Artificial Intelligence Review
https://doi.org/10.1007/s10462-022-10210-8

Overview of behavior recognition based on deep learning

Kai Hu[1,2] · **Junlan Jin**[1,2] · **Fei Zheng**[2,3] · **Liguo Weng**[1,2] · **Yiwu Ding**[1,2]

Abstract
Human behavior recognition has always been a hot spot for research in computer vision. With the wide application of behavior recognition in virtual reality and short video in recent years and the rapid development of deep learning algorithms, behavior recognition algorithms based on deep learning have emerged. Compared with traditional methods, behavior recognition algorithms based on deep learning have the advantages of strong robustness and high accuracy. This paper systemizes and introduces behavior recognition algorithms based on deep learning proposed in recent years, then focuses on a series of behavior recognition algorithms based on image and bone data; deeply analyzes their theories and performance, and finally, puts forward further prospects.

Keywords Behavior recognition · Deep learning · Skeleton data

Junlan Jin, Fei Zheng, Liguo Weng and Yiwu Ding have contributed equally to this work.

✉ Kai Hu
001600@nuist.edu.cn

Junlan Jin
20201249090@nuist.edu.cn

Fei Zheng
20181223108@nuist.edu.cn

Liguo Weng
002311@nuist.edu.cn

Yiwu Ding
20191223014@nuist.edu.cn

[1] School of Automation, Nanjing University of Information Science and Technology, No.219, Ningliu Road, 210044 Nanjing, Jiangsu, China

[2] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, No.219, Ningliu Road, 210044 Nanjing, Jiangsu, China

[3] Innovation Department of Industrial Internet, China Telecom Ningbo Branch, No.96 HeYi Road, 315000 Ningbo, Zhejiang, China

Published online: 21 June 2022

1 Introduction

With the wide application of virtual reality technology (Liu et al. 2021), human-computer interactions, intelligent security (Li et al. 2021), and other areas in real life (Li et al. 2021; Shen et al. 2019), human behavior recognition, which occupies a pivotal position in the field of computer vision, has become an important research topic (Liu et al. 2018, 2019). The purpose of the technology is to capture ongoing behavior in video or image sequences and judge its type.

At present, the research methods of behavior recognition are divided into two categories, one based on manual feature extraction (Yang and Tian 2014; Peng et al. 2014, 2016; Arandjelovic and Zisserman 2013; Duta et al. 2017) and the other on deep network learning features (Chen et al. 2021; Xia et al. 2021a; Xia et al. 2020a; Xia et al. 2021b). The method based on manual feature extraction often adopts the traditional machine learning method. Its advantages lie in need-based orientation, strong pertinence, and simple implementation. However, due to noise such as lighting, action similarity (jogging and running), and dynamic backgrounds in behavior recognition (Hong-Lei et al. 2018), the features extracted manually cannot meet the subsequent classification tasks. Therefore, the effect of manual feature extraction on behavior recognition is not very significant. Among the algorithms, improved dense trajectories (iDT) has the highest reliability, but its calculation speed is prolonged and it cannot meet real-time requirements.

In recent years, deep learning has been intensely developed and applied in many fields, such as image classification (Liu et al. 2021) and natural language processing (Li et al. 2021; Zhang et al. 2020). Because the principle of deep learning is to use neurons to simulate human thinking and other activities, which has the same mechanism as behavior recognition (Yi-Ming and Xiang 2018), researchers try to use deep learning to solve the problem of behavior recognition. The mainstream behavior recognition methods based on deep learning are two-stream convolutional networks (Simonyan and Zisserman 2014), 3D convolutional networks (C3D) (Tran et al. 2015), and long short-term memory (LSTM) recurrent networks (Donahue et al. 2015). In general, the advantage of behavior recognition methods based on deep learning is that they can effectively and automatically learn features from the data and realize the end-to-end learning process. The features known by these methods are more comprehensive, but the disadvantage is that they require a lot of data for training.

The behavior recognition methods based on mainstream deep learning algorithms have achieved fruitful research results. In many practical application scenarios, the data are generated from non-Euclidean space, and the human skeleton map also comprises irregular data. Skeleton features can be acquired through human body posture detection algorithms or high-precision depth cameras. The joints in the human body, connected by bone points, naturally form a graph structure. Using GCN to deal with human bone data has achieved good results. Therefore, the research of behavior recognition based on GCN is of great significance.

A recent review, titled "A Comprehensive Study of Deep Video Action Recognition" (Zhu et al. 2020), is a comprehensive examination of video behavior recognition papers based on deep learning. However, it does not record the latest paper and only introduces video imaging. Therefore, the main contributions of this paper are as follows: (1) We review the traditional behavior recognition algorithms based on machine learning. (2) We systematically introduce deep learning algorithms for behavior recognition based on image and skeleton, and explain the classical papers and the latest progress in detail. (3)

We describe the challenges in this field and the prospects for future research. Our work can help researchers quickly step into this field.

A more classic network is the two-stream convolutional network, proposed by Simonyan et al. (Simonyan and Zisserman 2014), which separates spatial stream from temporal stream features for learning and finally integrates them. This approach makes up for the lack of temporal stream features in traditional methods. Aiming at the problem that the two-stream convolutional network cannot model long-range temporal sequences, Wang et al. 2016 proposed combining that network with a uniform sparse sampling method to model the entire video sequence. Based on (Wang et al. 2016), Lan et al. 2017 added weights to the segments of the fusion part. Zhao and Snoek 2019 proposed an effective two-in-one stream network for spatial and temporal behavior detection, which solves the problems of large model size and computational complexity. Not all temporal and space dimensions have the same direction, and they cannot be treated equally. Therefore, Feichtenhofer et al. 2019 proposed a SlowFast network for video recognition to deal with them separately. Xiao et al. 2020b proposed a new architecture to process visual and audio signals. For audio signals of different lengths, Kazakos et al. 2021 suggested two processes of auditory recognition for shunt processing.

After the two-stream convolutional network was proposed, Tran et al. 2015 were inspired by its spatial and temporal features and proposed the 3D convolutional network (C3D), in which the temporal dimension is added to learn spatial and temporal features based on 2D convolution. To effectively extract dynamic human posture features in time series, Yan et al. 2019 proposed a concise pose-action 3D machine. To solve the problem of insufficient video processing capacity for long durations, Ren et al. 2021 proposed neural architecture search-temporal convolution. Because a 3D convolution network needs large computing resources and memory, Kondratyuk et al. 2021 proposed a three-step method to solve this problem. Yue-Hei Ng et al. 2015 combined the structure of the classical temporal series processing networks, LSTM network and two-stream convolutional network, to learn temporal features. The above approaches have generally achieved high accuracy in behavior recognition.

The behavior recognition approach based on skeleton data has been widely studied because of its strong adaptability to dynamic environments and complex backgrounds. The graph convolutional network (GCN), which extends convolution from images to graphs, has been successfully used in many applications. For the task of action recognition based on skeletons, Yan et al. 2018 first applied GCN to model skeleton data and constructed the spatio-temporal graph convolution network (ST-GCN). A novel two-stream adaptive graph convolution network (2s-AGCN) was proposed by Shi et al. 2019b. The topological graph of skeletal joints of this model can be learned adaptively by the BP algorithm. This data-driven method increases the flexibility of the model in graph construction. Li et al. 2019 proposed the actional-structural graph convolution network (AS-GCN) to mine potential joint correlations using a high-order adjacency matrix to obtain the physical structure between connected joints. Shi et al. 2019a suggested that a directed graph neural network (DGNN) could extract not only the information of joints and bones, but also directional correlation information between them. Li et al. 2019 proposed a spatio-temporal graph routing (STGR) scheme for skeleton-based action recognition, which adaptively learns intrinsic high-order connectivity relationships for physically distant skeletal joints. Si et al.2019 proposed an attention enhanced graph convolutional LSTM network (AGC-LSTM) to effectively extract distinctive spatio-temporal features. In the fifth section, the paper introduces behavior recognition algorithm based on GCN in detail.

Section 1 introduces the background of behavior recognition, Sect. 2 introduces the databases commonly used by researchers in this field, and Sect. 3 introduces the most effective iDT algorithm and its predecessor DT algorithm, as well as its sampling features and coding methods, which were the most effective methods before deep learning was applied to the field of behavior recognition. In Sect. 4, the image-based deep learning behavior recognition algorithms are summarized. This paper focuses on the two-stream convolutional network, 3D convolution network, and LSTM network and briefly introduces other networks. In Sect. 5, the deep learning algorithm of behavior recognition based on skeleton data is presented in detail. This section includes a variety of graph network models and algorithms, summarizes the critical aspects of network construction, and compares the performance of different algorithms on large-scale databases. Sect. 6 summarizes the network mentioned in Sects. 4 and 5 and discusses future research in behavior recognition. Figure 1 shows the overall framework of this paper.

2 Database

Research teams usually use human action databases to detect the accuracy and robustness of algorithms. A summary of datasets for behavior recognition is shown in Table 1. Databases play at least two critical roles: (1) when using them, researchers do not need to care about the process of acquisition and pretreatment, and (2) the performance of different

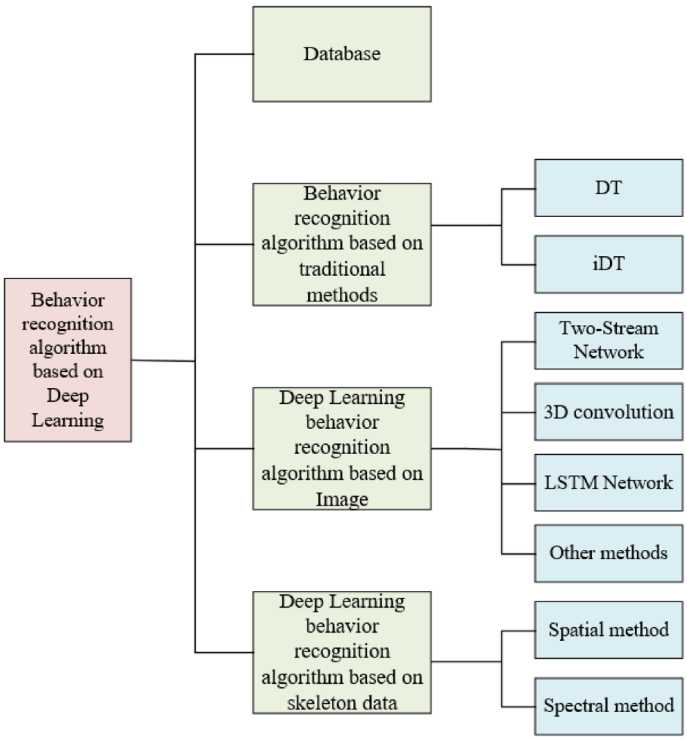

Fig. 1 Overall block diagram

Table 1 Introduction to common databases

Database	Year	Behavior category	Number of videos	Brief introduction	Resources
HDM05	2007	130	2337	Contains more than 3 hours of system recording and well-recorded motion capture data in C3D and ASF/AMC data formats.	http://resources.mpi-inf.mpg.de/HDM05/
UCF sports	2008	10	150	Derived from motion samples of mobile phones from BBC/ESPN radio and television channels.	www.crcv.ucf.edu/data/UCF_Sports_Action.php
UCF YouTube	2008	11	1600	Comes from video clips on YouTube.	www.crcv.ucf.edu/data/UCF_YouTube_Action.php
Northwestern-UCLA	2008	8	663	Contains 8 types of actions collected from 32 movies.	–
Hollywoods2	2009	12	3669	Extended version of Hollywood database, containing 12 behavior categories and 10 scenes extracted from 69 movies, close to situations in actual scenes.	www.di.ens.fr/~laptev/actions/hollywood2/
HMDB-51	2011	51	6849	Most videos come from movies, some from public data and YouTube.	Serrelab.clps.brown. edu/resource/hmdb-a-large-human-moti on-database/dataset
UCF-101	2012	101	13320	Contains 5 action categories, 101 specific small actions, and many action categories and samples.	www.crcv.ucf.edu/data/UCF101.php
Sports-1M	2014	487	1133158	Data come from sports videos collected on YouTube.	cs.stanford.edu/peo ple/karpathy/deepvideo
NTU-RGB+D	2016	60	56880	Used for 3D human activity analysis; consists of 56,880 RGB+D video samples captured from 40 human objects using Microsoft Kinect v2.	https://github.com/Hrener/3D-Action-recognition
Kinetics	2017	600	500000	Data come from YouTube and cover a wide range of actions, including human interaction, such as playing a musical instrument, shaking hands, hugging, etc.	deepmind.com/res earch/open-source /kinetics
NTU-RGB+D 120	2019	120	114480	Uses Microsoft Kinect v2 to collect RGB videos, depth sequences, skeleton data (3D positions of 25 major skeleton joints), and infrared frames.	http://rose1.ntu.edu.sg/Datasets/actionRecognition.asp

algorithms can be detected and compared under the same standard. The University of Central Florida released the UCF-101 database (Liu, 2017) in 2012. The samples in the dataset include various motion samples collected from TV broadcasts and video samples downloaded and saved from YouTube. There are five types of significant actions (human interaction, human interaction, body movement, musical instrument playing), 101 kinds of specific small activities, and 13,320 videos.

Brown University released the HMDB-51 database (Liu, 2017) in 2011. The video samples are from video clips on YouTube. There are 51 types of sample actions and 6849 videos. Each type of sample action contains at least 101 videos.

The above two datasets have large numbers of samples and complex backgrounds, which can test the accuracy of algorithms and detect their robustness. At present, the UCF-101 and HMDB-51 datasets are widely used in behavior recognition based on deep learning.

NTU RGB + D contains 60 action categories and 56,880 video samples. NTU RGB + D120 extends NTU RGB + D by adding 60 types and 57,600 video samples; that is, NTU RGB + D120 has a total of 120 categories and 11,4480 samples. Both datasets contain RGB videos, depth map sequences, 3D skeleton data, and infrared (IR) videos for each sample. Three Kinect V2 cameras are used simultaneously to capture each sample. The resolution of the RGB video is 1920×1080, the resolution of the depth map and the IR video is 512×424, and the 3D skeleton data contain the 3D coordinates of 25 human joints per frame.

In the above two datasets, more attention is paid to position changes of humans in action, so they are more suitable for the recently emerging behavior recognition algorithms based on skeleton detection.

3 Behavior recognition algorithm based on traditional methods

Behavior recognition is mainly divided into two methods: manual feature extraction and deep learning.

Although the performance of behavior recognition algorithms based on deep learning has surpassed that of dense trajectories (DT) and iDT algorithms, they have had a significant impact on development in this field. Many methods that achieve good performance adopt the idea of combining deep learning with the iDT algorithm, so DT and iDT algorithms are indispensable in behavior recognition. The iDT algorithm improves the DT algorithm, so the DT algorithm is introduced first.

3.1 DT algorithm

The basic framework of the DT algorithm is shown in Fig. 2 (Wang et al. 2013). The specific process is as follows:

Step 1. Each frame of the video is divided into multiple scales, the feature points of each scale are densely sampled by gridding, and some untraceable feature points are removed.

Step 2. The feature points are tracked to obtain the trajectory in the video sequence. The positions of a feature point on the continuous L frame image constitute a trajectory, and the tracking of feature points is carried out independently on each scale. In order to avoid the drift phenomenon, features should be recollected and retracted every L frames.

Step 3. Along the trajectory of a feature point with length L, a time–space volume is formed in the $N \times N$ region around the feature point in each image. The time–space volume

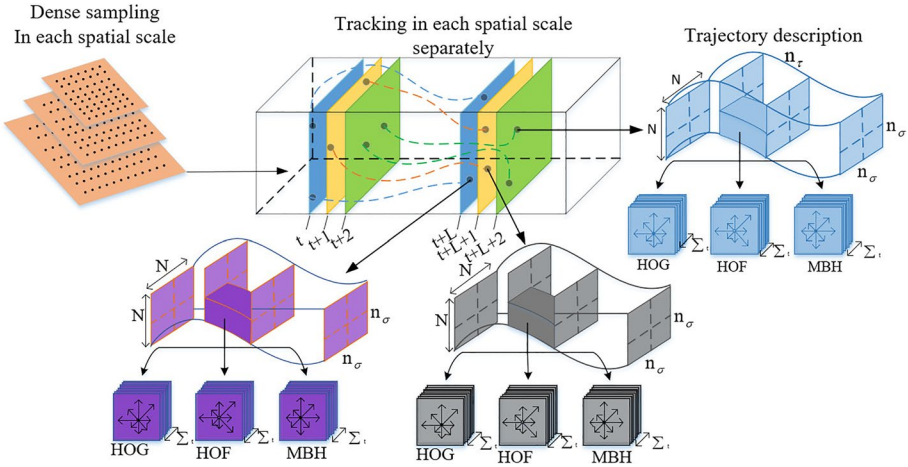

Fig. 2 Basic framework of DT algorithm (Wang et al. 2013)

is divided into $n_\sigma \times n_\sigma \times n_\tau$ block regions for feature extraction. The HOG feature is used to calculate the grayscale image gradient histogram. The HOF feature is used to calculate the optical flow information (direction and amplitude) histogram. The MBH feature is used to calculate the optical flow image gradient histogram. The L2 norm is used to normalize the features of HOG, HOF, and MBH.

Step 4. There are many tracks in a video, and each track corresponds to a set of features. Each feature group is encoded by a bag of features to get a certain length of feature coding for video classification.

Step 5. The RBF-x^2 core and the SVM trained by the one-to-many strategy are used to classify videos.

3.2 iDT algorithm

The iDT algorithm is an improved version of the DT algorithm, and its general process and framework are similar. Some features and noise are processed to enhance the performance of the algorithm:

(1) (Wang and Schmid 2013): The optical flow and trajectory in the background are eliminated by estimating the camera motion. Since the changes between two adjacent frames are small, the latter frame is assumed to be obtained by the projection transformation of the previous frame; as a result, the problem of camera motion estimation can be approximated to the problem of projection transformation matrix calculation by using the front and back frame images. The SURF and optical flow features are used to match feature points between images. The projection transformation matrix is estimated by the random sample consensus (RANSAC) algorithm.

(2): Because human action is more significant in the image, the projection matrix is not accurate enough to estimate the matching points of the human body. Therefore, the iDT algorithm uses the human body detector to detect the human position frame and remove the matching point pairs in the frame. The human action will not affect the estimation of the projection matrix.

In addition to these two improvements, the iDT algorithm also uses the L2 norm for feature normalization and the Fisher vector for feature coding to improve its accuracy and speed.

The main improvements of the iDT algorithm over the DT algorithm are its optimization of optical flow images and new ways of carrying out feature regularization and coding. These significantly improve the effectiveness, and the accuracy of the HMDB-51 dataset is increased from 46.6 to 57.2%.

The iDT algorithm specializes in feature extraction. Compared with deep learning methods, the iDT algorithm can extract behavior-related features more accurately, so it has greater stability. According to the advantages described above, deep learning algorithms can be effectively improved by combining them with the iDT.

4 Deep Learning behavior recognition algorithm based on Image

The main networks used for behavior recognition in deep learning are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Classical CNNs, such as AlexNet, VGG16, etc., achieve different feature extraction effects through different arrangements and combinations of convolution layers, pooling layers, and fully connected layers. The recurrent neural network takes sequence data as input and recurses in the changing direction of the sequence; additionally, the recurrent units have a chain-type connection (Goodfellow et al. 2016). CNN establishes weight connections between layers. RNN also establishes weight connections between neurons in layers, and its output is related to both the current input sequence and the previous output.

Most of the existing behavior recognition algorithms based on deep learning were developed based on two-stream convolutional networks, 3D convolutional networks, and RNN (especially LSTM). Following the continuous development of deep learning algorithms in the past two years, this paper divides behavior recognition algorithms based on deep learning into four categories: two-stream convolutional networks, 3D convolutional networks, LSTM networks, and other networks. Based on this classification, these algorithms are introduced as follows.

4.1 Deep Learning algorithm of behavior recognition based on Two-Stream Convolutional Network

Simonyan and Zisserman 2014 proposed the basic two-stream ConvNet architecture. As shown in Fig. 3, the network has two parallel networks of a spatial stream and a temporal stream (Huilan et al. 2018). It uses two independent CNN networks to process spatial and temporal information in a video separately. The input of the spatial stream network is a single frame image sampled from the video, and the input of the temporal stream network is optical flow information. The model combines the two networks to get the final recognition result. The network achieved 88 and 59.4% accuracy on the UCF-101 and HMDB-51 databases, respectively. The network model has good expansibility and has attracted the attention of researchers. Therefore, many improved algorithms have emerged with a focus on accuracy and robustness.

The temporal segment network (TSN), proposed by Wang et al. (Wang et al. 2016) adopts the uniform sparse sampling method to obtain information with long-range temporal structure, which can be applied to long video sequences. The network divides the

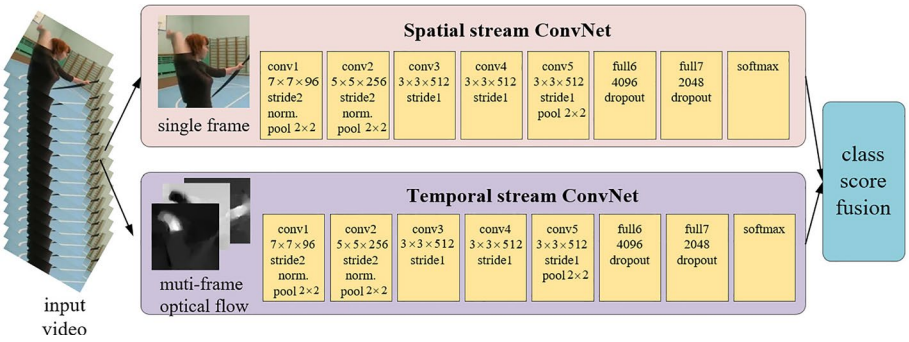

Fig. 3 Workflow of two-stream convolutional network (Simonyan and Zisserman 2014)

video into K segments, inputs each segment into the two-stream network, and then gets the classification result. It uses the weighted-average method to fuse all influences to get the final product, which overcomes the problem that the classical two-stream network can only handle short-term videos. Diba An et al. 2017 adopted another fusion method, proposing a temporal linear encoding (TLE) layer to fuse and encode the features extracted after video segmentation. It captures the interaction between all spatial features to get the dynamic process over a long time and can learn the convolution neural network model using limited samples.

The two-stream detection network based on RGB and flow provides state-of-the-art accuracy at the expense of large model size and heavy computation. Zhao and Snoek 2019 proposed embedding RGB and optical flow into a single two-in-one stream network with new layers, which reduces the computational cost of the traditional two-stream detection network by half while maintaining high accuracy. The method can be easily embedded into existing appearance- or two-stream action detection networks and trained end-to-end. The two-in-one network structure is shown in Fig. 4. With only half the computation and number of parameters, the two-in-one stream still achieves impressive results on UCF101-24, UCF sports, and J-HMDB.

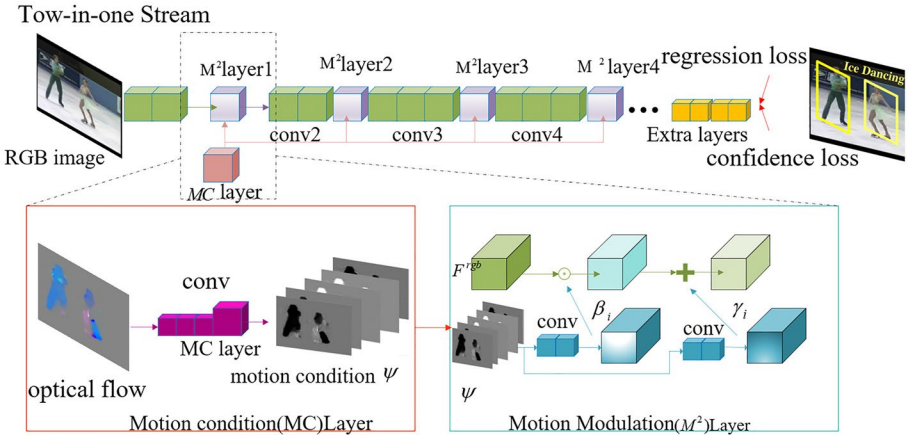

Fig. 4 Two-in-one network structure diagram (Zhao and Snoek 2019)

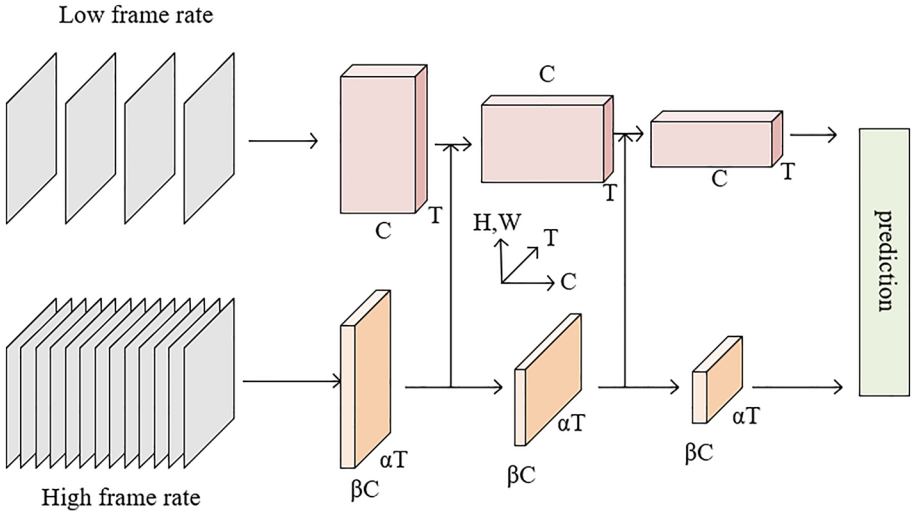

Fig. 5 SlowFast model (Feichtenhofer at al. 2019)

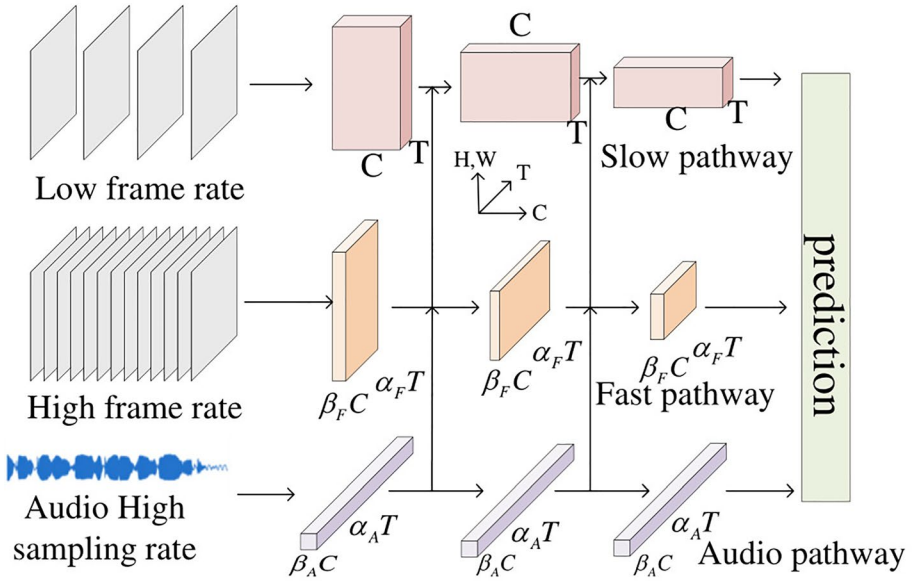

Fig. 6 Audiovisual SlowFast network (Xiao et al. 2020)

In the real world, most objects are static. Not all temporal and spatial dimensions are in the same direction, so they cannot be treated equally, but traditional convolution, such as 3D convolution, treats them equally. Based on this, Feichtenhofer et al. 2019 proposed a SlowFast network for video recognition. The model involves (1) a slow pathway, operating at a low frame rate, to capture spatial semantics, and (2) a fast pathway, operating at a high frame rate, to capture action information in the temporal dimension. As shown in

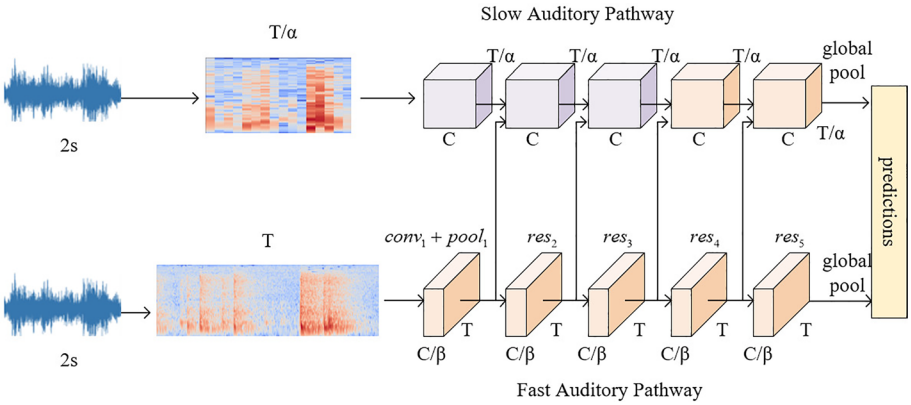

Fig. 7 SlowFast architecture (Kazakos et al. 2021)

Fig. 5, the T and C of the slow pathways are benchmarks for the fast pathway. For a video, the slow pathway samples T frames as input. Meanwhile, the fast pathway has to deal with high-frequency information, and uses no temporal downsampling layers in the whole process, so its input is always αT frame. To reduce the complexity and accuracy of the model, the number of fast pathway channels is reduced, which is β times the number of slow pathway channels. The two pathways are fused by lateral connections, followed by a complete connection layer for classification. Many experiments have proven its effectiveness.

For many video interpretation tasks, visual and audio signals are essential, because audio pathways generally train much faster than visual ones, which can lead to generalization issues during joint audiovisual training. To achieve unified audiovisual perception, Xiao et al. 2020c proposed a new architecture, audiovisual SlowFast network (AVSlowFast). As shown in Fig. 6, the model improves the SlowFast network by changing the number of input frames and channels of the fast pathway to αF and βF times those of the slow pathway, respectively. In addition, an audio pathway is added, which has a more acceptable temporal structure than the slow and fast pathways. The number of input frames and channels of the audio pathway is αF and βF times of the slow pathway. The model has shown its effectiveness on six datasets.

Different activities have different lengths of audio. Some may be momentary, while others may be repetitive over a more extended period. Based on this, Kazakos et al. 2021 proposed two streams for auditory recognition: a slow flow and a fast stream to achieve target recognition and temporal dimension sampling, respectively. The streams of the network are variants of residual networks. As shown in Fig. 7, each stream comprises an initial convolutional block with a pooling layer followed by four residual stages, where each step contains multiple residual blocks. The slow stream has high channel capacity, with β times more channels than the fast stream, while operating on a low sampling rate. As the input spectrogram is traversed temporally by α, the slow stream focuses on learning frequency semantics by restricting the temporal resolution and temporal kernels while keeping high channel capacity. The fast stream can focus on temporal patterns by having high temporal resolution and more temporal kernels but fewer channels.

As the earliest network in the behavior recognition algorithm based on deep learning, the two-stream network's division of video sequences into time and space provides

Table 2 Summary of deep learning algorithm of action recognition based on two-stream network

Algorithm name	Accuracy rate(%)			Improve	Code resources
	UCF-101	HMDB-51	UCF sports		
Two-Stream Convolutional Network (Simonyan and Zisserman 2014)	88	59.4	–	Deal with temporal and spatial features separately.	https://github.com/Yorwxue/Two-Stream-Convolutional-Networks
Deep Temporal Linear Encoding Network (Diba et al. 2017)	95.6	71.1	–	Combined with uniform sparse sampling method, segment network information with long temporal range is captured.	https://github.com/yjxiong/temporal-segment-networks
Two-in-one Network (Zhao and Snoek 2019)	Up to 75.48 on UCF101-24		92.74	Easy to embed existing appearance or two-stream action network and conduct end-to-end training.	https://github.com/qfei97/Two-in-One-ActionDetection
SlowFast Network (Xiao et al. 2020)	Up to 79.8 (top-1) and 93.9 (top-5) on Kinetics-400			Spatial information and temporal information should be treated separately.	https://github.com/facebookresearch/SlowFast
Audiovisual SlowFast Network(Feichtenhofer at al. 2019)	Up to 78.8 (top-1) and 93.6 (top-5) on Kinetics-400			Combination of visual and audio signals.	https://github.com/facebookresearch/SlowFast
Tow-stream audio recognition Network(Kazakos et al. 2021)	Up to 52.46 (top-1) and 78.12 (top-5) on VGG-Sound			Realize the shunt processing of audio of different lengths.	https://github.com/ekazakos/auditory-slow-fast

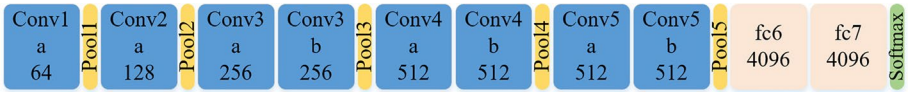

Fig. 8 Structure diagram of convolutional network (Tran et al. 2015)

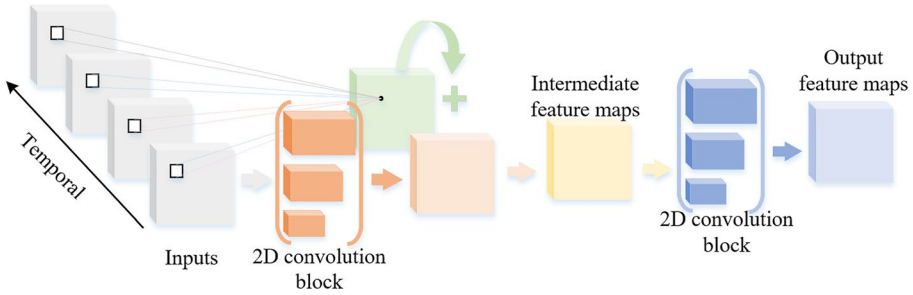

Fig. 9 MiCT model (Zhou et al. 2018)

researchers with much inspiration. Without considering training skills, such as network pretraining, the contribution of the two-stream network to the feature processing of video sequences is unmatched by C3D and LSTM networks. The deep learning algorithm for behavior recognition based on a two-stream network is summarized in Table 2.

4.2 Depth Learning algorithm for behavior recognition based on 3D convolution

In the beginning, Du et al. (Tran et al. 2015) proposed a 3D convolution neural network (C3D), in which 3D convolutions appeared for the first time, and applied it to behavior recognition. The optimal 3D convolution core size of $3 \times 3 \times 3$ was determined through a comparative experiment. As shown in Fig. 8, the network is composed of the following: 8 convolution layers, where all 3D convolution cores are $3 \times 3 \times 3$, with the same step size of $1 \times 1 \times 1$; 5 pooling layers, where the size of the first pooling layer is $1 \times 2 \times 2$ with a step size of $1 \times 2 \times 2$, and the size of the other pooling layers is $2 \times 2 \times 2$with a step of $2 \times 2 \times 2$, to retain more time information; 2 full connection layers with a size of 4096; and 1 softmax output layer. The best accuracy of this network on the UCF-101 database was 85.2%, and when combined with iDT, the accuracy was 90.4%. Under the same conditions, compared with the two-stream network, it improved accuracy by 1.6%. The network has good generalization performance and can be flexibly combined with other networks. Its calculation efficiency is high and it is easy to train and use.

The large number of parameters of 3D convolution increases the difficulty of optimization and memory usage. The computational cost of a 3D convolution neural network makes training it very difficult. Given the excellent performance of 2D convolution in two-dimensional image processing, Zhou et al. 2018 proposed a hybrid 3D/2D convolution module (MiCT) to process video data. As shown in Fig. 9, the MiCT module combines the hybrid 3D/2D series module and the cross-domain residual 3D/2D parallel module, which increases the depth of the 3D convolutional neural network and reduces the complexity of learning 3D features and fusing spatio-temporal features. MiCT enables 3D convolutional

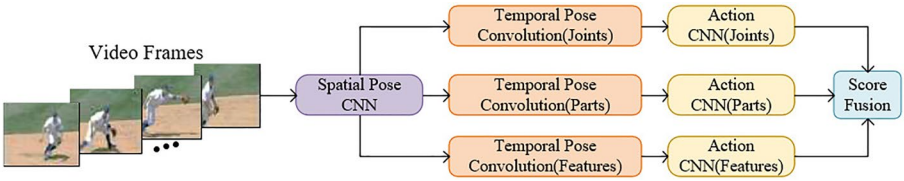

Fig. 10 Framework of Pose-Action 3D Machine (Yan et al. 2019)

neural networks to extract deeper spatio-temporal features with less 3D spatio-temporal fusion, smaller size, and faster speed.

In this context, 3D convolution can be used to model space and time series, but this modeling method cannot effectively extract the dynamic characteristics of human posture in time series. To address this, Yan et al. 2019 proposed a novel pose-action 3D (PA3D) machine. As shown in Fig. 10, it consists of three semantic modules: spatial pose CNN, temporal pose convolution, and action CNN. First, the spatial pose CNN extracts the joint, part affinity fields, and convolutional features to form three kinds of pose heatmaps, and sends them to the temporal pose convolution. For the heatmap stack of the same joint at different times, a cube with dynamic change in temporal sequence is obtained, which is fused by 1×1 convolution and dilation convolution, and finally sent into a convolutional and fully connected layer classification network. PA3D and I3D are highly complementary and superior to many other pose-based methods.

Although 3D convolution can deal with spatio-temporal information, it is insufficient for long-term video processing. In order to address this challenge, Ren et al. 2021 proposed a new processing framework called neural architecture search-temporal convolution (NAS-TC). They divided it into two stages: In the first phase, I3D is used as the backbone network to complete the computationally intensive feature extraction task. In the second stage, the neural architecture search (NAS) method is used to complete the lightweight task of extracting long-range temporal-dependent information. This method assigns parameters more reasonably to ensure the completion of long-range video processing tasks, and experiments show that the model improves the accuracy.

A 3D convolution network has high accuracy in video recognition, but it requires large computation and memory resources. Therefore, Kondratyuk et al. 2021) proposed a three-step approach to improve computational efficiency while substantially reducing the peak memory usage of 3D CNNs. (1) They designed a video network search space (NAS), which uses 3D convolution to embed overlapping segments of input video and to calculate their average. (2) They introduced the stream buffer technique, which allows 3D CNNs to embed streaming video sequences of arbitrary length for training and inference with a small constant memory footprint. (3) They proposed a simple ensembling technique by training two streaming MoViNets independently with the same total FLOPs as a single model and averaging their logits. This simple technique further improves accuracy without sacrificing efficiency.

The proposal of the C3D network expands the ideas of researchers beyond two-dimensional convolutions. 3D convolutions have good generalization performance, can be combined with many networks, and improve the performance of the original network. The most significant difference of C3D is that it reduces the number of network parameters and speeds up network training. The depth learning algorithm of behavior recognition based on 3D convolution is summarized in Table 3.

Table 3 Summary of deep learning algorithm of action recognition based on 3D convolution

Algorithm name	Accuracy rate(%)			Improve	Code resources
	UCF-101	HMDB-51	UCFSports		
C3D (Tran et al. 2015)	85.2%	–	–	Proposed 3D convolution with the temporal dimension.	https://github.com/facebook/C3D
MiCT (Zhou et al. 2018)	88.9%	63.8%	–	Extracts deeper spatio-temporal features.	–
PA3D (Yan et al. 2019)	Up to 82.1 on HMDB.			Extracts dynamic human posture features in time series.	–
NAS-TC (Ren et al. 2021)	Up to 76.83 on MultiTHUMOS.			Improves ability of video processing over a long time.	–
MoViNets (Kondratyuk et al. 2021)	Up to 84.8 (top-1) on Kinetics-600			Improves calculation rate and significantly reduces memory usage of 3D convolution .	https://github.com/tensorflow/models/tree/master/official/vision

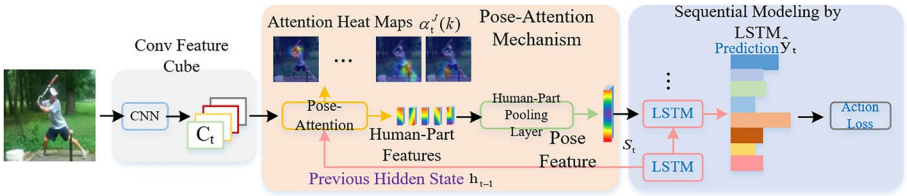

Fig. 11 Structure of RPAN network (Du et al. 2017)

4.3 Deep Learning algorithm of behavior recognition based on LSTM Network

In deep networks (Qu et al. 2021; Song et al. 2021; Xia et al.2021c; Xia et al. 2020a; Xia et al. 2020b), because LSTM can deal with temporal sequence information, LSTM networks are often used in behavior recognition. The internal structure of an LSTM unit consists of an amnesia gate, an input gate, and an output gate (Zhihui et al. 2017; Sanhong et al. 2017). The LSTM network allows gradients of key sequences to be transmitted all the time, avoiding the problem of gradient disappearance to some extent (Jiyang 2016).

To reduce the amount of computation and learn the global features of video, Ng et al. (Yue-Hei Ng et al. 2015) proposed a two-stream network model combined with LSTM. In this model, the pre-trained CNN network (AlexNet or GoogLeNet) on ImageNet is used to extract image and optical flow features of video frames. The extracted features are input into the LSTM network to obtain the final result (Yuhuan 2018). Although the performance of the network is fair, it provides a new idea for research behavior recognition: even if there is a lot of noise in optical flow images, it is helpful in classification when combined with LSTM.

Because the previous method combining convolutional neural network and LSTM cannot represent micro movements, and because other attention-based LSTM methods cannot train the LSTM network well, Du et al. 2017 proposed an end-to-end recurrent pose attention network (RPAN), which combines the attention mechanism with the LSTM network to represent more refined actions. The network structure diagram is shown in Fig. 11. The accuracy of the network reached 78.6 and 97.4% on the Sub-JHMDB and PennAction databases, respectively. We believe that the combination of LSTM and attention mechanism can achieve good results in behavior recognition, but the network structure is more complex.

Long et al. 2018 proposed a multimodal RNN framework, which divides visual features (including RGB image and optical flow features) and acoustic features into equal-length segments in LSTM, which reduces the amount of computation and improves the speed. The LSTM network is applied to extract different features. The multimodal LSTM network framework is shown in Fig. 12.

In order to make full use of spatial correlation in the video, Li et al. 2018 introduced the attention mechanism into LSTM and proposed the VideoLSTM model. This model obtains the spatial correlation from each frame of the video image through the convolutional neural network, and obtains a motion-based attention map. It finally locates the spatio-temporal position using the attention map according to the action label.

Wang et al. 2019 proposed an I3D-LSTM model by combining I3D with an LSTM network. I3D extracts spatial features and captures low-level action features between adjacent frames and then uses the LSTM network to model high-level temporal features. This method can learn low-level and high-level features well. The experimental results on the

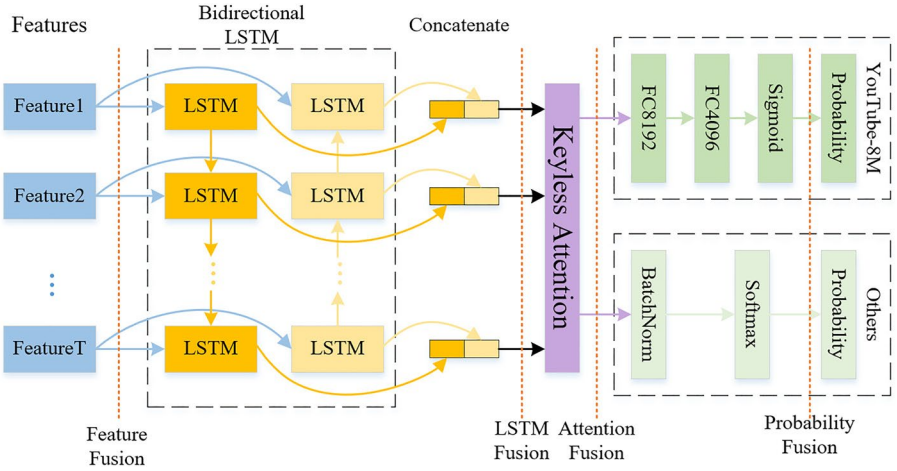

Fig. 12 Framework of multimodal LSTM network (Long et al. 2018)

UCF-101 video dataset verified that the I3D-LSTM model was more effective than other mainstream methods at that time.

He et al. 2021 proposed the DB-LSTM model, which uses dense hopping connections to strengthen the feature propagation and reduce the number of parameters. This network is also an extended form of the two-stream network. As shown in Fig. 13, the input is extracted frames and optical flow maps, which are fed into the stack representation learner (SRL) to produce the feature stack. The feature stack is utilized to model the temporal pattern. Finally, a fusion layer is employed to integrate the spatial and temporal clues to predict the final result. Experiments on two open datasets, UCF101 and HMDB51, showed that the model was superior to other action recognition methods when it was proposed.

In order to improve the differential discrimination of features, Muhammad et al. 2021 proposed a bi-directional long short-term memory (BiLSTM) based attention mechanism with a dilated convolutional neural network (DCNN) that selectively focuses on effective features in the input frame to recognize different human actions in videos. Figure 14 shows the overall architecture of the network. The model uses DCNN to extract

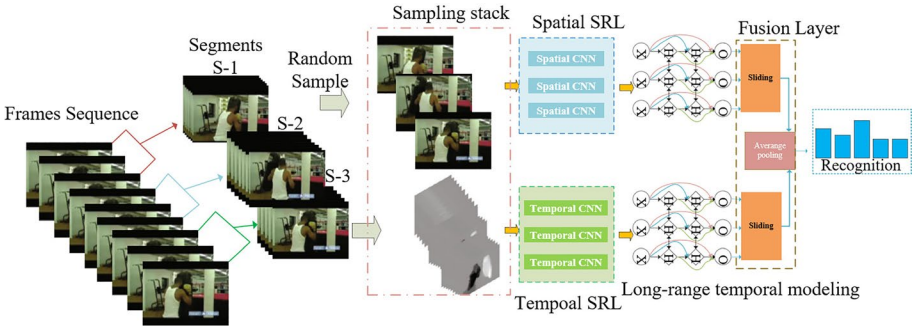

Fig. 13 DB-LSTM model framework (He et al. 2021)

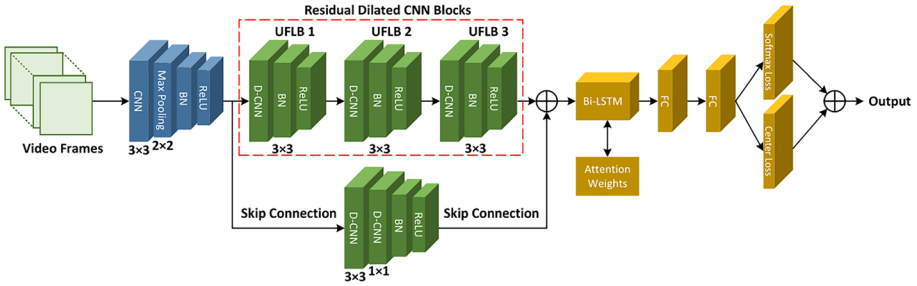

Fig. 14 Network structure diagram (Muhammad et al. 2021)

the CNN features from the input data. The fused features are sent into BiLSTM to learn the long-term dependencies, and then the context vector is obtained through the attention mechanism. In the end, the network outperformed the latest methods on the UCF11, UCF ,sports and J-HMDB datasets by 1 to 3%.

In order to take into consideration both spatial and motion features while constructing temporal dependencies, Majd and Safabakhsh 2020 proposed an extended version of LSTM units, named C2LSTM. The overall architecture of the network is shown in Figure 15. The network first feeds the video data into convolutional towers to extract spatial features. The outputs of all convolutional towers are then given to the C^2LSTM layer to extract spatial and temporal information as well as temporal dependence. C^2LSTM offers two improvements: (1) Its input gates and weights are 2D arrays and the multiplication operators are replaced with convolution, and (2) the input of previous time t-1 is used to calculate the correlation of subsequent input data. The network has proven to be effective on the UCF101 and HMDB51 datasets.

In the beginning, the accuracy of the LSTM dual-stream network model was not significantly higher than that of the two-stream network on the UCF-101 database. However, it reduced the amount of network computation. In the later development, it could be seen that most of the databases on which LSTM-based networks perform well were those integrated with human body actions. Compared with the Two-Stream Networks and C3D, the LSTM networks are more sensitive to slender limb actions. In the final analysis, the development of behavior recognition algorithms based on deep learning is inseparable from the contributions of Two-Stream networks, C3D networks, and LSTM networks.The deep learning algorithms for behavior recognition based on LSTM networks are summarized in Table 4.

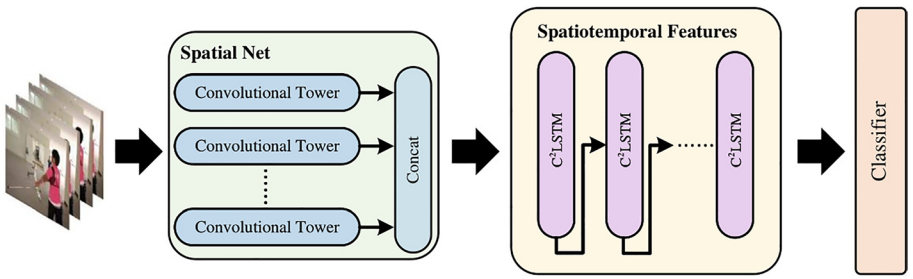

Fig. 15 Overall framework of C2LSTM network (Majd and Safabakhsh 2020)

Table 4 Summary of deep learning algorithm of action recognition based on LSTM network

Algorithm name	Accuracy rate(%)			Improve	Code resources
	UCF-101	HMDB-51	UCFSports		
Combined with LSTM two-stream network (Yue-Hei Ng et al. 2015)	88.6%	–	–	Reduces the amount of calculation and extract global features.	—
RPAN (Du et al. 2017)	Reached 78.6 on the Sub-JHMDB database		Reached 97.4 on the PennAction database	Refines the action characteristics and better train LSTM.	https://github.com/agethen/RPAN
Multimodal LSTM (Long et al. 2018)	94.8%	–	–	Integrated with acoustic features.	–
Video-LSTM (Li et al. 2018)	89.2%	56.4%	–	Adds attention mechanism to LSTM.	–
I3D-LSTM (Wang et al. 2019)	95.1%	–	–	Learns low-level and high-level features.	–
DB-LSTM (He et al. 2021)	96.1%	73.7%	–	Integrates merits of a dense network, bidirectional modeling.	–
DCNN-BiLSTM(Muhammad et al. 2021)	–	–	98.5%	Adds DCNN to BiLSTM.	–
C^2LSTM(Majd and Safabakhsh 2020)	92.8%	61.3%	–	Constructs time dependencies and consider spatial and motion characteristics.	–

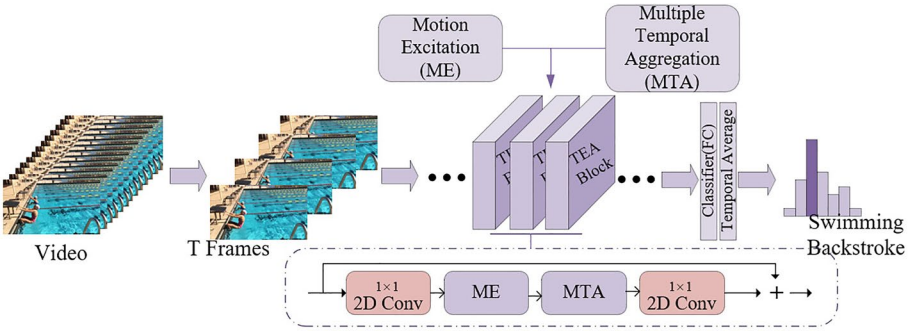

Fig. 16 TEA block temporal modeling frame diagram (Li et al. 2020)

4.4 Other deep learning algorithms for behavior recognition

In addition to the two-stream, 3D convolution, and LSTM networks, many other algorithms also have excellent performance in behavior recognition. These algorithms have high accuracy, fast speed, and good robustness.

Wu et al. 2018 proposed training the deep network directly on the compressed video to filter the noise and make the training easier. Li Chao et al. 2019 designed a joint spatio-temporal feature learning operation (CoST) to overcome the problem of a large number of 3D convolution network parameters and the limitation of real time. The network reduces the amount of calculation while maintaining accuracy, and its structure is relatively straightforward. Choutas et al. 2018 proposed encoding the action changes of human joints as key points; the resulting feature is called PoTion. The featured graph is input into the convolutional neural network for behavior recognition. Cho et al. 2018 proposed a new spatio-temporal fusion network (STFN), which can obtain local and global information of complementary data and is suitable for any video classification network. Li et al. 2020 proposed a motion excitation (ME) module and a multiple temporal aggregation (MTA) module, which are inserted into a standard ResNet block (He et al. 2016a, b) to build temporal excitation and aggregation (TEA) block for effective and efficient temporal modeling. The TEA block modeling framework is shown in Fig. 16.

The behavior recognition deep learning algorithms of other networks are summarized in Table 5.

5 Deep learning algorithms for behavior recognition based on skeleton data

With the continuous development of deep learning, convolutional neural network (CNN) (LeCun et al. 1998; Zhang et al. 2017), as a representative network of deep learning, can efficiently deal with single and two-dimensional data with regular structure. Although CNN has achieved great success in processing image sequences in Euclidean space, it cannot deal with the data in many scientific research fields generated in non-Euclidean space. In particular, human skeletal joints compose a topological graph that belongs to Euclidean structure. Skeleton data are less susceptible to appearance than RGB data and depth data. Because the skeleton is an advanced feature of the human body, it can better avoid the

Table 5 Summary of deep learning algorithm of action recognition based in other networks

Algorithm name	Accuracy rate(%)			Improve	Code resources
	UCF-101	HMDB-51	UCFSports		
Compressed video action recognition (Wu et al. 2018)	90.4%	59.1%	–	Uses compressed video for deep network training.	–
CoST (Li et al. 2019)	Reach 93.2 on Kinetics database			Joint learning of spatio-temporal features using 2D convolution.	–
PoTion (Choutas et al. 2018)	98.2%	82.3%	–	Encodes motion of human joints.	–
STFN (Cho and Foroosh 2018)	93.5%	70.4%	–	Obtain local and global information of complementary data, and apply it to any video classification network.	–
TEA (Li et al. 2020)	96.9%	73.3%	–	Uses the TEA module to capture short-term and long-term time evolution.	–

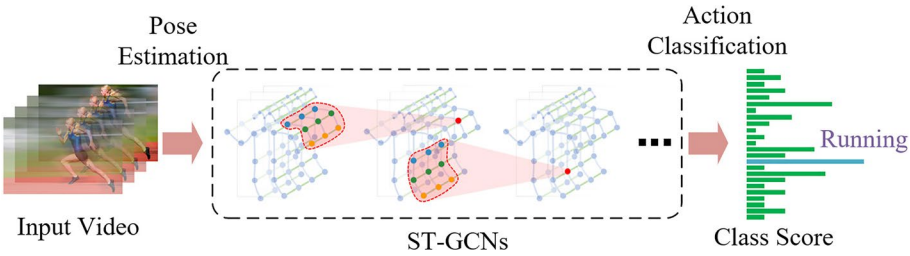

Fig. 17 ST-GCN structure (Yan et al. 2018)

noise caused by background occlusion, illumination change, and visual angle change in the process of calculation and storage. (Yan et al. 2017) Therefore, it has received extensive attention. For this kind of topological graph, GCN (Du et al. 2015) can be processed directly. In this paper, the convolution operation is extended to the graph domain, and the graph convolution is used to combine the extracted spatio-temporal features on the graph structure. Finally, the result of behavior recognition is obtained.

There are multiple modalities in human behavior recognition, such as appearance, depth, optical flow, and the skeleton. Among these modalities, dynamic human skeletons usually convey significant information that is complementary to others. Conventional approaches for modeling the skeleton only carry out temporal sequence analysis on feature vectors formed by temporal joint nodes, resulting in limited expression and difficulty with generalization. Therefore, Yan et al. (Yan et al. 2018) proposed the skeleton-based spatio-temporal graph convolution network (ST-GCN), and theirs was the first paper to use a graph convolution network to identify skeletal behavior recognition. It is also a pioneering and representative method. It summarizes how to use a spatio-temporal graph constructed based on the human skeleton sequence and perform the convolution operation on the graph.

The ST-GCN model, which is based on the human skeleton, represents the general sequence of behavior recognition, overcoming the deficiency of the RGB-based model. Using graph convolution to recognize the behavior of critical nodes can produce a hierarchical representation of skeletal joints and get better recognition results. The network opens up a new direction and possibilities for behavior recognition. The ST-GCN network structure is shown in Fig. 17. We can construct spatial and temporal graphs on skeletal sequences, apply and gradually generate higher-level feature maps on the graph, and finally, the action category is classified by the softmax classifier. On two large datasets, Kinetics and NTU-RGBD, it achieves substantial improvements over mainstream methods.

5.1 Adaptive graph structure

ST-GCN is a spatio-temporal model based on the skeletal structure. First, its graph topology is manually set and fixed across all layers and input samples, ignoring the underlying relationships between all nodes. Second, it fails to capture the features of crucial neighborhood nodes, ignoring the importance of different nodes, resulting in inaccurate final classification results. In order to make the graph structure more adaptable to varied data samples, Shi et al. 2019b, Li et al. 2019 and Yang et al. 2020 made a series of improvements based on the adjacency matrix of the graph structure.

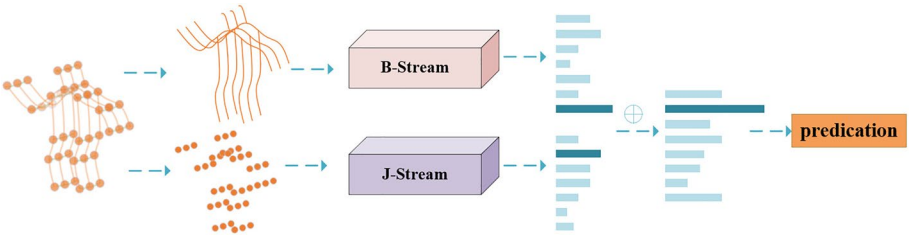

Fig. 18 Illustration of overall architecture of 2s-AGCN (Shi et al. 2019b)

The two-stream adaptive graph convolutional network (2s-AGCN) proposed by Shi et al. 2019b adaptively learns the topology of skeletal joints according to different graph convolution layers and BP algorithms in an end-to-end manner. Its hierarchical structure is better and it allows the addition of connections to dynamically adjust the graph structure, hence it is applied to recognition tasks. Typically, the length and direction of bones are naturally more informative and discriminative for action recognition. The model's two-stream structure uses the first-order information of the skeleton data (node information) and the second-order information, which brings notable improvement in recognition performance. Figure 18 shows the overall architecture of 2s-AGCN: first, the data of bones are calculated based on the data of joints; then, the joint and bone data are fed into the J-stream and B-stream, respectively; finally, the softmax scores of the two streams are added to obtain the fused score and predict the action label. The accuracy of this model is about 7% higher on the NTU-RGB+D dataset compared to the results of Yan et al. 2018.

Li et al. 2019) proposed the actional-structural graph convolution network (ASGCN), which uses an encoder-decoder structure to capture the implicit joint correlation and high-order neighborhood information. The network uses the high-order polynomials of the adjacency matrix to capture the inherent physical structure links between joints. It extends the skeleton graph and represents high-order and potential dependencies as structural and action links, respectively. For generalized graphs with action and structural links, the model uses actional-structural graph convolutions to capture spatial features. It also stacks multiple actional-structural graph and temporal convolutions. Figure 19 shows the structure of the AS-GCN network: the inferred actional graph A-links and extended structural graph S-links are fed to the AS-GCN blocks to learn spatial features, the multigraph AS-GCN extracts spatial and temporal information, the average of the adjacent features of each joint is weighted, and the last block is connected to the classifier to generate the final prediction. Finally, two datasets, NTU-RGB+D and Kinetics, were used to verify that the AS-GCN network is much better than the previous methods. AS-GCN also shows promising results for future pose prediction.

Yang et al. 2020 proposed a new pseudo graph convolutional network with temporal and channel-wise attention (PGCN-TCA) to capture the dependencies between connected joints and joints that are not physically connected. They also proposed a mixed temporal and channel-wise attention network to describe the importance of different frames and channels. In this way, PGCN-TCA can extract key frames and filter out input frames containing more features. A flowchart of the PGCN-TCA model is shown in Fig. 20. First, the relative coordinate and temporal differences are concatenated to form the network input, and second, the input is normalized by the normalization layer; the data are then input into 10 pseudo-graph convolution blocks. The pseudo-graph convolutional block with different numbers uses a temporal stride of 2 to reduce the sequence length, and a global average

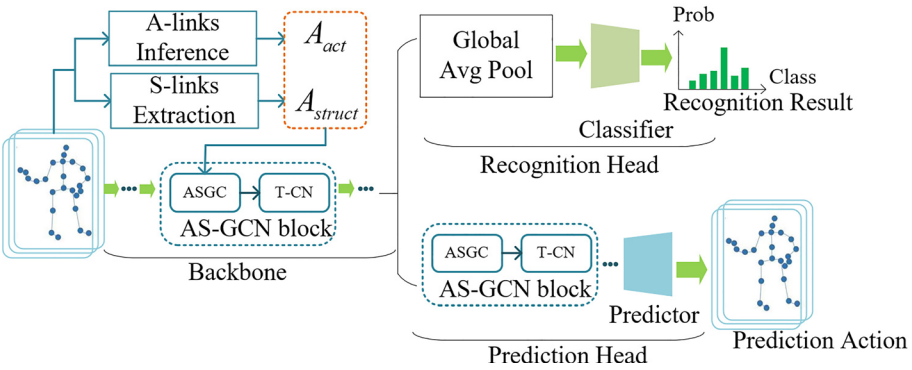

Fig. 19 AS-GCN network structure diagram (Li et al. 2019)

pooling layer is used after the last block; finally, the fully connected layer and softmax are used to calculate the score for each class. The model achieves competitive results on challenging datasets.

5.2 Feature information fusion

In action recognition tasks, information on both joints and bones in skeleton data has proven to be extremely useful. However, how to combine the two types of data to best exploit the relationship between joints and bones remains a problem to be solved. To address this issue, Shi et al. 2019a, Korban et al. 2020 and Zhang et al. 2020 considered the importance of the internal relationship between joint features and skeletal features.

Shi et al. 2019a represented skeleton data as a directed acyclic graph (DAG) based on the natural kinematic dependency between joints and bones in the human body. A novel directed graph neural network was designed specifically to extract the information of joints, bones, and their relationships and to make predictions based on the extracted features. In addition, to better fit the action recognition task, they made improvements based on 2s-AGCN (Shi et al. 2019b): In the first 10 rounds of training, $A = P + A_0$, is used and is not added to the following training, which achieves the optimal effect of the network. In addition, for improved performance, they combined the spatial information in the two-stream framework with the motion information of the same graph structure. Finally, they obtained recognition results by fusion

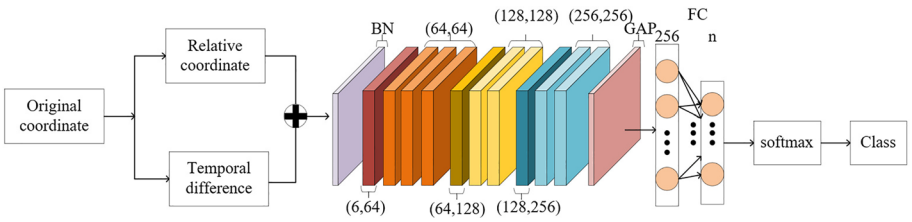

Fig. 20 Illustration of the PGCN-TCA architecture (Yang et al. 2020)

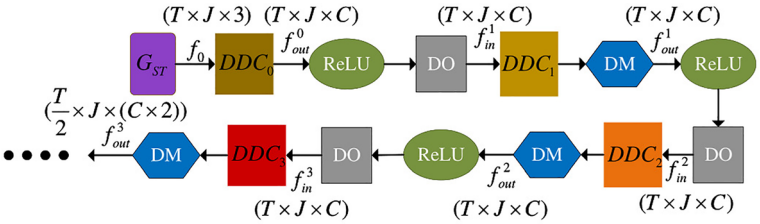

Fig. 21 DDGCN architecture diagram (Korban and Li 2020)

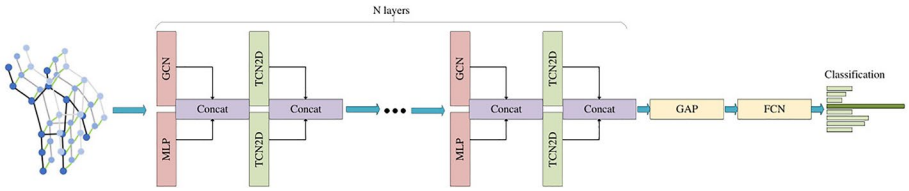

Fig. 22 Overall architecture of SFAGCN (Zhang et al. 2020)

and classification. Experiments show that the accuracy of the DGNN network is about 1% higher than that of 2s-AGCN.

Korban et al. 2020 proposed an end-to-end dynamic directed graph convolutional network (DDGCN). The DDGCN consists of three new feature modeling modules: dynamic convolutional sampling (DCS), dynamic convolutional weight (DCW), and directed graph spatio-temporal (DGST) feature extraction, which can adaptively learn spatio-temporal relations and model spatial and temporal sequence information. Among them, the DCS and DCW modules can effectively capture the spatio-temporal correlation between dynamic non-adjacent joints. The DSTG feature extraction module enhances the features of the action by including spatio-temporal ordered information. The DDGCN architecture shown in Fig. 21 consists of DDC blocks, dropout (DO) layers, a dimension modifier (DM), and ReLU. One of the DDC blocks is composed of DCS, DCW, and DSTG. Comprehensive experiments show that the DDGCN improves the accuracy of action recognition on multiple public datasets.

Zhang et al. 2020 proposed a novel structure-feature fusion adaptive GCN (SFAGCN) for skeleton-based action recognition. The model decouples the structure and feature information in the spatio-temporal graph and then fuses the decoupled information data. In this way, the topological structure and joint features of the skeleton graph can be effectively fused. In addition, the relevance of spatio-temporal data can be preserved by the fusion strategy, with data integrity ensured. The overall architecture of SFAGCN is shown in Fig. 22. The basic module of the space-time block is composed of TCN gate, SFAGCN, and adaptive fusion module; the extracted features are processed by global average pooling (GAP) and full convolutional network (FCN) and are classified by softmax classifiers. Taking 2s-AGCN and shift-GCN as the baseline, SFAGCN has achieved higher performance on the NTU-RGBD 60 dataset.

5.3 Optimization of receptive field

Compared with previous methods, ST-GCN has better performance, but there are still some problems. First, due to the inflexibility of attention mechanisms and narrow receptive fields, it is difficult to learn the relationship between non-physical connection nodes. Second, the skeletal structure with fixed physical connection limits the size of the receptive field, resulting in restricted performance improvement. In order to improve the receptive field and better extract higher-order information from human bone structure, the following studies were carried out from the aspects of optimizing the convolution mode (Cheng et al. 2020), enriching the feature information of the input (Si et al. 2019), and strengthening the node correlation (Li et al. 2019).

The shift graph convolutional network (Shift-GCN) was proposed by Cheng et al. 2020 for skeleton-based action recognition. It can significantly reduce computational cost. The improvement over ST-GCN (Yan et al. 2018) is reflected in the graph convolution of spatial information and the change in temporal sequence modeling methods. Shift-GCN is composed of spatial and temporal shift graph convolutions. Two kinds of spatial shift graph operation are proposed for spatial skeleton graphs, local and non-local. For local shift graph operation, the receptive field is specified with the physical body structure. Non-local shift graph operation makes the receptive field of each node cover the full skeleton graph and learns the relations between joints adaptively. Therefore, non-local spatial shift graph operation is computationally efficient and achieves good performance. Two kinds of temporal shift graph operations are proposed for temporal skeleton graph, naive and adaptive. Because adaptive temporal shift graph operation can adaptively adjust the receptive field, it can solve the problem of manually setting the receptive field of the naive temporal shift graph. It also optimizes the time modeling and reduces the computational complexity compared with conventional time models. The network notably exceeded the state-of-the-art methods at the time with more than 10 times less computation cost on the three skeleton-based action recognition datasets.

Si et al. 2019 proposed an attention enhanced graph convolutional LSTM network (AGC-LSTM), which consists of three elements: attention mechanism (A), GCN, and LSTM. In this model, the combination of graph convolution and LSTM is applied to bone-based behavior recognition for the first time. In the structure diagram of the AGC-LSTM unit block, LSTM is improved: the original input sequence data is changed into bone data, and the product is changed into the graph convolution. The AGC-LSTM network structure is shown in Fig. 23. First, the coordinate of each joint is transformed into a spatial feature with a linear layer; in order to capture spatio-temporal information, the difference between two consecutive frames is used as the frame difference feature. Next, three AGC-LSTM layers are used to model discriminative spatio-temporal features. Finally, the class of human action is predicted. The model has the ability to learn high-level semantic representation and a significantly reduced computation cost, but it can only improve the receptive field on small-scale graph structures. The best results were obtained on the NTU RGB+D and Northwestern-UCLA datasets when the algorithm was proposed.

Li et al. 2019 proposed a spatio-temporal graph routing (STGR) scheme, which adaptively learns the intrinsic high-order connectivity relationships for physically distant skeletal joints and effectively expands the receptive field. Specifically, the scheme is composed of two components: a spatial graph router (SGR) and a temporal graph router

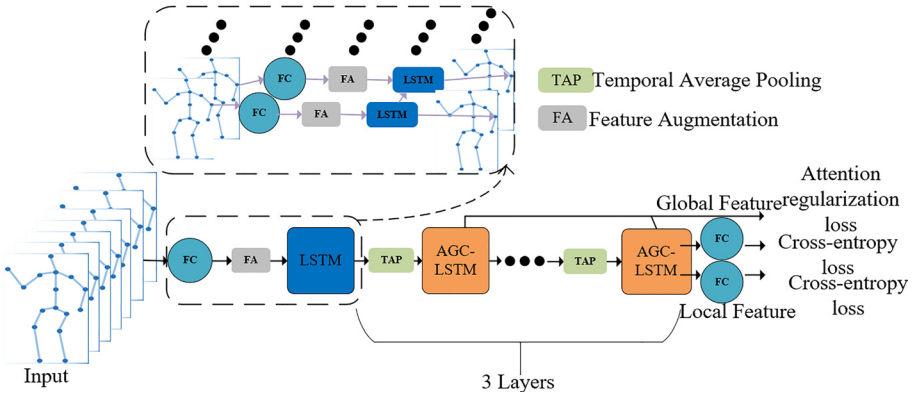

Fig. 23 AGC-LSTM network structure (Si et al. 2019)

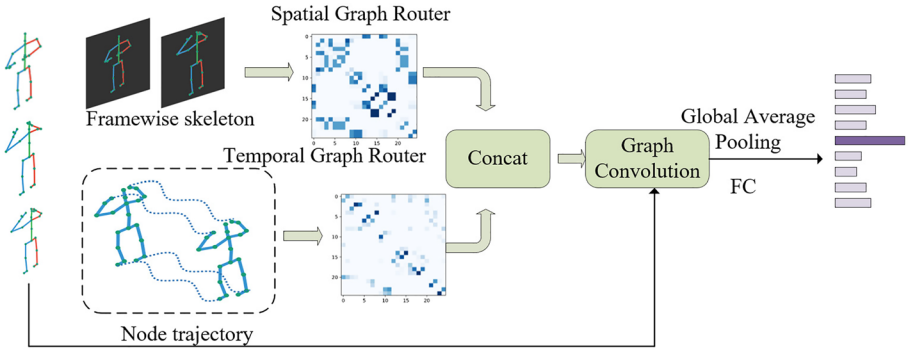

Fig. 24 STGR flowchart (Li et al. 2019)

(TGR). The SGR aims to discover the connectivity relationships among the joints, and the TGR explores the structural information by measuring the degree of correlation between temporal joint node trajectories. The proposed scheme is naturally incorporated into the framework of the GCN to produce a set of skeleton-joint connectivity graphs, which are further fed into the classification networks to better match action recognition tasks. The flowchart of STGR is shown in Fig. 24. The effectiveness of the method was proved on two benchmark datasets, NTU-RGB+D and Kinetics.

The deep learning behavior recognition algorithms based on skeleton data are summarized in Table 6

6 Conclusion and prospect

Human behavior recognition has many applications in today's society, and has received a great of attention by researchers in related fields. In order to help beginners understand the deep learning methods of behavior recognition and track the research hotspots, this paper conducted an integrated analysis of the popular behavior recognition methods based

Table 6 Summary of Deep Learning behavior recognition algorithm based on Skeleton data

Algorithm name	Accuracy rate(%)					Improve	Code resources
	Kinetics		NTU-RGB+D		HDM05		
	Top-1	Top-2	CS	CV			
ST-GCN (Yan et al. 2018)	30.7	52.8	81.5	88.3	–	Uses graph convolution for behavior recognition to form hierarchical representation of skeletal joints and achieve better recognition results	https://github.com/yysijie/st-gcn
2s-AGCN (Shi et al. 2019b)	36.1	58.7	88.5	95.1	–	Makes full use of second-order skeletal information	https://github.com/lshiwjx/2s-AGCN
AS-GCN (Li et al. 2019)	34.8	56.5	86.8	94.2	–	Captures implied joint correlations and obtains physical and structural links between joints	https://github.com/limaosen0/AS-GCN
DGNN (Shi et al. 2019a)	36.9	59.6	89.9	96.1	–	Extracts key nodes and bone information, and directional associations between them	https://github.com/kenziyuliu/Unofficial-DGNN-PyTorch
PGCN-TCA (Yang et al. 2020)	–		88	93.6	86.59±1.84	Extracts keyframes and filters out input frames that contain more features	–
DDGCN (Korban and Li 2020)	38.1	60.8	91	97.1	–	Provides spatial and temporal correlation in simulating human behavior	–
Shift-GCN (Cheng et al. 2020)	–		90.7	96.5	–	Uses shift map operations and lightweight point-by-point convolution	https://github.com/kchengiva/Shift-GCN
STGR (Li et al. 2019)	33.6	56.1	86.9	92.3	–	Adaptive learning of inherent high order connection of physically separated skeletal joints	–
SFAGCN (Zhang et al. 2020)	Kinetics 400		NTURGB+B 120		–	Integrates structural information of bone data and position information of joints	–
	38.3	60.6	87.3	88.5			
AGC-LSTM (Si et al. 2019)	–		89.2	95	–	Captures differences between spatial configuration and temporal dynamics	–

on deep learning in recent years. Behavior recognition methods based on deep learning do not require too much manual participation, and are trained and learned directly on the video dataset. Although these methods need a lot of data to support them, they can get more comprehensive features. For video sequences, deep learning can better deal with temporal sequences, so it is more suitable for feature extraction in behavior recognition than traditional methods. GCN is very effective in processing graph data in non-Euclidean space and has attracted more attention. In recent years, compared with unprocessed limb information, skeleton information is more stable and is free from background interference, so behavior recognition based on skeleton detection has also become an essential area of computer vision. The contribution of our paper is that in addition to introducing the traditional machine learning and image-based deep learning algorithms in the field of behavior recognition, we also focus on deep learning algorithms based on skeleton data. Different from previous reviews, our paper covers many classical papers and summarizes the latest progress in detail.

In recent years, researchers have made great achievements in the field of behavior recognition. However, there are still many challenges and directions worth pursuing in further research:

1. Temporal action detection problem. Behavior recognition is mainly used to classify segmented video clips, but in real life, most videos are unsegmented and of different lengths. The current trend is to find more accurate times to locate the start and end of an action and to determine the category of each action. The time span of temporal behavior segments can vary greatly; for example, for people of different ages raising hands, the shortest behavior segment is about a few seconds and the longest is more than dozens of seconds. Due to the large time span, it is difficult to detect temporal actions. Therefore, achieving fast and accurate action detection is the key to the successful application of behavior recognition, and it is also a research hotspot for the future.
2. Multi-action semantic recognition problem. At present, the existing behavior recognition algorithms can only deal with videos containing a single action. For video frames containing consecutive human behaviors with multiple semantics, it is difficult for the existing algorithms to complete the task of behavior recognition accurately and effectively. However, individuals in daily life tend to mix multiple actions at the same time, such as talking on the phone while walking. Therefore, how to develop an algorithm to effectively distinguish and recognize multiple action semantics in videos is one of the future research directions in this field.
3. Engineering realization problem. In recent years, the number of academic papers related to behavior recognition has increased, which has promoted rapid development in the field of behavior recognition. However, most current research is focused solely on improving recognition accuracy and ignores practical application issues. In real-world applications, such as video surveillance, smart homes, and so on, both the accuracy and real-time performance of behavior identification are critical. Therefore, in the pursuit of high precision, the model parameters, computational power consumption, behavior recognition speed, and other factors should also be taken into consideration. In view of the above problems, real-time behavior recognition should be widely studied in the future in order to be better applied in real life.

Behavior recognition methods based on deep learning have inspired many new methods and models recently. These methods are showing higher accuracy and real-time

performance. However, there are still many data samples, complex network models, and other problems that will need to be further improved in future development.

Acknowledgements Research in this article is supported by the National Natural Science Foundation of China (No. 61876079), the key special project of the National Key R &D Program (2018YFC1405703), the financial support of Jiangsu Austin Optronics Technology Co., Ltd. is deeply appreciated, and I would like to express my heartfelt thanks to those reviewers and editors who submitted valuable revisions to this article.

Author Contributions All authors drafted the manuscript, read, and approved the final manuscript.

Data Availability The data and code used to support the findings of this study are available from the corresponding author upon request (001600@nuist.edu.cn).

Declarations

Conflict of interest No potential conflict of interest were reported by the author.

References

Arandjelovic R, Zisserman A (2013) All about vlad. Proceedings of the ieee conference on computer vision and pattern recognition (pp.1578–1585)

Chen B, Xia M, Huang J (2021) Mfanet: a multi-level feature aggregation network for semantic segmentation of land cover. Remote Sensing 13(4):731

Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H (2020) Skeleton-based action recognition with shift graph convolutional network. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp.183-192)

Cho S, Foroosh H (2018). Spatio-temporal fusion networks for action recognition. Asian conference on computer vision (pp. 347-364)

Choutas V, Weinzaepfel P, Revaud J, Schmid C (2018) Potion: Pose motion representation for action recognition. Proceedings of the ieee conference on computer vision and pattern recognition (pp. 7024-7033)

Deng S, Fu Y, Wang H (2017) Multi-label classification of chinese books with lstm model. Data Analysis and Knowledge Discovery 1(7):52–60

Diba A, Sharma V, Van Gool L (2017) Deep temporal linear encoding networks. Proceedings of the ieee conference on computer vision and pattern recognition (pp. 2329-2338)

Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. Proceed- ings of the ieee conference on computer vision and pattern recognition (pp. 2625-2634)

Du W, Wang Y, Qiao Y (2017) Rpan: An end-to-end recurrent poseattention network for action recognition in videos. Proceedings of the ieee international conference on computer vision (pp. 3725-3734)

Du Y, Fu Y, Wang L (2015) Skeleton based action recognition with convolutional neural network. 2015 3rd iapr asian conference on pattern recognition (acpr) (pp. 579-583)

Duta IC, Ionescu B, Aizawa K, Sebe N (2017) Spatio-temporal vlad encoding for human action recognition in videos. International conference on multimedia modeling (pp. 365-378)

Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. Proceedings of the ieee/cvf international conference on computer vision (pp. 6202-6211)

Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning. MIT press Cambridge, USA

He J, Wu X, Cheng Z, Yuan Z, Jiang Y (2021) Db-lstm: Densely-connected bi-directional lstm for human action recognition. Neurocomputing 444:319–331

He K, Zhang X, Ren S, Sun J (2016a) Deep residual learning for image recognition. Proceedings of the ieee conference on computer vision and pattern recognition (pp. 770-778)

He K, Zhang X, Ren S, Sun J (2016b) Identity mappings in deep residual networks. European conference on computer vision (pp. 630-645)

Zhu H, Zhu C, Xu Z (2018) Research advances on human activity recognition datasets. Acta Automatica Sinica 44(6):978–1004

Luo H, Wang C, Lu F (2018) Survey of video behavior recognition. J Commun 39(6):169

Huang J (2016) Chinese word segmentation analysis based on bidirectional lstmn recurrent neural network. Nanjing University Jiangsu

Kazakos E, Nagrani A, Zisserman A, Damen D (2021) Slow-fast auditory streams for audio recognition. Icassp 2021-2021 ieee interna- tional conference on acoustics, speech and signal processing (icassp) (pp. 855-859)

Kondratyuk D, Yuan L, Li Y, Zhang L, Tan M, Brown M, Gong B (2021) Movinets: Mobile video networks for efficient video recognition. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 16020-16030)

Korban M, Li X (2020) Ddgcn: A dynamic directed graph convolutional network for action recognition. European conference on computer vision (pp. 761-776)

Lan Z, Zhu Y, Hauptmann AG, Newsam S (2017) Deep local video feature for action recognition. Proceedings of the ieee conference on computer vision and pattern recognition workshops (pp. 1-7)

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

Li B, Li X, Zhang Z, Wu F (2019) Spatio-temporal graph routing for skeleton-based action recognition. Proceedings of the aaai conference on artificial intelligence 33:8561–8568

Li C, Zhong Q, Xie D, Pu S (2019) Collaborative spatiotemporal feature learning for video action recognition. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 7872-7881)

Li D, Liu H, Zhang Z, Lin K, Fang S, Li Z, Xiong NN (2021) Carm: Confidence-aware recommender model via review representation learning and historical rating behavior in the online platforms. Neurocomputing 455:283–296

Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q (2019) Actional–structural graph convolutional networks for skeleton-based action recognition. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 3595-3603)

Li Y, Ji B, Shi X, Zhang J, Kang B, Wang L (2020) Tea: Temporal excitation and aggregation for action recognition. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 909- 918)

Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CG (2018) Videolstm convolves, attends and flows for action recognition. Comput Vision Image Understanding 166:41–50

Li Z, Liu H, Zhang Z, Liu T, Xiong NN (2021) Learning knowledge graph embedding with heterogeneous relation attention networks. IEEE Transactions on Neural Networks and Learning Systems

Liu S (2017) Video-based action recognition. Hebei Normal University

Liu H, Fang S, Zhang Z, Li D, Lin K, Wang J (2021) Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. IEEE Transactions on Multimedia

Liu H, Nie H, Zhang Z, Li Y (2021) Anisotropic angle distribution learning for head pose estimation and attention understanding in humancomputer interaction. Neurocomputing 433:310–322

Liu T, Liu H, Li Y, Zhang Z, Liu S (2018) Efficient blind signal reconstruction with wavelet transforms regularization for educational robot infrared vision sensing. IEEE/ASME Transactions on Mechatronics 24(1):384–394

Liu T, Liu H, Li Y, Chen Z, Zhang Z, Liu S (2019) Flexible ftir spectral imaging enhancement for industrial robot infrared vision sensing. IEEE Transac Indus Informatics 16(1):544–554

Long X, Gan C, Melo G, Liu X, Li Y, Li F, Wen S (2018) Multimodal keyless attention fusion for video classification. Proceedings of the aaai conference on artificial intelligence (Vol. 32)

Majd M, Safabakhsh R (2020) Correlational convolutional lstm for human action recognition. Neurocomputing 396:224–229

Muhammad K, Ullah A, Imran AS, Sajjad M, Kiran MS, Sannino G et al (2021) Human action recognition using attention based lstm network with dilated cnn features. Future Generation Comp Syst 125:820–830

Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding 150:109–125

Peng X, Zou C, Qiao Y, Peng Q (2014) Action recognition with stacked fisher vectors. Springer, Cham, pp 581–595

Qu Y, Xia M, Zhang Y (2021) Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow. Comput Geosci 157:104940

Ren P, Xiao G, Chang X, Xiao Y, Li Z, Chen X (2021) Nas-tc: Neural architecture search on temporal convolutions for complex action recognition. arXiv preprint arXiv:2104.01110

Shen X, Yi B, Liu H, Zhang W, Zhang Z, Liu S, Xiong N (2019) Deep variational matrix factorization with knowledge embedding for recommendation system. IEEE Transactions on Knowledge and Data Engineering 33(5):1906–1918

Shi L, Zhang Y, Cheng J, Lu H (2019a) Skeleton-based action recognition with directed graph neural networks. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 7912-7921)
Shi L, Zhang Y, Cheng J, Lu H (2019b) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 12026-12035)
Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 1227-1236)
Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199
Song L, Xia M, Jin J, Qian M, Zhang Y (2021) Suacdnet: Attentional change detection network based on siamese u-shaped structure. Int J Appl Earth Obser Geoinformat 105:102597
Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. Proceedings of the ieee international conference on computer vision (pp. 4489-4497)
Wang H, Kläser A, Schmid C, Liu C-L (2013) Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vision 103(1):60–79
Wang H, Schmid C (2013) Action recognition with improved trajectories. Proceedings of the ieee international conference on computer vision (pp. 3551-3558)
Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. European conference on computer vision (pp. 20-36)
Wang X, Miao Z, Zhang R, Hao S (2019) I3d-lstm A new model for human action recognition. Iop Conference Series: Mater Sci Engin 569:032035
Wu C, Zaheer M, Hu H, Manmatha R, Smola AJ, Krähenbühl P (2018) Compressed video action recognition. Proceedings of the ieee conference on computer vision and pattern recognition (pp. 6026–6035)
Xia M, Cui Y, Zhang Y, Xu Y, Liu J, Xu Y (2021a) Dau-net: a novel water areas segmentation structure for remote sensing image. Int J Remote Sensing 42(7):2594–2621
Xia M, Qu Y, Lin H (2021b) Panda: parallel asymmetric network with double attention for cloud and its shadow detection. J Appl Remote Sens 15(4):046512
Xia M, Wang K, Song W, Chen C, Li Y et al (2020a) Non-intrusive load disaggregation based on composite deep long short-term memory network. Expert Syst Applicat 160:113669
Xia M, Wang T, Zhang Y, Liu J, Xu Y (2021c) Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. Int J Remote Sensing 42(6):2022–2045
Xia M, Zhang X, Weng L, Xu Y et al (2020b) Multi-stage feature constraints learning for age estimation. IEEE Transact Informat Forensic Sec 15:2417–2428
Xiao F, Lee YJ, Grauman K, Malik J, Feichtenhofer C (2020c) Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740
Yan A, Wang Y, Li Z, Qiao Y (2019) Pa3d: Pose-action 3d machine for video recognition. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 7922-7931)
Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the aaai conference on artificial intelligence (Vol. 32)
Yan Y, Xu J, Ni B, Zhang W, Yang X (2017) Skeleton-aided articulated motion generation. Proceedings of the 25th acm international conference on multimedia (pp. 199-207)
Yang H, Gu Y, Zhu J, Hu K, Zhang X (2020) Pgcn-tca: Pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition. IEEE Access 8:10040–10047
Yang X, Tian Y (2014) Action recognition using super sparse coding vector with spatio-temporal awareness. European conference on computer vision (pp. 727-741)
Chen Y, Gao X (2018) The latest progress of deep learning. Comput Sci Appl 08(04):565–571
Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: Deep networks for video classification. Proceedings of the ieee conference on computer vision and pattern recognition (pp. 4694-4702)
Zhang Y (2018) Text sentiment analysis based on multiple lstm structures. Beijing University of Posts and Telecommunications
Zhang S, Gong Y, Wang J (2017) The development of deep convolution neural network and its applications on computer vision. Chinese J Comput 40(9):1–29
Zhang Z, Li Z, Liu H, Xiong NN (2020) Multi-scale dynamic convolutional network for knowledge graph embedding. IEEE Transactions on Knowledge and Data Engineering

Zhang Z, Wang Z, Zhuang S, Huang F (2020) Structure-feature fusion adaptive graph convolutional networks for skeleton-based action recognition. IEEE Access 8:228108–228117
Zhao J, Snoek CG (2019) Dance with flow: Two-in-one stream action detection. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 9935-9944)
Ren Z, Xu H, Feng S, Zhou H, Shi J (2017) Sequence labeling chinese word segmentation method based on lstm networks. Appl Res Comput 34(5):1321–1324
Zhou Y, Sun X, Zha Z, Zeng W (2018) Mict: Mixed 3d/2d convolutional tube for human action recognition. Proceedings of the ieee conference on computer vision and pattern recognition (pp. 449–458)
Zhu Y, Li X, Liu C, Zolfaghari M, Xiong Y, Wu C, Li M (2020). A comprehensive study of deep video action recognition. arXiv preprint arXiv:2012.06567