same time, two different attentional models are designed to learn the temporal focus of the actions, and finally, a joint optimization module is adopted to train a more robust LSTM network. The network is trained to be more robust through a joint optimization module. VideoLSTM proposed by Li et al. [46] combines convolutional and motion-based attention into a soft-attention LSTM to localize actions through action category labeling and temporal attention smoothing operations to better capture spatial and motion information.

The action recognition algorithm combining CNN and LSTM networks can effectively utilize the advantages of the two models for extracting image features and processing time-series data, and the combination of the two can analyze the action information more comprehensively, deal with the differences between different people and changes in the speed of action, and have a better generalization ability. At the same time, the introduction of LSTM can capture the long-term dependencies in the

time series, which is conducive to more accurately recognize complex action sequences. However, there are some drawbacks of this type of model, for example, the algorithm is better for recognizing some basic actions such as walking and running, but less accurate for recognizing some subtle differences in actions. In addition, it requires a large amount of data to train the model, and the training time is long, and since LSTM can only handle fixed-length sequences, there may be some recognition errors.

3.4. CNN-based and Attention Mechanisms

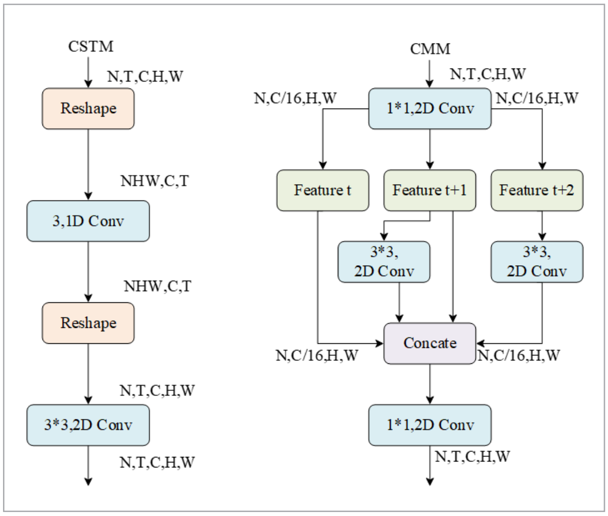
Recently, with the excellent results of the attention mechanism in 2D detection tasks, many researchers have begun to try to introduce the attention mechanism into the video domain as well. Wang et al. [95] proposed Non-local considering that both convolutional and cyclic operations deal with one localized block at a time, and using the non-local operation as a generalized family of building blocks used to capture long term dependencies. Its non-local operation computes the response at a location as a weighted sum of features at all locations. Jiang et al. [34] proposed the STM module, which consists of a channel-based spatio-temporal module (CSTM) for presenting spatio-temporal features and a channel-based motion module (CMM) for efficiently encoding motion features. The model structure is shown in Figure 8, in which STM blocks are used to replace the original residual blocks in the ResNet architecture to form a simple and efficient STM network by introducing a very limited additional computational cost.

In order to help the model allocate more attentional resources to the target region during the feature learning process, and thus suppress redundant information. Cai et al. [6] proposed a three-dimensional residual attention network (3DRAN) based on CBAM [97], a convolutional attention module proposed by Woo et al. The module sequentially infers the attention map along two independent dimensions, channel and space, and multiplies the attention map by the input feature map to reweight key features. In addition, this module is a lightweight and generalized module that can be seamlessly integrated into any CNN architecture. On the other hand, 3DRAN builds on its foundation by extending the 2D residual attention structure to 3D space, which consists of an attention mechanism and a 3D ResNets architecture, capable of capturing spatio-temporal information in an end-to-end manner.

Li et al. [42] argued that temporal modeling is the key to action recognition in video, based on which they proposed a Temporal Excitation and Aggregation (TEA) module, which consists of a Motion Excitation (ME) module and a Multi-Temporal Aggregation (MTA) module, the former mainly calculates the temporal difference between feature level and spatio-temporal features to motivate the motion-sensitive channels of the features. The latter mainly deforms the local

Figure 8

CSTM-CMM Model Structure



convolution into a set of sub-convolution to form a hierarchical residual structure, which effectively simulates the long-range temporal relations on long-distance frames. Moreover, the two components of the TEA module are complementary in temporal modeling, and their effectiveness and efficiency are competitive with the best previous approaches. Unlike TEA's attention mechanism in the temporal dimension only, Liu et al. [49] constructed a two-stream residual spatio-temporal attention network (R-STAN) based on temporal and spatial attention mechanisms as well as residual networks, in which each tributary stream is constructed by stacking residual spatio-temporal attention blocks (R-STABs), which enables R-STAN to have the ability to generate attention-aware features along both temporal and spatial dimensions. The R-STAN has the ability to generate attention-aware features along both temporal and spatial dimensions, and combines the characteristics of residual learning to construct a very deep network to learn the spatio-temporal information in the video, which induces the attention-aware features from the R-STAB to change adaptively, in order to reduce the redundant information.

Liu et al. [50] proposed a TEINet network with a Motion Enhancement Module (MEM) and a Temporal Interaction Module (TIM) at its core, where the former is used to enhance motion-related features while suppressing irrelevant information (e.g., background), and the latter supplements the temporal context information in the form of channels. By decoupling the modeling of channel correlation and temporal interactions to learn temporal features, the temporal structure can be captured flexibly and efficiently for model inference.

In order to capture complex action variations, Li et al. [43] first proposed a spatio-temporal deformable convolution module (STDA) with an attention mechanism. The module can utilize both long-range temporal dependencies across multiple frames and long-range spatial dependencies within each frame, thus enabling the extraction of discriminative global information at both the inter-frame and frame levels. Unlike traditional convolutional localization in the local regularity sense field, it can further capture temporal and spatial irregularities by learning different convolutional filter offsets with attentional information to significantly improve the overall recognition performance.

Zhu et al. [118] proposed a novel spatio-temporal mesh transformer (STMT) to directly model mesh sequences. The model uses a hierarchical converter with intra-frame offset attention and inter-frame self-attention. The attention mechanism allows the model to engage freely between any two vertex patches to learn non-local relationships in the spatio-temporal domain. Masked vertex modeling and future frame prediction serve as two self-supervised tasks to fully activate bidirectional and autoregressive attention in the layered converter. The proposed method achieves state-of-the-art performance compared to skeleton-based models in common MoCap benchmark tests.

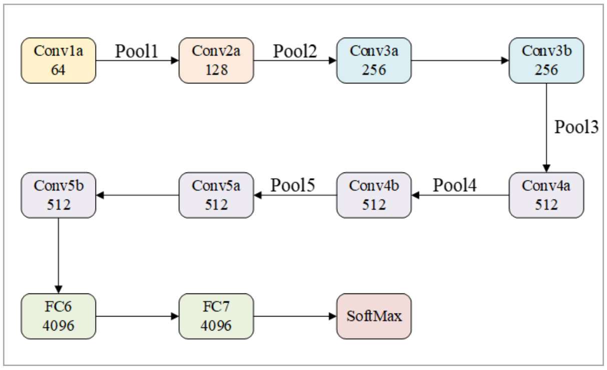
3.5. Based on 3D CNN and Its Variants

3.5.1. The Emergence of 3D CNNs

Although the above methods have achieved good results, but the core is still operating on a two-dimensional basis, the recognition of video actions has an additional temporal dimension, and thus the use of 3D CNN for video action recognition has gradually become a new research hotspot. The general approach of this algorithm is to stack a series of video frames in the temporal dimension to form a cube structure as the input to the model, and then learn the temporal and spatial features of the video information adaptively through a convolutional neural network under the supervision of a given action category label. This method avoids the operation of extracting temporal and spatial features separately in the dual-stream model, and it does not need to design a spatio-temporal feature fusion module, and captures different video features directly from the video data at the same time, so that it is easier and more convenient to acquire the corresponding feature information, and the efficiency of the model is improved.

Ji et al. [33] were the first to use 3D CNN method for action recognition, which is the general approach mentioned above, but due to the shallow depth of the model, it is not sufficient to show its potential. Tran et al. [74] improved on this foundation and proposed C3D based on VGG16, which uses 3D convolution on adjacent frames to model spatio-temporal features in a unified way. The performance of the modified model on standard benchmarks is not satisfactory, but the model has a strong generalization ability, and it is generally used as a general-purpose feature extractor for a wide range of video tasks, and the structure of the network is shown in Figure 9.

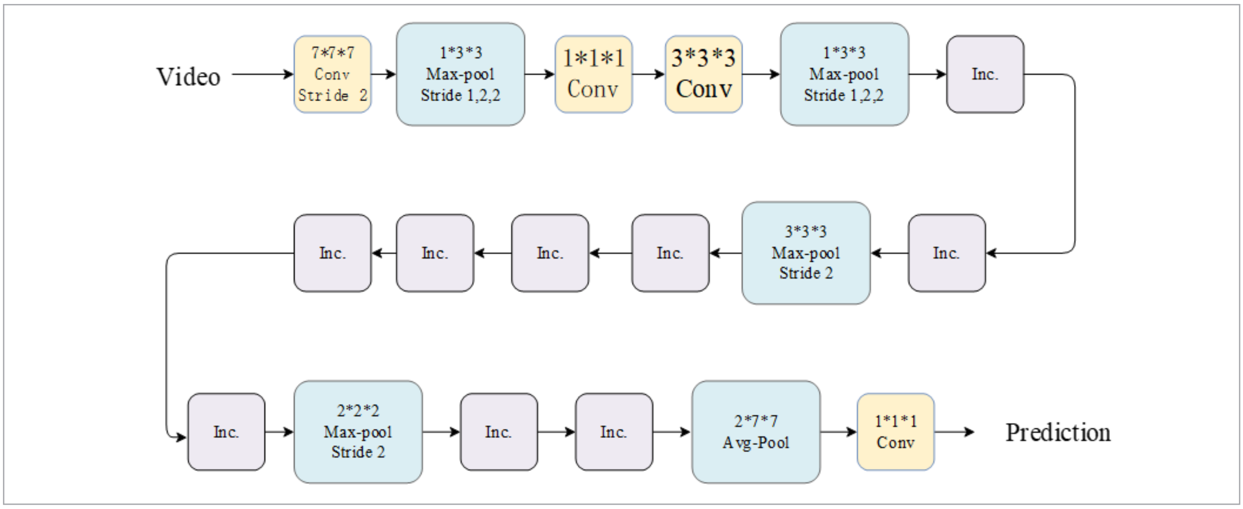
Figure 9
C3D Model Structure



The ensuing problem is that the large number of 3D convolutional operations generates a large number of parameters, increasing the computational effort and making the network difficult to optimize. In order to solve this problem and further improve the performance of video behavior recognition, some other works [9, 17, 115, 116] investigated 3D CNN models with dual- or multi-stream design. For example, Carreira et al. [9] proposed an I3D model based on 2D CNNs, whose model structure is shown in Figure 10. The model takes a video clip as input, extracts a video into a series of video frames, and expands the 2D CNN into a 3D CNN, which makes it possible to extract both temporal and spatial information simultaneously. Meanwhile, for the model weights, the model inflates the pre-trained 2D model weights from ImageNet to the corresponding weights in the 3D model, so that the 3D CNN does not have to train a new network from scratch.

To address the input problem of fixed-size video frames in video recognition, Yang et al. [106] proposed a 3D-dense connected convolutional network

Figure 10
I3D Model Structure



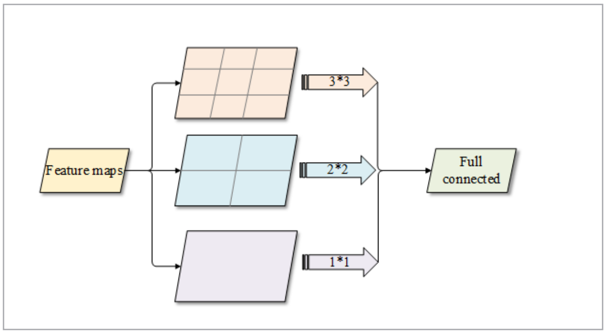
(3D-DenseNet-SPP) based on spatial pyramid pooling (Figure 11), which adopts a migration learning approach, and is pre-trained on a large-scale dataset, Kinetics400, then fine-tuned with a small dataset, thus reducing the training difficulty of the model. This model uses a migration learning approach, and after pre-training on a large-scale dataset, Kinetics400, the model is fine-tuned with a small dataset to reduce the difficulty of training. Meanwhile, Huang et al. [30] argued that pre-training an effective 3D ConvNet on large-scale datasets usually requires an expensive pre-training process, in order to avoid this situation, they designed a 2D inflated convolution operation and a parallel 3D ConvNet architecture to direct the pretraining parameters of the 2D convolutional network to the 3D convolution. In addition, the model is further enhanced by cumulative gradient descent and video sequence decomposition.

While traditional studies have only explored relatively shallow 3D architectures, Hara et al. [25], in order to determine whether current video datasets can support the training of ultra-deep convolutional neural networks (CNNs) with spatio-temporal 3D kernels, the model is directly based on 2D ResNet, which is modeled after I3D by replacing all the 2D convolutional kernels and pooling kernels with 3D ones. Meanwhile, inspired by the fact that 2D CNNs greatly contribute to generalized feature representation after pre-training on the ImageNet dataset, the authors propose for the first time a variety of 3D CNNs modeled on the Kinetics dataset (ResNet, ResNext-101) starting from zero and ranging from relatively shallow to very deep, demonstrating that deep 3D CNNs in combination with the Kinetics dataset together will trace the successful history of 2D CNNs and ImageNet, and will further stimulate the advancement of computer vision.

Inspired by SENet [29], STCNet [11] constructed a new block for modeling the correlation between 3D CNN channels based on temporal and spatial elements, which can be added as residual units to different parts of 3D CNNs, and embedded this block into ResNet and ResNext architectures, which resulted in a significant performance improvement, and its model is shown in Figure 11. The model is shown in Figure 12. The work also transfers the knowledge from the pre-trained 2D CNNs to the randomly initialized 3D CNNs for stable weight initialization, which is later fine-tuned on the target dataset, greatly reducing the training cost.

Figure 11

Network Architecture with Pyramid Structure Pooling Layers



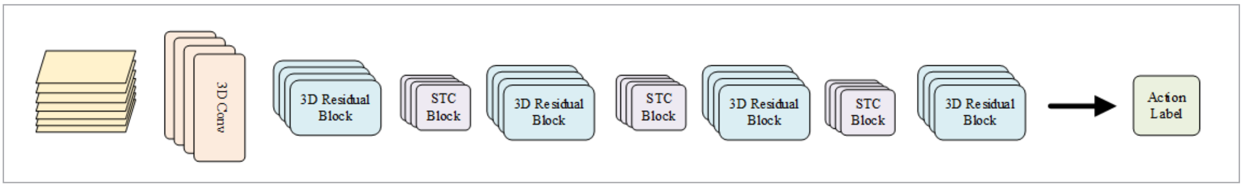
3.5.2. Variants of 3D CNN

3D convolution-based methods are very effective for modeling discriminative features in both spatial and temporal dimensions for video behavior recognition. However, many 3D CNN-based frameworks contain a large number of parameters, increasing the computational effort. To reduce the complexity of 3D convolutional network training.

Therefore, some studies [61, 70, 77] aim to decompose 3D convolution. Qiu et al. [61] proposed a pseudo-3D network model: P3D (Pseudo-3D Re-sidual Networks), which simulates 3D convolution by means of 2+1, i.e., a combination of traditional 2D convolu-

Figure 12

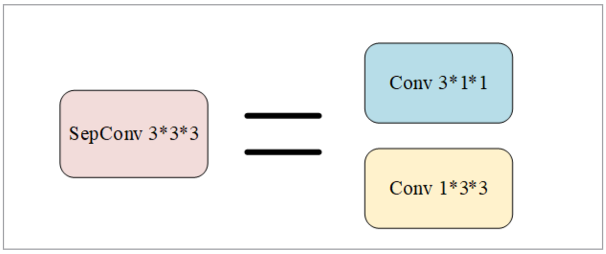
3D STC-ResNet model structure



tion and 1D convolution. Similarly, Tran et al. [77] proposed a method called R(2+1)D to decompose 3D convolution. The core idea of these two methods is to decompose a 3D convolution kernel into two independent operations along the temporal and spatial dimensions, i.e., a 3 × 3 × 3 convolution kernel is decomposed into a two-dimensional spatial convolution kernel (1 × 3 × 3) and a one-dimensional temporal convolution kernel (3 × 1 × 1), and its equivalent substitution is shown in Figure 13: this method greatly reduces the complexity of the model, and further enhances the efficiency of the model. Sun et al. [70] proposed a factorized spatio-temporal convolutional network (FstCN), which decomposes the original 3D convolutional kernel learning into a sequential process of learning a 2D spatial kernel in the lower layer (spatial convolutional layer), and then learning a 1D temporal kernel in the upper layer (temporal convolutional layer).

Figure 13

Equivalent transformations for 3D convolution



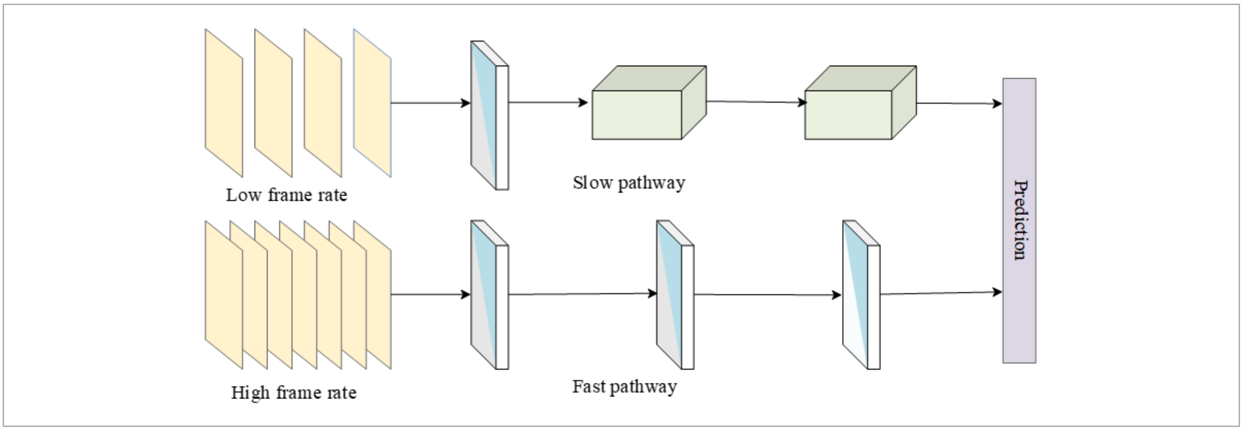
Another approach to reduce model complexity is to mix 2D convolution and 3D convolution. Zhou et al. [115] proposed a hybrid convolutional model, MicT, which integrates 2D CNN and 3D CNN modules to generate deeper and more informative feature maps while reducing the training complexity of each round of spatio-temporal fusion. They also proposed a new end-to-end 3D network MiCT-Net based on MiCT to better explore the spatio-temporal information in human behavior. Meanwhile, Xie et al. [104] also proposed a new model S3D based on I3D, which makes a series of attempts based on I2D and I3D, e.g., Bottom-Heavy-I3D and Top-Heavy-I3D, where the former one uses 3D time-domain convolution at the lower level of the network and 2D convolution at the higher level, and the latter one is just the opposite. In the former case, 3D time-domain convolution is used at the lower level of the network, while 2D convolution is used at the higher level. Experimental results show that the latter model is faster and more accurate.

The CSN proposed by Tran et al. [76] based on the idea of group convolution is also a good variant of 3D CNN, which decomposes 3D convolution by separating channel interaction and spatio-temporal interaction, which improves the recognition accuracy and reduces the computational cost. Secondly, 3D channel-separated convolution also provides a form of regularization, which has lower training accuracy but higher testing accuracy than conventional 3D convolution. Li et al. [41] proposed a new collaborative spatio-temporal network (CoST) based on C2D and C3D, which retains the advantages of the original C2D to learn spatal and temporal features independently and the C3D to jointly learn unconstrained parameters, and collaboratively encodes spatial-temporal features by sharing constraints and applying weights to the learnable parameters. By sharing the convolutional kernels of different views, the spatial and temporal features are learned collaboratively so as to benefit from each other, and the performance of the model is greatly improved by weighted summation of the complementary features. Luo et al. [53] proposed a new decomposition method to decompose the feature channel into spatial and temporal groups in parallel, called spatio-temporal aggregation (GST). This decomposition is similar to the two-stream network, which uses one path to model spatial information and the other to model temporal information, so that the two groups focus on the static and dynamic cues, respectively, and then spatial-temporal features are spliced together. At the same time, this decomposition not only results in fewer parameters, but also in higher parameter efficiency, which enables quantitative analysis of the contributions of spatial and temporal features in different layers.

SlowFast, as an efficient network with fast and slow paths, is modeled as shown in Figure 14, which uses two-stream convolution as inputs; the slow path runs at a low frame rate to capture detailed semantic information, i.e., the spatial information, while the fast path runs at a high temporal resolution to capture the fast-changing motion, i.e., the spatial information.

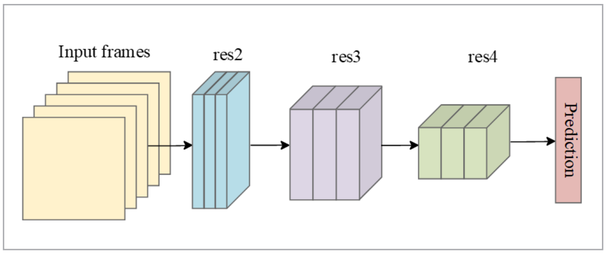
In order to merge the temporal and spatial semantic information of the different paths, the fast and slow paths are also unilaterally connected in a lateral direction to merge the features of the fast path into the

Figure 14
SlowFast model structure



slow path and match the different temporal dimensions of the two paths. Although SlowFast is also a two-stream input, it is different from the traditional two-stream structure, which has more temporal and spatial streams, while the SlowFast model adopts the same backbone network to simulate different temporal velocities rather than spatial and temporal models. The absence of operations such as calculating optical flow makes the model more efficient and lighter. On this basis, the team further introduced the X3D [16] model, which is gradually expanded along multiple network axes (e.g., time, space, width, and depth), and finally formed a 2D image classification architecture. X3D adopts a simple step-by-step network expansion approach, which expands a single axis at each step, thus achieving good accuracy in the complexity trade-off. The model structure is shown in Figure 15. In order to expand X3D to a specific target complexity, it uses progressive forward expansion and then backward contraction, which is very light in terms of network width and parameters, but achieves more advanced performance.

Figure 15
X3D Model Structure



Meanwhile, there are some works [96, 111] that combine 3D convolution with LSTM to enhance the acquisition of contextual information and global attention. Wang et al. [96] proposed a new model I3D-LSTM based on I3D combined with LSTM, which pre-trained the 3D CNN model on the Kinetics dataset to improve the generality of the model, and then introduced the Long Short-Term Memory Network (LSTM) to model the high-level temporal features generated by the pre-trained 3D CNN model, which realized the modeling of the sequence of low-level and high-level spatial-temporal features and the sequence of high-level spatial-temporal features. and high-level temporal feature sequences are modeled. Similar to I3D-LSTM, Zhang et al. [111] proposed a long term 3D convolutional fusion network (LT3D-CFN), which replaces the CNNs in the dual-stream network with 3D CNNs in order to extract features from the spatial and temporal dimensions of a single video clip. In addition, long term correlations are established between clips of a single motion video by adding a deep LSTM network.

3.5.3. Dense Connectivity for 3D CNNs

The TSN model in the dual-stream approach realizes the recognition of long videos by segmenting the video to randomly extract frames. In 3D CNN, the simplest way to solve this problem is by superimposing multiple short videos. This is achieved by using a convolution on each short video, e.g., using a 3×3×3 convolution kernel, but some information is lost due to the superposition of multiple convolution kernels, resulting in a loss of accuracy. In order to solve this problem, Varol et al. [80] proposed the model LTC, which not only demonstrated that the LTC-CNN model with increasing time horizon improves the accuracy of action recognition, but also proved the importance of high-quality optical flow estimation for learning accurate action models by examining the effect of different low-levels on the representations, such as the original values of video pixels and the optical flow vector field. In the same period, Diba et al. [12] proposed a T3D network based on DenseNet using a densely connected structure, in order to make the 3D convolutional network can be better initialized, and at the same time proposed a migration learning method, transferring the knowledge from the pre-trained 2D CNNs as a 3D CNN for the stable initialization of the weights, which greatly reduces the number of training samples, accelerates the training This greatly reduces the number of training samples and speeds up training, while maintaining the integrity of the original temporal information as much as possible, so as to make the final prediction. Related researchers have also investigated knowledge distillation to improve the motion representation of 3D CNN frameworks. For example, Stroud et al. [69] introduced a distilled 3D network (D3D) consisting of a student network and a teacher network, and designed a model that uses the teacher network to pre-train on the optical flow, and then the student network is trained on the RGB video, and also extracts the knowledge of the teacher network that was trained on the sequence of the optical flow, and adjusts the spatial flow to simulate the temporal flow, effectively combining the two models into a single flow, which greatly improves the speed of prediction, while maintaining the integrity of the original temporal information. The spatial flow is adjusted to simulate the temporal flow, which effectively combines the two models into one flow and greatly improves the inference effect of the model.

3D CNN-based video behavior recognition methods typically perform spatio-temporal processing over limited time intervals via window-based 3D convolutional operations, where each convolutional operation focuses only on a relatively short period of contextual information in the video. Meanwhile, RNN-based methods recursively process video sequence elements, and thus cannot model relatively long-term spatio-temporal dependencies. However, Transformer can directly attend to the completion of video sequences through its scalable self-attention mechanism, and thus can effectively learn remote spatio-temporal relationships in videos. Therefore many recent works have also investigated Transformer-based video behavior recognition in RGB videos. In the next section, we review the Transformer-based approach.

3.6. Based on Transformer with ViT

In Video Behavior Recognition, Transformer is suitable for modeling and processing sequence data by introducing a self-attention mechanism to construct global connections. In video behavior recognition, the video is usually regarded as a time sequence, and each moment (or frame), can be described as a vector, using Transformer to model the video. The structure of Transformer-based video behavior recognition model is shown in Figure 16, the core of the model is the Transformer block in the box, the Transformer block consists of an encoder and a decoder, and the encoder mainly consists of a number of self-attentive blocks to encode the input sequence. The decoder shares a similar architecture to the encoder, except for the additional encoder-decoder attention mechanism in each block. This design allows the Transformer to perform well in long-term dependency modeling.

In recent years, with the excellent results of Transformer in NLP, through its scalable self-attention mechanism, thus it can effectively learn long-distance spatio-temporal relations in videos. Therefore, many recent works have also investigated Transformer-based video behavior recognition in RGB videos. Liu et al. [51] proposed Video Swin Transformer based on Transformer, whose model structure is shown in Figure 17. It advocates the localized generalized bias in video Transformer. Meanwhile, it also introduces a cross-channel attention mechanism and a hierarchical multi-scale feature fusion technique to better capture the temporal and spatial information in the video. In addition, the model also utilizes the power of pre-trained images, and its method possesses better speed and accuracy than previous methods that use spatio-temporal decomposition to globally compute self-attention. Meanwhile, Arnab et al. [3] proposed

Figure 16

Video Behavior Recognition Model Based on Transformer

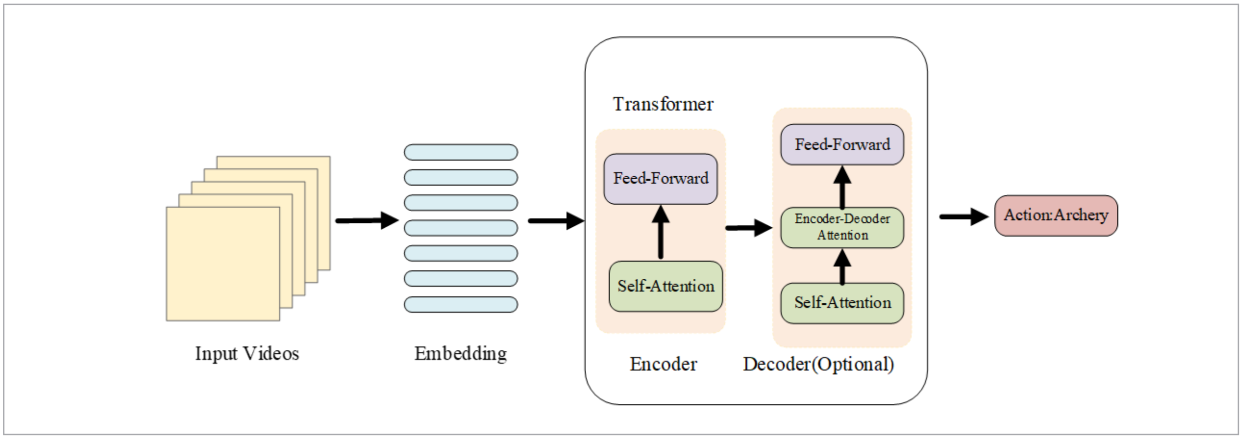
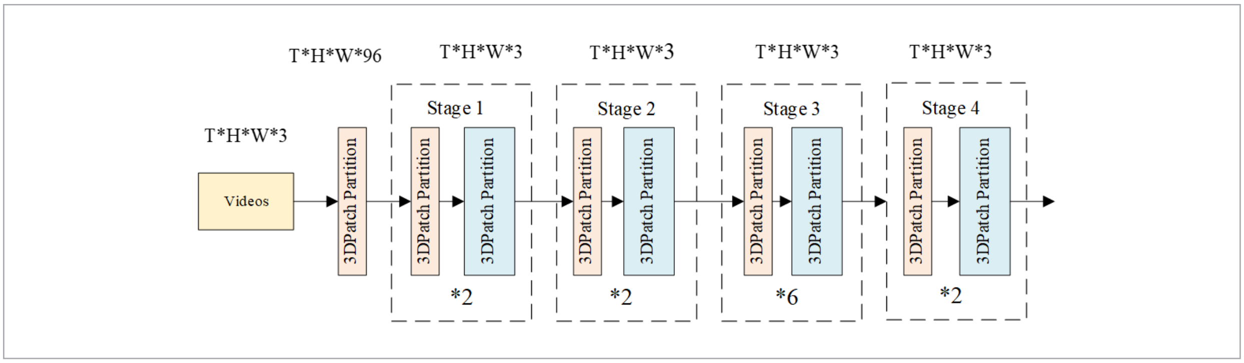


Figure 17

Video Swin Transformer Model Structure



a pure Transformer-based video classification model. This model extracts spatio-temporal markers from the input video, which are then encoded by a series of transformers. In order to deal with long sequences of Token encountered in the video, their variant operation is applied to the model to enable it to decompose the spatial and temporal dimensions of the input.

Truong et al. [78] proposed a spatio-temporally oriented attention architecture (DirecFormer) based on depth-separable convolution and attention, which utilizes the magnitude of attention between frames as well as the direction of attention to learn the correct order of frames within an action video, which can improve the model's receptive field and the efficiency of feature extraction while maintaining the depth of the model. Yan et al. [105] proposed a multiview transformer, a model consisting of multiple individual encoders, each of which is dedicated to a single input representation. Horizontal connections between the individual encoders are used to efficiently fuse information from different representations of the input video.

Piergiovanni et al. [57] proposed a method that can turn ViT coding into an efficient video model. It can seamlessly process both image and video inputs. The input is sparsely sampled for training and inference. At the same time, the model is easily scalable to accommodate large-scale pre-training of ViT without full fine-tuning. Qing et al. [60] proposed a new motion recognition scheme (MAR) based on VideoMAE [73], which reduces redundant computation by discarding a certain percentage of patches and running on only a portion of the video. Meanwhile, in order to enable ViT to easily perceive the details beyond the corresponding patches, it proposes a unit-run masking module to preserve the spatio-temporal correlation in the video. In order to solve the problem of not being able to achieve accurate classification, it also proposes a bridging classifier to bridge the semantic differences between the reconstructed ViT coded features and the features dedicated for classification.

Although ViT has been breaking the records of many visual tasks, it is computationally intensive, memory-intensive, and not friendly to embedded devices. In order to make video behavior recognition available to lightweight devices, Nguyen et al. [55] used the modernized structure of ConvNet to design a new action recognition backbone, which consists of 2D convolution, without using any 3D convolution, remote attention plug-in, which greatly reduces the computational effort without losing computational performance, and thus can be deployed on lightweight devices. The network consists of only 2D convolution, without using any 3D convolution, remote attention plug-in, which greatly reduces the computational volume without losing computational performance, and enables lightweight, which can be deployed on lightweight devices. Meanwhile, Zhang et al. [112] proposed VidTr, which has a video classification mechanism with separable attention. Compared with common 3D networks, VidTr is able to aggregate spatio-temporal information by stacking attention and provide better performance with higher efficiency. To further optimize the model, it proposes standard deviation-based topK pooling, which reduces the computation by discarding non-informative features along the temporal dimension, and is more effective for prediction work that requires long-term temporal inference of behaviors. In the improved correlated behavior recognition algorithm based on ViT, directly applying spatio-temporal converter on video data will bring heavy computational and memory burden due to the significant increase in the number of patches and the secondary complexity of self-attention computation. Xiang et al. [102] proposed a time patch shifting (TPS) method for efficient 3D self-attention modeling in Transformer for video action recognition. With no additional cost, TPS shifts a portion of Patch with a specific pattern in the time dimension, thus converting the original spatial self-attention operation into a temporal one. At the same time, TPS is a plug-and-play module that can be inserted into existing 2D converter models to enhance spatio-temporal feature learning.

Most of the existing vision base models focus only on image pre-training, and compared to the image domain which has many base models, there are few base models in the video behavior recognition domain. To fill this gap, Wang et al. [84] proposed a generalized video base model, InternVideo, which exploits generative and discriminative self-supervised learning, effectively explores masked video modeling and video language comparison learning as pre-training objectives, and selectively coordinates video representations of these two complementary frameworks in a learnable manner to facilitate a variety of video applications.

As the size of large pre-trained models continues to grow, and standard fully fine-tuned task-based adaptation strategies become prohibitively expensive in terms of model training and storage, parameter-efficient transfer learning has emerged. Some researchers have used techniques such as knowledge distillation and transfer learning in this area as well. Wang et al. [93] proposed the Masked Video Distillation (MVD) technique, which is a simple but effective two-stage masked feature modeling framework for video representation learning. It starts with low-level features to pre-train video and picture models, and then uses the resultant features as targets for masked feature modeling. Student models are extracted from video and image teachers through mask modeling. Experiments show that the video Transformer pre-trained by spatio-temporal co-teaching outperforms the model refined by a single teacher on numerous video datasets. Meanwhile, Pan et al. [56] proposed a new spatio-temporal adapter (ST-Adapter) for parameter-efficient fine-tuning of each video task. The model has built-in spatio-temporal reasoning in a compact design so that pre-trained image models do not require temporal knowledge, i.e., reasoning about dynamic video content can be performed at a very small parameter cost, which can be drastically reduced without decreasing the efficiency.

Wang et al. [89] proposed a video architecture approach called temporal difference network (TDN) in order to alleviate the problem of temporal modeling in video motion recognition. The core of the approach is to design an effective temporal module (TDM) by utilizing temporal difference operators and systematically evaluating its impact on short-term and long-term motion modeling to capture multi-scale temporal information for effective action recognition. TDN adopts a two-level differential modeling paradigm. Specifically, for local motion modeling, temporal differences over consecutive frames are used to provide a finer-grained motion pattern for the 2D CNN, whereas for global motion modeling, temporal differences across segments are combined to capture the tele-structure motivated by the motion features, with a performance comparable to the best performance on the Something-Something V1, V2, and Kinetics-400 datasets.

Modern deep learning models compute self-attention by performing spatio-temporal 3D convolution, decomposing 3D convolution into spatial and temporal convolution separately, or along the temporal dimension under the premise that feature mappings across consecutive frames can be well aggregated by default. Arguing that this premise may not always be particularly applicable to regions with large deformations, Long et al. [52] proposed a new block of inter-frame attention, known as stand-alone inter-frame attention (SIFA), which rescales offset predictions based on the difference between two frames to reshape the deformable design. Taking each spatial position in the current frame as a query, the local deformable neighborhood in the next frame is considered as a key/value. Then, SIFA measures the similarity between the query keys as an independent attention to the weighted average of the temporally aggregated values. And further, the SIFA modules are inserted into ConvNets and Vision Transformer to design SIFA-Net and SIFA-Transformer, respectively. Wu et al. [98] proposed a memory-enhanced multi-scale vision transformer (MeMViT) in order to realize the long video recognition. The converter replaces most methods that attempt to process multiple frames at once with a method that processes the video online and caches it to memory at each iteration. With memory, the model can be modeled over time with reference to previous contexts, temporarily supporting video lengths up to 30 times longer than those supported by existing models, with only a 4.5% increase in computational effort, whereas traditional methods require >3000% of the computational power to perform the same operation.

3.7. Based on Skeletal Joint Points with Graph Convolution

Skeletal sequences encode the motion trajectories of human joints, which characterize a rich set of information about human motion. Skeletal data can be obtained by applying pose estimation algorithms to RGB video or depth maps, and can also be collected by motion capture systems. The use of skeletal data for video behavior recognition has many advantages, as it provides information about body structure and pose, its intrinsically simple and informative representation, scale invariance, and its robustness to changes in clothing texture and background. Due to these advantages, combined with its accuracy and the availability of low-cost depth sensors, skeleton-based video behavior recognition has attracted much attention from the research community in recent years. Meanwhile, GCN can capture the connectivity model of intra-graph dependencies through message passing between nodes. It is just able to capture the high-level spatial structure and dynamic temporal information of skeleton data. The work of introducing graph convolutional neural network GCN combined with human skeleton keypoints into behavior recognition.

Thakkar et al. [72] proposed Partial Graph Convolutional Network (PB-GCN) based on Deformable Part's Model (DPM), which plots the human skeleton into four subgraphs and shares joints between them Learning uses a recognition model based on Partial Graph Convolutional Networks, which uses relative coordinates and temporal displacements to improve performance. Shi et al. [64] designed a directed graph neural network for extracting information about joints, bones, and their relationships to represent skeletal data as a directed acyclic graph based on kinematic dependencies between joints and bones in the natural human body. The model can adapt itself to the topology of the graph during the training process. Li et al. [44] introduced an encoder and decoder structure in order to capture richer dependencies, extending the existing skeleton graph to represent higher-order dependencies. An action-structure graph convolutional network (AS-GCN) that uses action-structure graph convolution and temporal convolution as the basic building blocks is further proposed to learn spatial and temporal features for action recognition, which helps to capture more detailed action patterns through self-supervision.

Shi et al. [65] proposed a two-stream adaptive graph convolutional network (2S-AGCN), in which the topology of the graphs can be learned either uniformly or individually in an end-to-end manner via a BP algorithm. This data-driven approach increases the flexibility of the model in graph construction and brings more generalization to accommodate various data samples. In addition, the paper proposes a dual-stream framework for modeling both first- and second-order information, which significantly improves the recognition accuracy. However, since the graph convolution operation is a local operation, it cannot fully investigate the non-local joints that are crucial for recognizing actions. Therefore, Zhang et al. [113] proposed a context-aware graph convolution network (CA-GCN), which utilizes asymmetric correlation measures and higher-level representations to compute contextual information for greater flexibility and better performance. In addition, the network simplifies the network by considering the context term of each vertex by integrating the information of all other vertices in addition to computing the local graph convolution, thus achieving a performance comparable to the then optimal network. Ye et al. [107] constructed a GCN network by stacking multiple Context Coding Networks (CeN), which learns the dependency between two joints by merging the contextual features of the remaining joints. As an advantage of CeN, dynamic graph topology is constructed for different input samples and different depths of graph convolutional layers. The final results achieve state-of-the-art performance on three challenging datasets.

Yu et al. [108] proposed a base multimodal network (MMNet), which fuses skeleton and RGB modal data to improve the accuracy of integrated recognition by effectively applying mutually complementary information. It uses a spatio-temporal graph convolutional network as a skeleton modality to learn the attentional weights, and then transfers these attentional weights to the network of RGB modalities, which then efficiently capture the mutually complementary features in different RGB -D video modalities complement each other and provide more discriminative features for HAR. Duan et al. [15] proposed a bone-based action recognition method, PoseConv3D, which relies on 3D heatmap volumes rather than graphical sequences. Compared with the GCN-based method, PoseConv3D is more efficient in learning spatio-temporal features, more reliable against pose estimation noise, and has better generalization ability. In addition, the model can handle multiplayer scenarios without additional computational cost, and its hierarchical features can be easily integrated with other modalities in the early fusion stage, providing a good design space for performance improvement. Ahn et al. [2] in order to solve the problem of cross-modal data requiring separate models and balanced feature representations, proposed a spatio-temporal transformer (STAR), which can efficiently use two cross-modal features as recognizable vectors. The STAR transformer encoder consists of a full spatio-temporal attention (FAttn) module and a proposed sawtooth spatio-temporal attention (ZAttn) module. Similarly, the sequential decoder consists of a FAttn module and a proposed Binary Spatiotemporal Attention (BAttn) module. STAR-transformer further gains a great deal by correctly arranging the pairing of FAttn, ZAttn, and BAttn modules to learn effective multi-feature representations of spatio-temporal features. Qi et al. [59] proposed an efficient graph convolutional network based on multi-order feature information (MFGCN) for human skeletal action recognition, which introduces angular features (called fourth-order features), which are implicitly embedded into other third-order features by encoding the angular features in order to robustly capture detailed features in the spatiotemporal dimension and enhance the ability to differentiate between similar actions; second, a content-adaptive approach is used to construct an adjacency matrix to dynamically learning the topology between skeleton joints; finally, a spatio-temporal information sliding extraction module (STISE) is developed to improve the interconnectivity of spatio-temporal information.

In summary, skeletal morphology provides information about body structure, which represents human behavior in a simple, efficient, and informative way. However, video behavior recognition using skeleton data still faces challenges because of its very sparse representation, noisy skeleton information, and lack of shape information which is important when dealing with human-computer interaction. Therefore, some existing work on video behavior recognition also focuses on the use of depth maps.

4. Comparison of Models and Analysis of Their Effects

This section compares the dominant methods on the benchmark dataset. First, in Section 4.1, we present standard evaluation schemes. Next, we divide the common benchmarks into four categories based on model characteristics, and in Section 4.2, we compare the accuracy of hand-crafted and using convolutional neural network models on common datasets such as UCF101, HMDB51, and Kinetics400. In Section 4.3, we extend the NTU RGB+D dataset and compare the accuracy of LSTM and its variant models. Similarly, in Section 4.4 and Section 4.5, we extend the Skeleton-Kinetics dataset and the Something-Something V2 dataset, respectively, and then compare the accuracy on the Skeleton-+GCN model and the Transformer and ViT models, respectively. Finally, in Section 4.6, we compare the mAP of the models.

4.1. Evaluation Program

During model training, a video frame clip is usually randomly selected to constitute a small batch of samples. However, for a fair comparison, we need a standardized process in the evaluation phase. For 2D CNNs, a widely adopted evaluation scheme is to take 25 frames uniformly in each video, crop each frame at the corners and center, then perform horizontal flipping for data enhancement, and finally average the prediction scores of all samples. This means that we use 250 frames per video for inference. For 3D CNNs, a widely adopted evaluation scheme called the 30-view strategy, which uniformly draws 10 clips in the relevant dataset and performs three data enhancement schemes for each video clip. Specifically, the shorter spatial side is extended to 256 pixels, three 256 × 256 frames are taken to cover the spatial dimension, and the scores are averaged for prediction.

In terms of evaluation metrics, we report the accuracy of single-label action recognition, as well as the mAP (mean average precision) of multi-label action recognition.

4.2. Manual and Convolutional Neural Networks

In chronological order, we first provide in Table 5 the results of early hand-crafted and subsequent initial attempts at dual/multistream networks using deep learning. We find that before deep learning was widely used, most methods were still based on hand-crafted, e.g., IDT, spatio-temporal trajectory-based, and spatio-temporal point-of-interest-based models whose results were not particularly good. In the absence of motion/time modeling, the performance of DeepVideo is not as effective as other modeling methods. However, DeepVideo was sort of the first time that the behavior recognition task was transferred from traditional methods (non-CNN) to deep learning. Subsequently, more and more models started to use deep learning, and the dual-stream model used both data modalities of RGB images and optical streams for the first time, which led to a significant performance improvement, and many subsequent models continued to follow the structure of the dual-stream model. For example, the TDD model uses trajectory pooling to extract motion-aware CNN features based on the dual-stream.

Next, we compare 3D CNN-based methods in the lower part of Table 5. C3D does not perform as well as the two-stream model in 2D CNNs, although it uses a large amount of data during the training process, which may be due to the difficulty of optimizing the 3D kernel. Inspired by this, several papers, including I3D, P3D [61], R(2+1)D [77], and S3D [104], decompose the 3D convolutional filter into a 2D spatial kernel and a 1D temporal kernel to simplify training. In addition, I3D introduces an inflation strategy that avoids training from scratch. They inflate 3D model weights from a trained 2D network. By using these techniques, they can achieve good performance without optical flow.

By comparing the results of the experimental data in Table 5, we can see that not all cases are necessarily better for 3D CNNs than for 2D CNNs. Many models absorb the advantages of both, and some of them are able to obtain higher recognition accuracy than the two-stream network (above), and their performance is comparable to that of 3D CNN. Since these models are based on 2D CNNs and do not use optical flow, they offer efficiency advantages in both training and inference, and most of them are also real-time methods. Therefore, based on the need for efficiency, we believe that 2D CNN + time-domain modeling is a promising research direction. Here, temporal modeling can be attention-based, stream-based, or 3D kernel-based.

Table 5

Performance of manuals, CNNs/and their variants, dual/multistream, 3D CNNs, and related variants

Model Category	Model Name/ Literature Index	Pre-training	Modal (computing, linguistics)	Accuracy on relevant datasets		
				UCF101	HMDB51	Kinetics400
craft	IDT [83]	ImageNet	RGB	86.4	61.7	
	time and space track [5]					
	TSPI [39]					
	Orbit [88]					
Dual/ multistream and CNN variants	DeepVideo [37]	ImageNet	RGB	65.4	-	
	Two-Stream [66]	ImageNet	RGB+Flow	88.0	59.4	
	LSF CNN [82]	ImageNet	RGB+Flow	94.8	70.2	
	Literatures [90]	ImageNet	RGB+Flow	91.4	-	
	Siamese [94]					
	TDD [88]					
	Literatures [19]	-	RGB+Flow	94.2	68.9	
	Literatures [85]	-	RGB+Flow	94.6	68.9	
	TSN [91]	Sports-1M	RGB+Flow	94.2	69.4	
	SR-CNNS [86]					
	Literatures [100]	Kinetics-400	RGB+Audio	92.2	-	
	AVSlowFast [103]	Kinetics-400	RGB+Audio	93.6	-	
	Literatures [87]					
	Literatures [4]	Kinetics-400	RGB+Flow	95.5	72.5	
3D CNN and its variants	C3D [74]	Sports1M	RGB	82.3	56.8	59.5
	I3D [9]	ImageNet, K400	RGB	95.6	74.8	71.1
	3D-DenseNet-SPP [106]	K400	RGB	88.94	/	/
	Literatures [30]	ImageNet	RGB	92.7	69.1	/
	Two-stream I3D [25]	K400	RGB	98.0	80.7	/
	STCNet [11]	Sports1M	RGB	96.5	74.9	/
	P3D [61]	Sports1M	RGB	93.7	/	71.6
	R(2+1)D [77]	K400	RGB	96.8	74.5	72.0
	MicT [115]	ImageNet+Sports1M	RGB	88.9	63.8	/
	S3D [104]	ImageNet+ K400	RGB	96.8	/	72.0
	CSN [76]	Sports1M	RGB	/	/	77.0
	CoST [41]	ImageNet	RGB	/	/	82.7
	SlowFast [17]	K400	RGB	/	/	79.8
	X3D [16]	IamgeNet	RGB	/	/	80.4
	LTC [80]	ImageNet	RGB+Flow	92.7	67.2	/
	T3D+TSN [12]	Sports1M	RGB	93.2	63.5	/

4.3. LSTM and Its Variants

We found that although 2D CNN can handle spatial information very well, its effect on temporal information is still unsatisfactory, such as TDD and TSN models have captured the spatial information very well, and utilized convolutional neural networks to acquire spatial features layer by layer, because the temporal dimension has not been well utilized, so that the above models do not have much change in the action, and the features captured are more similar to the image recognition task, and lack of significance of the video action recognition task.

LSTM can be a good solution to the problem of time-domain modeling, and it can extract action information over a long time span. Although LRCN, as the pioneer in this field, did not surpass the then popular dual stream network in terms of effect, the idea of stripping the temporal and spatial information and sending the feature maps after convolutional operation as input to LSTM for the next step of learning, especially for the complex dynamic modeling of temporal information, the method still plays a great role in the effectiveness of the method.

Most of the subsequent LSTM-based variants of the model continue to follow the idea of LRCN, and the accuracy of the model on UCF101, HMDB51, and other datasets is shown in Table 6.

Table 6
LSTM and its related variants

Model name or method	Pre-training	Input modal	Accuracy on relevant datasets		
			UCF101	HMDB51	NTU RGB+D
LRCN [14]	ImageNet	RGB+Flow	82.3	/	/
DB-LSTM [79]	Sports-1M	RGB+Flow	91.21	87.64	/
Gammulle [21]	ImageNet	RGB+Flow	94.6	69.0	/
Literatures [45]	Sports-1M	RGB+Flow	91.9	64.1	/
GCA-LSTM [48]	Kinetics400	RGB		66.2	84.0
I3D-LSTM [96]	Kinetics400	RGB+Flow	95.1	/	/
IP-LSTM+IDT [109]	Sports1M	RGB+Flow	91.4	68.2	/
Proposed Method [92]	/	RGB	84.1	/	/
Srivastava [68]					
LiteEval [101]					
VideoLSTM [46]					

4.4. Bone + GCN

Before the large-scale application of skeletal data, most of the data used for video behavior recognition are RGB images obtained by frame extraction of video, but whether using RGB images or extracting optical streams from video, 2D CNN, 3D CNN, and LSTM will generate a large amount of data, and also generate a large amount of redundant information, which makes the training of the model is very difficult, and greatly increases the difficulty of training the model.

This greatly increases the difficulty of model training. However, with the large-scale application of motion capture systems, as well as the application of pose estimation algorithms for RGB videos and depth maps, which can collect a large amount of human skeletal information, the advantages of skeletal data and graph convolutional neural networks mentioned in Section 3.7 have enabled the development of related models. We extend the Skeleton-Kinetics and NTU RGB + D (60/120) skeletal datasets, and summarize the accuracy of 10 relevant models on this dataset, as shown in Table 7.

Table 7
Performance of Bone+GCN and related variants

Model name or method	Ppre-training	Input modal	Accuracy on relevant datasets		
			HMDB51	Skeleton-Kinetics (Top-1/Top-5)	NTU RGB + D(60/120)
PB-GCN [72]	ImageNet	RGB+Skeleton	88.17	-	93.2/-
DGNN [64]	Sports1M	RGB+Skeleton	-	36.9/59.6	96.1/-
AS-GCN [44]	Kinetics-400	RGB+Skeleton	-	-	94.2/-
2S-AGCN [65]	Kinetics-400	RGB+Skeleton	-	36.1/58.7	95.1/-
CA-GCN [113]	Kinetics-400	RGB+Skeleton	-	33.3/55.4	96.0/-
Dynamic-GCN [107]	Kinetics-400	RGB+Skeleton	-	37.9/61.3	96.0/88.6
PointNet++ [58]	-	RGB+Skeleton	-	-	85.1/-
CTR-GCN [10]	Kinetics-400	RGB+Skeleton	-		96.8/90.6
PoseC3D [15]	Kinetics-400	RGB+Skeleton	85.0	-	97.1/90.3
MMNet [108]					

4.5. Transformer and ViT

In recent years, the hot Transformer and ViT have also been applied to the field of video behavior recognition, its scalable self-attention mechanism can naturally learn long-distance spatio-temporal relationships in video, and its internal self-attention blocks and decoder and encoder structures also enable the model to perform better in long-term dependency modeling. Video Swin Transformer extracts spatio-temporal tokens from the input data, and then encodes and decodes them by a series of converters, obtaining the corresponding long sequence of tokens, which is greatly improved by various attention blocks. Subsequent related variant models are also mostly based on this foundation for tinkering and performing variant operations to decompose the spatial and temporal dimensions of the input, and the accuracy of the related variant models on some datasets is shown in Tables 6-8.

Table 8
Transformer and the performance of ViT

Model name or method	Pre-training	Modes	Accuracy on relevant datasets (Average Accuracy / %)			
			UCF101	HMDB51	Kinetics(400/600)	SSv2(Top-1/Top-5)
Video Swin Transformer [51]	ImageNet	RGB	-	-	84.9/85.9	69.6/92.7
ViViT [3]	Sports1M	RGB	82.3	56.8	84.9/85.8	65.9/89.9
TubeViT [57]	ImageNet	RGB	95.6	74.8	90.9/91.8	76.1/95.2
MAR [60]	Kinetics-400	RGB	95.9	74.1	85.3/-	74.7/94.9
VidTr [112]	Kinetics-400	RGB	96.7	74.4	79.1/-	63.0/-
VidConv [55]	Kinetics-400	RGB	-	-	80.5/86.1	65.1/89.6
TPS [102]	Kinetics-400	RGB	-	-	82.5/-	69.8/93.0
InternVideo [84]	Kinetics-400	RGB	91.85	89.3	91.1/91.3	77.2/-
MVD [93]	Kinetics-400	RGB	97.5	79.7	87.2/-	77.3/-
ST-Adapter [56]	CLIP	RGB	-	-	87.2/-	72.3/93.9
TDN [89]	Kinetics-400	RGB	-	-	79.4/-	68.2/91.6
SIFA-Transformer [52]	Kinetics-400	RGB	-	-	83.1/84.5	69.8/93.1

4.6. Performance Comparison of Different Architectural Models

In the previous subsection, the accuracy of different models is mainly compared, but in real-life scenarios, it is not enough to rely on the accuracy alone to deploy the models, in this subsection, we further evaluate some of the above mentioned models from the perspective of mAP, in the RGB and Optical Flow (OF) of the original video as well as other data modalities, we mainly compare the algorithms on UCF101 and HMDB51 data sets. We performed the algorithm comparison mainly on the UCF101 and HMDB51 datasets; Table 9 describes the comparison results of different recognition algorithms in RGB and OF as well as other data modalities, where the mean accuracy percentage (mAP) is used as a criterion for the behavioral recognition algorithms.

Table 9 compares the average accuracy of current mainstream behavior recognition algorithms on UCF101 and HMDB51 datasets based on the model structures of Two-Stream, 3D CNN, CNN-LSTM, and Transformer. The data modalities mainly include RGB and optical flow OF. based on the results in the table, we can see that not all model structures are the best, but the Two-Stream and 3D CNN models are slightly better relative to CNN-LSTM. In each model structure, subsequent studies have made some improvements on the former. For example, Feichtenhofer et al. [20] changed the fusion method of the model based on Simonyan et al. [66], Wang et al. [91] introduced a sparse temporal sampling strategy, and Wang et al. [86] used an end-to-end Faster-RCNN network instead of standard spatial streaming, which further improves the recognition accuracy of the model. Among the RNN-LSTM models, the CNN-LSTM model pioneered by Donahue et al. [14] learns the temporal and spatial streams separately and achieves good results, on which Srivastava et al. [68] maps the input video into a fixed-length representation and decodes it by a decoder, which also further improves the accuracy of the model.

Table 9
mAP of different model structures on several datasets

Model	Model Structure	Input	UCF101 mAP/ %	HMDB51 mAP/ %
Wan et al. [82]	Two-Stream	RGB+OF	87.96	71.71
Simonyan et al. [66]	Two-Stream	RGB+OF	88.0	59.4
Wang et al. [86]	Two-Stream	RGB+OF	88.9	61.3
Wang et al. [88]	Two-Stream	RGB+OF	91.5	64.5
Feichtenhofer et al. [20]	Two-Stream	RGB+OF	92.5	65.4
Wang et al. [91]	Two-Stream	RGB+OF	94.2	69.4
Feichtenhofer et al. [18]	Two-Stream	RGB+OF	94.6	70.3
Tran et al. [75]	3D CNN	RGB	85.8	54.9
Diba et al. [12]	3D CNN	RGB	93.2	63.5
SlowFast [17]	2D CNN	RGB	94.5	68.8
Tran et al. [77]	3D CNN	RGB	97.3	78.7
Donahue et al. [14]	CNN-LSTM	RGB+OF	73.8	68.3
Srivastava et al. [68]	CNN-LSTM	RGB+OF	75.8	44.0
Liu et al. [48]	GCA-LSTM	Skeleton	76.8	71.9
Ng et al. [54]	CNN-LSTM	RGB+OF	88.6	-
Liu et al. [51]	Transformer	RGB	-	67.1
Truong et al. [78]	Transformer	RGB	67.3	68.2

There are also many influencing factors when comparing the performance of the same model structure on different datasets, and different model structures on the same dataset. For example, according to the experimental results of the SlowFast paper, the average accuracy (mAP) of the SlowFast model on the UCF101 dataset is about 94.5%, whereas on the HMDB51 dataset, the average accuracy (mAP) is about 68.8%. Compared with its performance on the UCF101 dataset, its performance on HMDB51 is degraded, probably due to the fact that the video clips in the HMDB51 dataset are shorter and relatively complex, and there are more motion blur and fast movements in the videos, which makes it difficult to recognize them accurately. In contrast, the average accuracy (mAP) of the TDD model on the UCF101 dataset is about 91.5%, which is relatively close to the performance of the TDD model on the UCF101 dataset compared to the SlowFast model. However, the TDD model requires a larger amount of computation and more time and computational resources for training and inference compared to the SlowFast model, and this result is also affected by many other factors, such as hyper-parameter selection, data preprocessing, and so on. Therefore, in specific applications, it is necessary to choose the suitable model according to the task requirements, and adjust and optimize it according to the actual situation.

From the analysis of the previous subsections about the video behavior recognition models with different structures, we can see that the 2D convolution-based video behavior recognition models usually require less number of parameters and are computationally fast, but they are usually ineffective on some similar scenes or video tasks that are very dependent on timing information. 3D convolutional-based recognition models, on the other hand, take timing information into account, but this also introduces a higher number of parameters, which may reduce the training speed and inference speed of the network. There are some Transformer-based video recognition models that have outperformed convolutional network-based video recognition models in terms of accuracy, but in terms of model training, more data is required to train Transformer from scratch compared to CNN. This is because the CNN implementation already implies some a priori knowledge about the image, such as the translation invariance of the image; whereas Transformer requires time-consuming pre-training on a large-scale training dataset to learn these rules. Therefore, when choosing a video behavior recognition model, not only the recognition effect of the model should be considered, but also factors such as the complexity of the model, the amount of computation, and the training requirements need to be taken into account.

5. Future Work

From methods based on traditional manual feature extraction, to the extensive use of deep learning, and now the introduction of the attention mechanism, each development aims to solve the previous problems or improve the efficiency. Traditional manual methods are not only less efficient but also more limited as they require relevant prior knowledge. The emergence of deep learning alleviates this problem by extracting high-dimensional features from raw video data through convolutional operations, which is further combined with temporal-spatial streaming to obtain richer contextual semantic information about the video data and increase the descriptive capability of the model. However, as the depth of the model increases, and the data volume grows, problems such as deep convolutional operations are often difficult to effectively capture continuous time-space information are gradually exposed. In addition, the introduction of the attention mechanism is further addressing this problem. One of the key tasks in video behavior recognition is the video feature extraction method. In addition, video behavior recognition technology faces problems such as illumination, occlusion, and change of viewing angle in practical applications. In the future, video behavior recognition technology needs to further improve the accuracy, robustness and real-time performance, as well as enhance the fusion processing of multimodal data to achieve more accurate and efficient video behavior recognition. In addition, attention needs to be paid to privacy protection to ensure the legality and compliance of video behavior recognition technology.

Possible future research directions and their challenges are listed below:

1 Efficient models: Although existing models have been able to achieve high accuracy, most of them are limited to laboratory settings. This is because

most of the methods are developed offline, which means that the input video data is manually set rather than a random video stream online. In addition, many models are poorly time-sensitive, making it difficult to achieve real-time detection. Therefore, how to design a suitable model becomes a key issue.

2 New datasets: current datasets used for video actions tend to be somewhat subjective, and many video data do not take into account motion on the timeline. For example, some actions can be recognized by only some frames in the video, which makes the video data lose the meaning of time stream. In addition, some video datasets may cause the model to learn some wrong knowledge due to the lack of accuracy of their accompanying annotation files. This calls for a more accurate dataset to orient the research to take more account of temporal issues, fine-grained behavioral issues in human activities, and to improve the ability of the model in modeling specific action information, thus advancing the progress of video behavior recognition.

3 Multimodal features and their fusion: most of the current research on human action recognition has mainly considered visual features in videos, and many of them are only based on frames extracted from videos. However, video behavior recognition has strong temporal correlation. Therefore, multimodal data can be used as an aid for fusion detection, which helps the model to further acquire deep features. How to synergistically utilize the complementarity between multimodal features to select appropriate modal data for training based on the differences in data and the needs of the problem is also a future research focus.

6. Conclusion

In recent years, video behavior recognition technology has been constantly innovated, and its application areas have become more and more extensive, such as automatic driving, intelligent security and other related fields are also increasingly concerned about the development of related technologies. In this paper, we provide a comprehensive overview of the process of video behavior recognition from the relevant dataset, video preprocessing, feature extraction, and then the design of the relevant model, and summarize and analyze the advantages and disadvantages of various methods according to the characteristics of the relevant model. Finally, we summarize some existing problems and possible future research directions in the field of video behavior recognition, hoping to provide subsequent researchers with a better understanding of the current research status in the field of video behavior recognition.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Acknowledgments

The authors would like to thank the faculty for providing excellent research directions and guidance during the writing of the paper, and the research participants for providing support to help the authors complete the paper. All individuals included in this section have agreed to acknowledge.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S. YouTube-8M: A Large-Scale Video Classification Benchmark. ArXiv, 2016. https://doi.org/10.48550/-arXiv.1609.08675
2. Ahn, D., Kim, S., Hong, H., Chul Ko, B. STAR-Transformer: A Spatio-Temporal Cross Attention Transformer for Human Action Recognition. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision

(WACV), 2023, 3319-3328. https://doi.org/10.1109/WACV56688.2023.00333

3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C. ViViT: A Video Vision Transformer. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 6816-6826. https://doi.org/10.1109/ICCV48922.2021.00676

4. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A. Action Recognition with Dynamic Image Networks. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12), 2799-2813. https://doi.org/10.1109/TPAMI.2017.2769085

5. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R. Actions as Space-Time Shapes. Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, 2, 1395-1402. https://doi.org/10.1109/ICCV.2005.28

6. Cai, J., Hu, J. 3D RANs: 3D Residual Attention Networks for Action Recognition. The Visual Comp-uter, 2020, 36, 1261-1270. https://doi.org/10.1007/s00371-019-01733-3

7. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A. A Short Note about Kinetics-600. ArXiv, 2018. https://doi.org/10.48550/arXiv.1808.01340

8. Carreira, J., Noland, E., Hillier, C., Zisserman, A. A Short Note on the Kinetics-700 Human Action Dataset. ArXiv, 2019. https://doi.org/10.48550/arXi-v.1907.06987

9. Carreira, J., Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 IEEE Conference on Computer Vision and Pat-tern Recognition (CVPR), 2017, 4724-4733. https://doi.org/10.1109/CVPR.2017.502

10. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W. Channel-Wise Topology Refinement Graph Con-volution for Skeleton-Based Action Recognition. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 13359-13368. https://doi.org/10.1109/ICCV48922.2021.01311

11. Diba, A., Fayyaz, M., Sharma, V., Arzani, M.M., Yousefzadeh, R., Gall, J., Van Gool, L. Spatio-Temporal Channel Correlation Networks for Action Classification. Computer Vision - ECCV 2018, 2018, 11208, 299-315. https://doi.org/10.1007/978-3-030-01225-0_18

12. Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R., Van Gool, L. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. ArXiv, 2017.

13. Diba, A., Sharma, V., VanGool, L. Deep Temporal Linear Encoding Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, 1541-1550. https://doi.org/10.1109/CVPR.2017.168

14. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4), 677-691. https://doi.org/10.1109/TPAMI.2016.2599174

15. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B. Revisiting Skeleton-Based Action Recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 2959-2968. https://doi.org/10.1109/CVPR52688.2022.00298

16. Feichtenhofer, C. X3D: Expanding Architectures for Efficient Video Recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 200-210. https://doi.org/10.1109/CVPR42600.2020.00028

17. Feichtenhofer, C., Fan, H., Malik, J., He, K. Slowfast Networks for Video Recognition. 2019 IEEE/CVF International Conference on Computer Vision, 2019, 6201-6210. https://doi.org/10.1109/ICCV.2019.00630

18. Feichtenhofer, C., Pinz, A., Wildes, R. P. Spatiotemporal Residual Networks for Video Action Recognition, 2016. https://doi.org/10.1109/CVPR.2017.787

19. Feichtenhofer, C., Pinz, A., Wildes, R. P. Spatiotemp-oral Multiplier Networks for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 7445-7454. https://doi.org/10.1109/CVPR.2017.787

20. Feichtenhofer, C., Pinz, A., Zisserman, A. Convoluti-onal Two-Stream Network Fusion for Video Action Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 1933-1941. https://doi.org/10.1109/CVPR.2016.213

21. Gammulle, H., Denman, S., Sridharan, S., Fookes, C. Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, 177-186. https://doi.org/10.1109/WACV.2017.27

22. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B. ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, 3165-3174. https://doi.org/10.1109/CVPR.2017.337

23. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I.,

Memisevic, R. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, 5843-5851. https://doi.org/10.1109/ICCV.2017.622

24. Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 6047-6056. https://doi.org/10.1109/CVPR.2018.00633

25. Hara, K., Kataoka, H., Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 6546-6555. https://doi.org/10.1109/CVPR.2018.00685

26. He, J. Y., Wu, X., Cheng, Z. Q., Yuan, Z., Jiang, Y. G. DB-LSTM: Densely-Connected Bi-Directional LSTM for Human Action Recognition. Neurocomputing, 2021, 444, 319-331. https://doi.org/10.1016/j.neucom.2020.05.118

27. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

28. Heilbron, F. C., Escorcia, V., Ghanem, B., Niebles, J. C. ActivityNet: A large-scale video benchmark for human activity understanding. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 961-970. https://doi.org/10.1109/CVPR.2015.7298698

29. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E. Squeeze-and-Excitation Networks.In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 42(8), 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

30. Huang, Y., Guo, Y., Gao, C. Efficient Parallel Inflated 3D Convolution Architecture for Action Recognition. In IEEE Access, 2020, 8, 45753-45765. https://doi.org/10.1109/ACCESS.2020.2978223

31. Ioffe, S., Szegedy, C. Batch Normalization: Acceler-ating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015, 37, 448-456.

32. Jain, V., Gupta, G., Gupta, M., Sharma, D.K., Ghosh, U. Ambient Intelligence-Based Multimodal Human Action Recognition for Autonomous Systems. ISA Transactions, 2023, 132, 94-108. https://doi.org/10.1016/j.isatra.2022.10.034

33. Ji, S., Xu, W., Yang, M., Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1), 221-231. https://doi.org/10.1109/TPAMI.2012.59

34. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J. STM: Spatiotemporal and Motion Encoding for Action Recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 2000-2009. https://doi.org/10.1109/ICCV.2019.00209

35. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A. The Kinetics Human Action Video Dataset. ArXiv, 2017. https://doi.org/10.48550/arXiv.1705.06950

36. Kar, A., Rai, N., Sikka, K., Sharma, G. AdaScan: Adaptive Scan Pooling in Deep Convolutional Neural Networks for Human Action Recognition in Videos. 2017 IEEE Conference on Computer Vision and Pat-tern Recognition (CVPR), 2017, 5699-5708. https://doi.org/10.1109/CVPR.2017.604

37. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.F. Large-Scale Video Classification with Convolutional Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, 1725-1732. https://doi.org/10.1109/CVPR.2014.223

38. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T. HMDB: A large video database for human motion recognition. 2011 International Conference on Computer Vision, 2011, 2556-2563. https://doi.org/10.1109/ICCV.2011.6126543

39. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B. Learning Realistic Human Actions from Movies. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, 1-8. https://doi.org/10.1109/CVPR.2008.4587756

40. Li, A., Thotakuri, M., Ross, D. A., Carreira, J., Vostr-ikov, A., Zisserman, A. The AVA-Kinetics Localized Human Actions Video Dataset. ArXiv, 2020.

41. Li, C., Zhong, Q., Xie, D., Pu, S. Collaborative Spatiotemporal Feature Learning for Video Action Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 7864-7873. https://doi.org/10.1109/CVPR.2019.00806

42. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L. TEA: Temporal Excitation and Aggregation for Action Recognition. 2020 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), 2020, 906-915. https://doi.org/10.1109/CVPR42600.2020.00099

43. Li, J., Liu, X., Zhang, M., Wang, D. Spatio-Temporal Deformable 3D ConvNets with Attention for Action Recognition. Pattern Recognit, 2020, 98. https://doi.org/10.1016/j.patcog.2019.107037

44. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 3590-3598. https://doi.org/10.1109/CVPR.2019.00371

45. Li, Q., Qiu, Z., Yao, T., Mei, T., Rui, Y., Luo, J. Action Recognition by Learning Deep Multi-Granular Spatio-Temporal Video Representation. Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 2016, 159-166. https://doi.org/10.1145/2911996.2912001

46. Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.M. VideoLSTM Convolves, Attends and Flows for Action Recognition. Computer Vision and Image Understanding, 2018, 166, 41-50. https://doi.org/10.1016/j.cviu.2017.10.011

47. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., Kot, A. C. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10), 2684-2701. https://doi.org/10.1109/TPAMI.2019.2916873

48. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C. Global Context-Aware Attention LSTM Networks for 3D Action Recognition. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 3671-3680. https://doi.org/10.1109/CVPR.2017.391

49. Liu, Q., Che, X., Bie, M. R-STAN: Residual Spatial-Temporal Attention Network for Action Recognition. In IEEE Access, 2019, 7, 82246-82255. https://doi.org/10.1109/ACCESS.2019.2923651

50. Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., Lu, T. TEINet: Towards an Efficient Architecture for Video Recognition. Proc-eedings of the AAAI conference on artificial intell-igence, 2020, 11669-11676. https://doi.org/10.1609/aaai.v34i07.6836

51. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H. Video Swin Transformer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recogn-ition (CVPR), 2022, 3192-3201. https://doi.org/10.1109/CVPR52688.2022.00320

52. Long, F., Qiu, Z., Pan, Y., Yao, T., Luo, J., Mei, T. Stand-Alone Inter-Frame Attention in Video Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 3182-3191. https://doi.org/10.1109/CVPR52688.2022.00319

53. Luo, C., Yuille, A. Grouped Spatial-Temporal Aggregation for Efficient Action Recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 5511-5520. https://doi.org/10.1109/ICCV.2019.00561

54. Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G. Beyond Short Snippets: Deep Networks for Video Classification. 2015 IEEE Conference on Computer Vision and Pat-tern Recognition (CVPR), 2015, 4694-4702. https://doi.org/10.1109/CVPR.2015.7299101

55. Nguyen, C.H., Huynh, S., Nguyen, V., Nguyen CyberCore, N.A., Chi Minh, H., Nam, V. VidConv: A Modernized 2D ConvNet for Efficient Video Recognition. ArXiv, 2022. https://doi.org/10.48550/arXiv.2207.03782

56. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H. ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning. Neural Information Processing Systems, 2022.

57. Piergiovanni, A.J., Kuo, W., Research, G., Angelova, A. Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recogn-ition (CVPR), 2023, 2214-2224. https://doi.org/10.1109/CVPR52729.2023.00220

58. Qi, C.R., Yi, L., Su, H., Guibas, L. J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 5099-5108.

59. Qi, Y., Hu, J., Han, X., Hu, L., Zhao, Z. MFGCN: An Efficient Graph Convolutional Network Based on Multi-Order Feature Information for Human Skeleton Action Recognition. Neural Computing and Applications, 2023, 35, 19979-19995. https://doi.org/10.1007/s00521-023-08814-4

60. Qing, Z., Zhang, S., Huang, Z., Wang, X., Wang, Y., Lv, Y., Gao, C., Sang, N. MAR: Masked Autoencoders for Efficient Action Recognition. IEEE Transactions on Multimedia, 2024, 26, 218-233. https://doi.org/10.1109/TMM.2023.3263288

61. Qiu, Z., Yao, T., Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. 2017 IEEE International Conference on Computer Vision

(ICCV), 2017, 5534-5542. https://doi.org/10.1109/ICCV.2017.590

62. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

63. Shahroudy, A., Liu, J., Ng, T. T., Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, 1010-1019. https://doi.org/10.1109/CVPR.2016.115

64. Shi, L., Zhang, Y., Cheng, J., Lu, H. Skeleton-Based Action Recognition with Directed Graph Neural Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 7904-7913. https://doi.org/10.1109/CVPR.2019.00810

65. Shi, L., Zhang, Y., Cheng, J., Lu, H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 12018-12027. https://doi.org/10.1109/CVPR.2019.01230

66. Simonyan, K., Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, 1, 568-576.

67. Soomro, K., Zamir, A.R., Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. ArXiv, 2012. https://doi.org/10.48550/arXiv.1212.0402

68. Srivastava, N., Mansimov, E., Salakhutdinov, R. Unsupervised Learning of Video Representations Using LSTMs. Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015, 37, 843-852.

69. Stroud, J., Ross, D., Sun, C., Deng, J., Sukthankar, R. D3d: Distilled 3d Networks for Video Action Recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, 614-623. https://doi.org/10.1109/WACV45572.2020.9093274

70. Sun, L., Jia, K., Yeung, D.-Y., Shi, B. E. Human Action Recognition Using Factorized Spatio-Temporal Convol-utional Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 2015, 4597-4605. https://doi.org/10.1109/ICCV.2015.522

71. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. Going Deeper with Convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

72. Thakkar, K., Narayanan, P. J. Part-Based Graph Convolutional Network for Action Recognition. BMVC 2018, 2018, 270.

73. Tong, Z., Song, Y., Wang, J., Wang, L. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 2022, 10078-10093.

74. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

75. Tran, D., Ray, J., Shou, Z., Chang, S.-F., Paluri, M. ConvNet Architecture Search for Spatiotemporal Fe-ature Learning. ArXiv, 2017.

76. Tran, D., Wang, H., Feiszli, M., Torresani, L. Video Classification with Channel-Separated Convolutional Networks. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 5551-5560. https://doi.org/10.1109/ICCV.2019.00565

77. Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y., Paluri, M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 6450-6459. https://doi.org/10.1109/CVPR.2018.00675

78. Truong, T. D., Bui, Q. H., Duong, C. N., Seo, H. S., Phung, S. L., Li, X., Luu, K. Direcformer: A Directed Attention in Transformer Approach to Robust Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 19998-20008. https://doi.org/10.1109/CVPR52688.2022.01940

79. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W. Action Recognition in Video Sequences Using Deep Bi-Directional LSTM with CNN Features. In IEEE Access, 2017, 6, 1155-1166. https://doi.org/10.1109/ACCESS.2017.2778011

80. Varol, G., Laptev, I., Schmid, C. Long-Term Temporal Convolutions for Action Recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6), 1510-1517. https://doi.org/10.1109/TPAMI.2017.2712608

81. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. Attention Is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 6000-6010.

82. Wan, Y., Yu, Z., Wang, Y., Li, X. Action Recognition Based on Two-Stream Convolutional Networks with Long-Short-Term Spatiotemporal Features. In IEEE Access, 2020, 8, 85284-85293. https://doi.org/10.1109/ACCESS.2020.2993227

83. Wang, H., Schmid, C. Action Recognition with Improved Trajectories. 2013 IEEE International Conference on Computer Vision, 2013, 3551-3558. https://doi.org/10.1109/ICCV.2013.441

84. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y. InternV-ideo: General Video Foundation Models via Generative and Discriminative Learning. ArXiv, 2022. https://doi.org/10.48550/arXiv.2212.03191

85. Wang, Y., Long, M., Wang, J., Yu, P. S. Spatiotemporal Pyramid Network for Video Action Recognition. In Proceedings of the Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 2097-2106. https://doi.org/10.1109/CVPR.2017.226

86. Wang, Y., Song, J., Wang, L., Van Gool, L., Hilliges, O. Two-Stream SR-CNNs for Action Recognition in Videos. British Machine Vision Conference, 2016. https://doi.org/10.5244/C.30.108

87. Wang, L., Ge, L., Li, R., Fang, Y. Three-Stream CNNs for Action Recognition. Pattern Recognition Letters, 2017, 92, 33-40. https://doi.org/10.1016/j.patrec.2017.04.004

88. Wang, L., Qiao, Y., Tang, X. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, 4305-4314. https://doi.org/10.1109/CVPR.2015.7299059

89. Wang, L., Tong, Z., Ji, B., Wu, G. TDN: Temporal Difference Networks for Efficient Action Recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 1895-1904. https://doi.org/10.1109/CVPR46437.2021.00193

90. Wang, L., Xiong, Y., Wang, Z., Qiao, Y. Towards Good Practices for Very Deep Two-Stream ConvNets. ArXiv, 2015. https://doi.org/10.48550/arXiv.1507.02159

91. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., van Gool, L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. Computer Vision. ECCV 2016, 2016, 9912, 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

92. Wang, L., Xu, Y., Cheng, J., Xia, H., Yin, J., Wu, J. Human Action Recognition by Learning Spatio-Temporal Features with Deep Neural Networks. In IEEE Access, 2018, 6, 17913-17922. https://doi.org/10.1109/ACCESS.2018.2817253

93. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., Jiang, Y.-G. Masked Video Distillation: Rethinking Masked Feature Modeling for Self-Sup-ervised Video Representation Learning. 2023 IEE-E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 6312-6322. https://doi.org/10.1109/CVPR52729.2023.00611

94. Wang, X., Farhadi, A., Gupta, A. Actions ~ Transformations. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 2658-2667. https://doi.org/10.1109/CVPR.2016.291

95. Wang, X., Girshick, R., Gupta, A., He, K. Non-Local Neural Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

96. Wang, X., Miao, Z., Zhang, R., Hao, S. I3D-LSTM: A New Model for Human Action Recognition. IOP Conference Series: Materials Science and Engineering, 2019, 569(3). https://doi.org/10.1088/1757-899X/569/3/032035

97. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. CBAM: Convolutional Block Attention Module. Computer Vision. ECCV 2018, 2018, 11211, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

98. Wu, C. Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C. MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 13577-13587. https://doi.org/10.1109/CVPR52688.2022.01322

99. Wu, Z., Fu, Y., Jiang, Y.G., Sigal, L. Harnessing Object and Scene Semantics for Large-Scale Video Understanding. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 3112-3121. https://doi.org/10.1109/CVPR.2016.339

100. Wu, Z., Jiang, Y.G., Wang, X., Ye, H., Xue, X. Multi-Stream Multi-Class Fusion of Deep Networks for Video Classification. Proceedings of the 24th ACM International Conference on Multimedia, 2016, 791-800. https://doi.org/10.1145/2964284.2964328

101. Wu, Z., Li, H., Zheng, Y., Xiong, C., Jiang, Y.G., Davis, L.S. A Coarse-to-Fine Framework for Resource Efficient Video Recognition. International Journal of Computer Vision, 2021, 129, 2965-2977. https://doi.org/10.1007/s11263-021-01508-1

102. Xiang, W., Li, C., Wang, B., Wei, X., Hua, X.S., Zhang, L. Spatiotemporal Self-Attention Modeling with Temporal Patch Shift for Action Recognition. Computer Vision. ECCV 2022, 2022, 13663, 627-644. https://doi.org/10.1007/978-3-031-20062-5_36

103. Xiao, F., Lee, Y. J., Grauman, K., Malik, J., Feichtenhofer, C. Audiovisual SlowFast Networks for Video Recognition, ArXiv, 2020.

104. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification. Computer Vision. ECCV 2018, 2018, 11219, 318-335. https://doi.org/10.1007/978-3-030-01267-0_19

105. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C. Multiview Transformers for Video Recognition. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022, 3323-3333. https://doi.org/10.1109/CVPR52688.2022.00333

106. Yang, W., Chen, Y., Huang, C., Gao, M. Video-Based Human Action Recognition Using Spatial Pyramid Pooling and 3D Densely Convolutional Networks. Future Internet, 2018, 10(12), 115. https://doi.org/10.3390/fi10120115

107. Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., Tang, H. Dynamic GCN: Context-Enriched Topology Learning for Skeleton-Based Action Recognition. Proceedings of the 28th ACM International Conference on Multimedia, 2020, 55-63. https://doi.org/10.1145/3394171.3413941

108. Yu, B. X. B., Liu, Y., Zhang, X., Zhong, S. H., Chan, K. C. C. MMNet: A Model-Based Multimodal Network for Human Action Recognition in RGB-D Videos. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3), 3522-3538. https://doi.org/10.1109/TPAMI.2022. 3177813

109. Yu, S., Xie, L., Liu, L., Xia, D. Learning Long-Term Temporal Features with Deep Neural Networks for Human Action Recognition. In IEEE Access, 2020, 8, 1840-1850. https://doi.org/10.1109/ACCESS.2019.2962284

110. Zeiler, M. D., Fergus, R. Visualizing and Understanding Convolutional Networks. Computer Vision. ECCV 2014, 2014, 8689, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53

111. Zhang, Y., Hao, K., Tang, X., Wei, B., Ren, L. Long-Term 3D Convolutional Fusion Network for Action Recognition. 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2019, 216-220. https://doi.org/10.1109/ICAICA.2019.8873471

112. Zhang, Y., Li, X., Liu, C., Shuai, B., Zhu, Y., Brattoli, B., Chen, H., Marsic, I., Tighe, J. VidTr: Video Transformer Without Convolutions. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 13557-13567. https://doi.org/10.1109/ICCV48922.2021.01332

113. Zhang, X., Xu, C., Tao, D. Context Aware Graph Convolution for Skeleton-Based Action Recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 14321-14330. https://doi.org/10.1109/CVPR42600.2020.01434

114. Zhao, H., Jin, X. Human Action Recognition Based on Improved Fusion Attention CNN and RNN. 2020 5th International Conference on Computational Intelligence and Applications (ICCIA), 2020, 108-112. https://doi.org/10.1109/ICCIA49625.2020.00028

115. Zhou, Y., Sun, X., Zha, Z.J., Zeng, W. MiCT: Mixed 3D/2D Convolutional Tube for Human Action Reco-gnition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 449-458. https://doi.org/10.1109/CVPR.2018.00054

116. Zhu, Y., Lan, Z., Newsam, S., Hauptmann, A. Hidden Two-Stream Convolutional Networks for Action Recognition. Computer Vision. ACCV 2018, 2019, 11363, 363-378. https://doi.org/10.1007/978-3-030-20893-6_23

117. Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., Li, M. A Comprehensive Study of Deep Video Action Recog-nition. ArXiv, 2020. https://doi.org/10.48550/arXi-v.2012.06567

118. Zhu, X., Huang, P.-Y., Liang, J., de Melo, C.M., Hauptmann, A. STMT: A Spatial-Temporal Mesh Transformer for MoCap-Based Action Recognition. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 1526-1536. https://doi.org/10.1109/CVPR52729.2023.00153