The Visual Computer
https://doi.org/10.1007/s00371-023-03073-9

ORIGINAL ARTICLE

BDNet: a method based on forward and backward convolutional networks for action recognition in videos

Chuanjiang Leng[1] · Qichuan Ding[1] · Chengdong Wu[1] · Ange Chen[1] · Huan Wang[1] · Hao Wu[2]

Accepted: 27 July 2023

Abstract
Human action recognition analyzes the behavior in a scene according to the spatiotemporal features carried in image sequences. Existing works suffers from ineffective spatial–temporal feature learning. For short video sequence, the critical challenge is to extract informative spatiotemporal features from a limited-length video. For long video sequences, combining long-range contextual information can improve recognition performance. However, conventional methods primarily consider modeling the action's spatiotemporal features along a single direction, which is difficult to consider context information and ignores the information from the opposite direction. This article proposes a bi-directional network to simulate the bi-directional Long Short-Term Memory (Bi-LSTM) processing of time series data. Specifically, two 3D Convolutional Neural Networks (3D CNNs) extract spatiotemporal features along the forward and backward image sequence of action for each modality individually. After integrating the features of each branch, a dynamic-fusion strategy is applied to obtain a video-level prediction. We conducted comprehensive experiments on the action recognition dataset UCF101 and HMDB51 and achieved 98.0% and 81.4% recognition accuracy, respectively, with a reduction of three quarters of the inputting RGB images.

Keywords Action recognition · Convolutional neural network · Bi-directional network · Spatiotemporal features

1 Introduction

1.1 Backgrounds

The rapid development of imaging equipment results in massive videos generation. This proposes a significant new challenge in analyzing human action in videos for searching, ranking, and intelligent recommendation tasks [1–4]. The primary action recognition methods can be categorized as either deep learning feature-based and hand-crafted feature-based methods [5, 6]. In the last decade, depth learning methods have been widely used in human action recognition thanks to their capability of automatically extracting spatiotemporal features from image sequences, which significantly improves recognition accuracy over traditional methods. The commonly used depth network structures include Convolutional Neural Networks (CNNs) [7–12], Recurrent Neural Networks (RNNs) [13–15], and Graph Convolution Networks (GCNs) [16–18], among which RNNs and GCNs cannot directly process raw video images, while CNNs can directly extract features from images and achieve excellent performance, thus are widely used in action recognition. In practical deployments, especially in low-power mobile and wearable devices, Human action recognition (HAR) needs to achieve a trade-off between recognition accuracy and recognition efficiency [19–21].

1.2 Limits of priors

In practice, video adds a time dimension compared with static images, thus there is a temporal relationship between different images in the video. However, images varied in time reflect the motion information of action and are represented by the optical flow images. Therefore, the two-stream network structures [7, 22–25] that combine the optical flow and RGB images are commonly used to recognize human

✉ Qichuan Ding
dingqichuan@mail.neu.edu.cn

✉ Chengdong Wu
wuchengdong@mail.neu.edu.cn

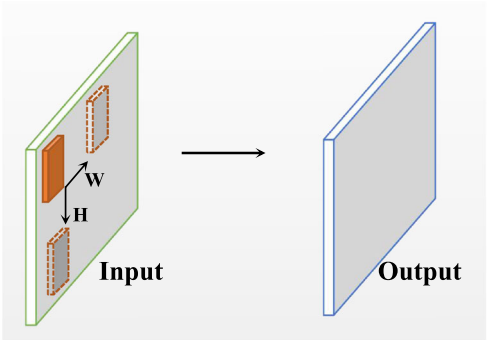
[1] Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110169, China

[2] Swinburne University of Technology, Sydney, Australia

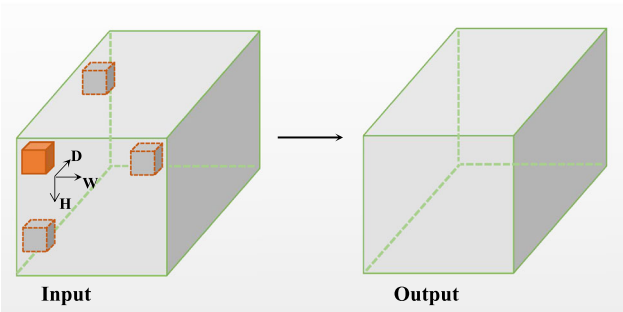
Published online: 17 October 2023

action. Existing methods [22, 24] use CNNs as a backbone network which can extract the action's spatial and temporal features. The commonly used 2D convolution is shown in Fig. 1a, which respectively operates along the width and height of the image, resulting in missing abundant temporal information for an action sequence. However, the 3D convolution operation, shown in Fig. 1b, operates along the width, height, and depth of the video volume, which models the spatiotemporal feature of action jointly and obtains better performance [26]. Although these methods have achieved promising performance, they only operate convolution along a single direction of the image sequence, making it hard to apply long-range sequence context information for recognition and also ignoring other information from the video.

Another problem in a video-based two-stream action recognition network is that it does not distinguish the importance of spatial and temporal features for each video. Weighing fusion is frequently adopted to integrate spatial and temporal features [7, 8], whereas the weight is fixed for each video. Intuitively, to recognize some actions, such as playing the piano, spatial features should take a more significant proportion in the recognition process because the player's actions change slightly, so finding motion is difficult in actions. For the action with violent motion change, such as playing soccer, the temporal feature is more critical in recognition due to the background and characters in the video do not change much. Therefore, a relatively small number of RGB images are able to provide abundant spatial features for recognition. Other fusion methods, such as max pooling and convolution fusion, both adopt a fixed manner to fuse spatial and temporal for all videos.



(a) 2D Convolution



(b) 3D Convolution

Fig. 1 2D and 3D convolution operations. **a** Applying 2D convolution on image results in an image. **b** Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal

1.3 Research motivation

Generally speaking, video is sequence time series data in which images can be organized from front to back (forward) or back to front (backward). In contrast, videos in different directions show differences in spatiotemporal characteristics. Bi-LSTM has achieved excellent performance in natural language processing because it can integrate information from two directions of time series data and solve the problem of long-range dependence. However, the Long Short-Term Memory (LSTM) cannot directly process the raw images due to its weakness in feature extraction of static images compared to CNN. In order to model temporal features of action, some methods [27–30] use LSTMs to process the extracted image features by CNNs, which makes the network structure more complex. However, these methods still lose a lot of temporal information due to using 2D CNNs to extract features from images. Inspired by bi-directional LSTM, an intuitive idea is to consider CNN to simulate simultaneously integrating contextual spatiotemporal information from both directions of an action sequence. Therefore, this study proposes to use 3D convolution to extract spatiotemporal features from different directions along the time dimension to avoid the loss of a large amount of temporal information.

1.4 Main contribution

This study proposes a novel bi-directional network (BDNet) for action recognition to address the problems mentioned above. Specifically, two 3D CNNs are used to extract the spatial feature along the forward and backward image sequences of the action, respectively. Then two categories of features are integrated to obtain the spatial-level feature. Meanwhile, the temporal-level feature is also obtained using optical flow images similarly. Subsequently, a dynamic fusion strategy is used to integrate the spatial- and temporal-level features to obtain the video-level features. Finally, the proven Softmax classifier is applied to predict the action class.

The key contributions of this study are as follows:

1. A bi-directional network is proposed to model the spatial and temporal features along the forward and backward

action sequence respectively. Specifically, for the spatial stream, one 3D CNN is used to extract the feature of the forward action sequence, and the other extracts the feature in the opposite direction. For the temporal stream, a similar operation is executed using optical flow images. Subsequently, integrating of both features obtains spatial- and temporal-level features for the two-stream network.
2. A dynamic-fusion strategy that can distinguish the importance of spatial and temporal features is proposed to fuse each video's spatial-level and temporal-level features. Specifically, the different importance of spatiotemporal features in the video can be distinguished by the proportion of the optical flow image with rich motion information in the video.
3. The bi-directional structure enlarges the perception scope of the network and achieves the balance between accuracy and efficiency. That is, the proposed method achieves a comparable accuracy but uses fewer inputs, which improves the recognition speed of the network.

2 Related work

The duration of action in action recognition is typically different due to the discrepancy in action and actors. Fully exploring the information from limited image sequences is a problem that deserves studying in human motion recognition. Previous works [5, 6, 31–33] mainly focus on designing hand-crafted features, such as HOG3D [34], HOF [35], and iDT [31], i.e., for human action recognition. Subsequently, deep learning-based methods [22–25, 36] have become the mainstream, and many deep networks have been proposed and achieved promising performance. Among them, the two-stream network [22–24] is widely used in some current action recognition algorithms because it combines spatial and motion information of action. To make full use of the information carried by videos, some works [22, 37, 38] attempt to use audio, caption, and other modalities as supplementary information to enhance recognition performance. To deeply learn the action's spatial appearance and temporal cues, Li et al. [39] proposed decomposing volumetric video data along three orthogonal views and performing 2D convolution, respectively. Shou et al. [36] used current residuals and motion vectors in the compressed video to replace flow computation to save computational cost, which improves efficiency, but it sacrifices recognition accuracy to some extent. Wang et al. [22] explored the RGB difference and warped optical flow fields as additional modalities for two-stream Convolutional Networks (ConvNets). Wang et al. [40] exploited the conjoint information in multiple heterogeneous sources cooperatively using RGB visual features and depth features to train a convolutional neural network and deeply aggregate the two kinds of features for action recognition. Visual signals are rarely present without accompanying audio, so fusing audiovisual information can improve some single-modality visual analysis systems. Min et al. [41] explored the audio and video (A/V) quality, which is used to compare and develop A/V quality measurement models and algorithms. It provided a measurement of multi-modality information for the action recognition system, which can determine the contribution of each modality for each action sequence. To make a comprehensive judgment for recognizing action, Long et al. [37] leveraged the audio as an additional information source to alleviate action recognition by attention fusion for RGB, Flow, and Audio features. Nagrani et al. [38] noticed the link between spoken words and actions in movies and proposed to learn this correlation without additional supervision. Although additional modes provide much supplementary information and improve the accuracy of identification, these methods cannot enlarge the receptive field of the network and enhance the ability to contact the context.

Furthermore, as time series data, the backward sequence also provides complementary information compared with the forward sequence of videos. Many works have used a bi-directional structure for detection and recognition [27–30]. To model long-term temporal dynamics within and between actions, Singh et al. [30] proposed to use the multi-stream CNN followed by a bi-directional Long Short-Term Memory layer. Ullah et al. [28] learned the sequential information among CNN frame features using a Deep Bi-directional LSTM which increases the depth by stacking multiple layers together in the forward and backward pass of it. He et al. [27] stacked and integrated the spatial and temporal features extracted by CNNs with a bi-directional LSTM to model the visual and temporal associations in both forward and backward directions. Ullah et al. [29] adopted a multi-layer Bi-directional Long Short-term Memory model accurately classify anomalous/normal events in complex surveillance scenes that used the features extracted from the sequence of frames with the CNN model. The above methods achieved a promising performance, but they adopted an LSTM using features extracted by CNNs, which was limited to 2D features of images. Therefore, we attempted to use 3D convolution to simulate the bi-directional structure, which can model spatiotemporal features better and consider backward information of video sequences.

Another model the long-range contextual independence is adding a self-attention mechanism to action recognition. Girdhar et al. [42] introduced the Action Transformer model, which aggregates features from the spatiotemporal context around the person for recognizing and localizing human actions in video clips. Liu et al. [43] proposed the Swin Transformer designed for the image domain and leveraged

the power of pre-trained models, which advocated an inductive bias of locality in video Transformers and led to a better speed-accuracy trade-off. To explicitly model different spatiotemporal resolutions, Yan et al. [44] proposed the multiview transformers for video recognition by separating encoders to represent different views of the input video with lateral connections to fuse information across views. Mazzia et al. [45] introduced an Action Transformer which exploits 2D pose representations over small temporal windows, limits computational and energy requests, and provides a low latency solution for accurate and effective real-time performance. These attention-based methods demonstrate outstanding advantages, such as capturing global and local connections in one step and capturing long-term dependencies without being limited by sequence length. In addition, compared to CNN and RNN, the model has fewer parameters and lower complexity. However, these methods remain shortcomings, such as challenging to capture positional information, namely, learning the sequential relationships in the sequence. For multi-modality information fusion, an effective method is to fuse modal features, i.e., to process the raw data in different methods and then fuse the processed results [22, 37, 38]. Some early methods [22–24] adopted weighted fusion to integrate spatial and temporal features. Subsequently, attention-based methods were adopted to pay more attention to most information features across different modalities. However, these methods were usually complex and needed vast calculations to update weights. More recently, many methods [27–30] used LSTM to integrate spatiotemporal features of different modalities. Although they acquired excellent recognition performance, they also need complex structures and more calculations.

To tackle the problems existing in the previous study, BDNet is proposed in this study to enlarge the network's receptive fields and enhance the ability to contact the context for action recognition. Inspired by the Bi-LSTM network, the backward information of the sequence data can help the network to contact further context information and expand the receptive field of the network. Considering that 3D convolution can extract spatiotemporal features along the time dimension, we attempt to use 3D convolution to extract forward and backward features of video to simulate the bi-directional process for enlarging the perceptional range of the network. Here, two 3D CNNs extract the features of the action along the forward and backward image sequences, respectively. Then, the bi-directional features are integrated to obtain the spatial or temporal features. Finally, a dynamic-fusion strategy assigns different importance scores to each video's spatial and temporal features while avoiding recalculating the weights in the attention mechanism or more calculations in complex LSTM. Experimental results show that the BDNet is superior to previous methods for action recognition.

3 Approach

3.1 Network architecture

The structure of the BDNet is shown in Fig. 2. In the two-stream network framework, the spatial and the temporal stream branches, BDSNet and BDTNet, are designed to process RGB and Flow images. Each branch uses two 3D CNNs to extract image features along the forward and backward input image sequence. Then, the features of the two directions are aggregated to double the aware scope of the network and combine the features of different times to integrate longer-range context information. Then, we use a dynamic-fusion strategy to assign a unique weight value to the features obtained from each branch, which measures the importance of spatial and temporal features. Finally, we get the result of action recognition by a Softmax classifier.

3.2 Bi-directional two-stream network

The core idea of the bi-directional two-stream network is inspired by Bi-LSTM, which simultaneously trains two sequences of LSTM units, one for forward and another for backward sequence. It combines both LSTM units' output to obtain the outcome for each Bi-LSTM [47]. The structure of Bi-LSTM is shown in Fig. 3a. Similarly, we construct a Bi-CNN named BDNet, shown in Fig. 3b. Specifically, for a given video sequence containing T frames, the RGB images sequence $V \in R^{T\times W\times H\times 3}$ is a matrix where H and W denote the height and weight of the image, respectively. In contrast, the backward RGB sequence $V' \in R^{T'\times W\times H\times 3}$ is opposite to the RGB sequence, which T denotes the opposite sequence's T frames. The construction of optical flow and reverse optical flow image sequences $F \in R^{T\times W\times H\times 2}$ and $F' \in R^{T'\times W\times H\times 2}$ are similar to RGB and reverse RGB image sequences. We calculate the backward optical flow images $F' \in R^{T'\times W\times H\times 2}$ opposite the action sequence using the same $TV-L^1$ [48]. The bi-directional two-stream network models an action sequence as follows:

$$BDSNet\,(T_1, \cdots, T_K) = H\left(G_1\,(T_1, \cdots, T_k;\, W_1),\, G_2\left(T_1', \cdots, T_k';\, W_2\right)\right) \quad (1)$$

$$BDTNet(F_1,\, F_2, \cdots, F_k) = H(G_3(F_1,\, F_2, \cdots, F_k;\, W_3),\, G_4(F_1',\, F_2',\, \ldots,\, F_k';\, W_4)) \quad (2)$$

where $G_1(\cdot)$, $G_2(\cdot)$, $G_3(\cdot)$ and $G_4(\cdot)$ denote 3D ConvNets with parameters W_1, W_2, W_3 and W_4. T_i, T_i', F_i and F_i' respectively represent i_{th} frame in the forward and backward RGB and Flow sequences. H is the function that integrates the forward and backward features for each branch. For the

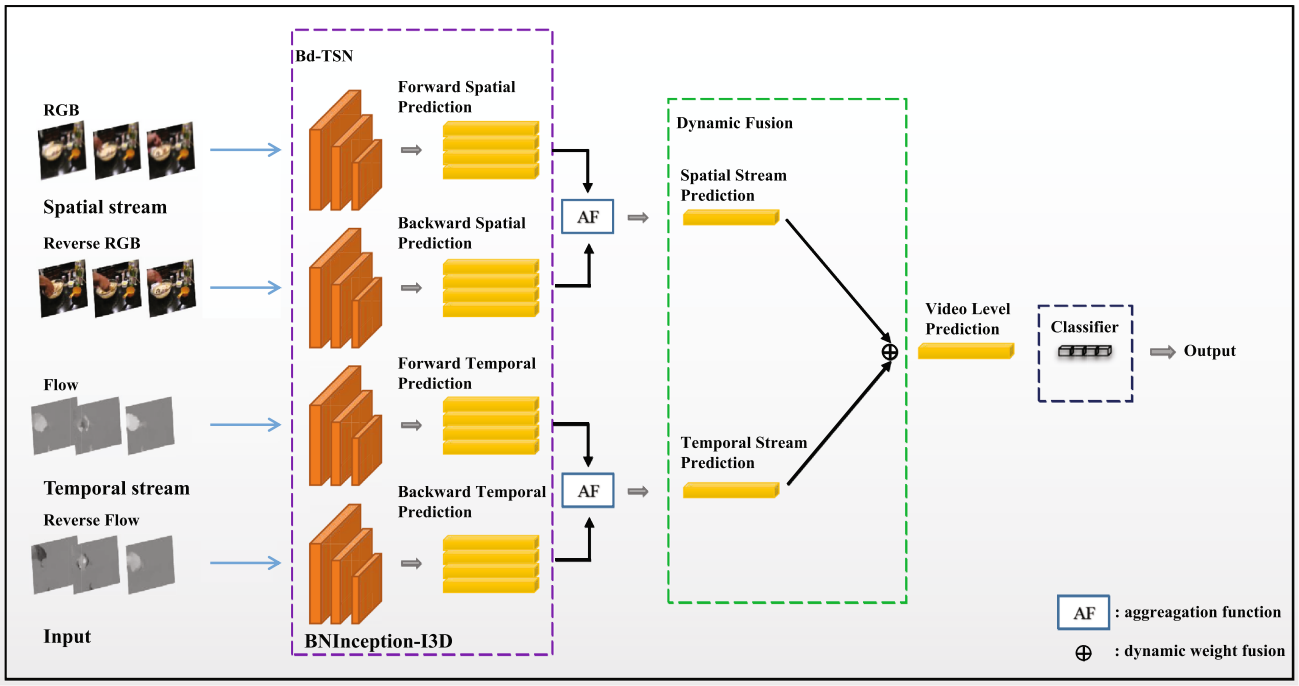


Fig. 2 The flowchart of the proposed BDNet for action recognition. RGB and Reverse RGB images are input into the Inception with Batch Normalization Inflated 3D ConvNet (BNInception-I3D) [46], followed by an aggregation function (AF) in the Spatial branch. AF is an aggregate function that integrates the features from the forward and backward action sequences. A similar operation is in the temporal branch. After obtaining the spatial and temporal features, a dynamic-fusion strategy is used to assign importance scores for the spatial and temporal features. Finally, a frequently-used Softmax classifier is used to produce a prediction result

loss function, we use standard categorical cross-entropy loss, which is formed as:

$$L(y, G) = -\sum_{i=1}^{N} y_i \left(G_i - \log \sum_{j=1}^{N} \exp G_j \right), \tag{3}$$

where N denotes the number of action classes and y_i is the label for the action class i.

3.3 Dynamic-weight fusion strategy

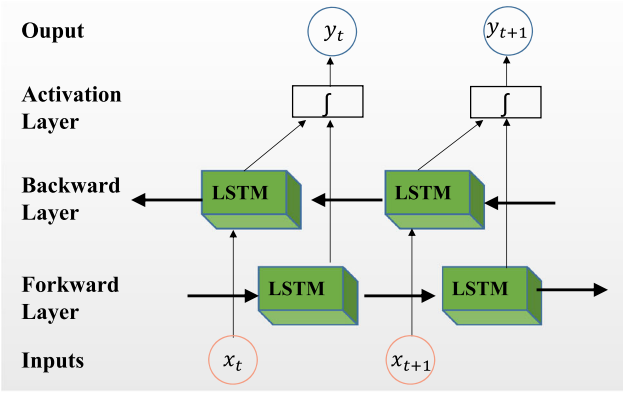
A critical step for the two-stream network is feature fusion. Most previous works adopt a fixed fusion weight for all videos. However, the spatial and temporal features should not share a fixed weight due to their different importance for different videos. Optical flow describes the object's motion, so the number of grayscales in the optical flow images can measure the motion information. And the fewer optical flow images containing fewer grayscale levels in a video, the smaller the proportion of temporal features in later fusion, and the spatial features dominate in recognition. On the contrary, spatial and temporal features can consider equally important in late fusion. Based on this, we construct a dynamic fusion strategy. Specifically, judge whether the video contains motion information according to the number of grayscales of optical flow images in a video first. Then, count the number of such images and calculate their proportion in all images as a weight for the video's spatiotemporal features fusion. The calculation can be described by Eqs. (4) and (5).

$$\eta = \frac{\sum_i \min(\lfloor \mathrm{H}(F_i)/g_t, 1 \rfloor)}{n}, \; i = 1, \cdots n \tag{4}$$

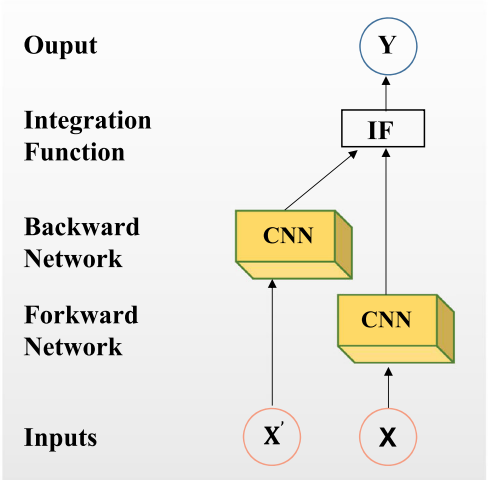
$$\begin{aligned} y = &(1-\eta) \cdots BDSNet(T_1, T_2, \cdots, T_k) \\ &+ \eta \cdots w \cdots BDTNet(F_1, F_2, \cdots, F_k) \end{aligned} \tag{5}$$

where η presents the proportion of informative frames, min() denotes the minimal function, n is the video frame number, $H(\cdot)$ is the grayscales statistics function, g_t denotes a grayscale threshold, $\lfloor \cdot \rfloor$ is the operation of down-rounding, w is a constant weight value.

After that, a Softmax function is used to predict the final output based on the integrated spatial–temporal feature.



(a) Bi-directional LSTM



(b) Bi-directional CNN

Fig. 3 The bi-directional LSTM and bi-directional CNN structures. **a** Operation of the bi-directional LSTM structure. **b** Operation of the bi-directional CNN structure

4 Experiments

4.1 Dataset

We used the open dataset UCF101 and HMDB51 to evaluate the proposed method. The UCF101 dataset contains 13,320 videos of 101 action categories, mainly divided into five categories: Body motion, Human interactions, Human object interactions, Playing musical instruments, and sports. All videos are collected from YouTube with a fixed frame rate of 25 FPS and image resolution of 320 × 240, and the total length of these video clips is over 27 h. In contrast, the HMDB51 dataset contains 6849 realistic video clips collected from various sources, including movies and web videos. We follow a standard evaluation scheme using three different training/testing splits, which split the dataset into three parts, one for testing and the others for training, and measure the method's performance with the average accuracy over these three splits.

4.2 Implementation details

PyTorch is used to implement the proposed method. We adopt the BNInception-I3D as the backbone to construct the BDNet. Here, a classic two-stream structure is used for action recognition, where the BDSNet and BDTNet are trained to process RGB and optical flow images, respectively. For all backbones, we initialized the network parameters with the pre-trained model in Kinetics and the mini-batch stochastic gradient descent algorithm to update the model parameters. The network's input is a 64-frame clip constructed by randomly selecting one frame from every five frames. When the video frame number is insufficient, copy the video repetitively until it meets the video length condition. The optical flow images are calculated with $TV-L^1$ using CUDA.

When training the BDSNet, we set the values of epochs, momentum, batch size, weight decay, and learning rate to 120, 0.7, 20, 5×10^{-4}, and 0.001, respectively. For BDTNet, we set these parameters to 340, 0.7, 24, 5×10^{-4}, and 0.001. The learning rate drops to 1/10 of the current value after 40, 80, 120, and 290, respectively. The experiments are conducted on the platform with an Intel(R) Xeon(R) Silver 4114 CPU @2.20 GHz and four TITAN GPUs(12G).

4.3 Ablation study

In this part, we conducted detailed ablation studies on the UCF101 Dataset to verify the effectiveness of the proposed method. Here, for the convenience of description, we named the network branch that processes RGB images as the spatial stream network (SpaNet), and the network branch that processes optical flow images as the temporal stream network (TempNet). Additionally, we named the following branch networks in our BDNet.

1. *BS-For* It can be regarded as the baseline of the spatial stream, which is used to extract RGB images' features along the forward action sequence.
2. *BS-Back* This variant extracts the backward spatial features.
3. *BT-For* It is the baseline of the temporal stream, which is used to process the forward Flow images.
4. *BT-Back* This variant extracts the backward temporal features.
5. *BDSNet* This variant is the complete version of our spatial stream, which integrates the forward and backward spatial features.
6. *BDTNet* This variant integrates the forward and backward temporal features.

Table 1 The recognition (%) of all variants on the three splits of the UCF101 dataset

Method	Split-1	Split-2	Split-3
BDSNet-F	95.2	95.3	94.9
BDSNet-B	94.8	94.8	94.9
BDTNet-F	94.7	95.2	95.8
BDTNet-B	94.5	95.0	94.6
BDSNet	95.3	95.3	95.2
BDTNet	95.2	95.5	95.9

As multiple network variants were included in the proposed BDNet, a preliminary experiment was conducted to evaluate the performance of different variants on the UCF101. The experimental results are recorded in Table 1. As can be seen from Table 1, the spatial and temporal stream networks achieve 95.2%, 95.3%, 94.9%, 94.7%, 95.2%, and 95.8% accuracy on three splits, respectively. Although the Inflated 3D ConvNet (I3D) network is promising in recognizing actions with a single network, its receptive field is still limited due to the video length and cannot be enlarged. However, as supplementary information, the backward sequence performs similarly to the forward consequence. After integrating the features of the forward and backward action sequence, the spatial (BDSNet) and temporal network (BDTNet) achieves 95.3%, 95.3%, 95.2%, 95.2%, 95.5%, 95.9% accuracy on three splits, respectively. Compared with the base model (BS-For and BT-For), the adding networks improved the accuracies by 0.1% 0.5%. That is because a bi-directional structure can enlarge the perception range of the network, and the backward Flow images calculated with $TV - L^1$, which contains some different motion information, positively affect action recognition.

A critical step for the proposed bi-directional network is to integrate the spatial and temporal features of the forward and backward networks. General integration methods [8] include Max fusion, Sum fusion, and weigh fusion. In this section, we examined the impacts of various integrations of the forward and backward networks. Table 2 respectively lists the recognition accuracies on Split-1 of the UCF101 obtained by the BS-For, BS-Back, BT-For, and BT-For, and the integrating results of forward and backward features acquired by Max fusion, Sum fusion, and Weigh fusion. These single networks achieve recognition accuracy of 95.2%, 94.8%, 94.7%, and 94.5%. For the spatial stream, three fusion methods integrating forward and backward networks' features achieve 95.0%, 95.1%, and 95.3%, respectively, while the temporal stream is 95.1%, 95.1%, and 95.2%. As can be seen, the results of various integration strategies are similar, and the weighted fusion achieves a slightly higher performance. Compared with the recognition results of a single network, the recognition accuracies of weighted fusion increase by 0.1%, 0.5%, 0.5%, and 0.7%, respectively. For spatial stream networks, the features of aggregated forward and backward networks do not significantly improve the recognition results. However, the three aggregation methods significantly improve the recognition results for the temporal stream network. A reasonable explanation is that the spatial information in action changes little, and the action background and actors do not change in an extensive range. A small number of RGB images can well represent the spatial features of the action. As the optical flow image represents the motion information, it is easier to capture the change of motion information in temporal. The bi-directional structure can make the network perceive the context information in a broader range, consistent with the conclusion that Bi-LSTM can process the temporal sequence data well. Therefore, we use weighted fusion to aggregate forward and backward spatiotemporal features in subsequent research.

Adopting a bi-directional structure models the spatiotemporal features of action that can enlarge the network's perception scope, which means fewer frames can achieve comparable performance compared with the baseline model. Next, we evaluated the recognition performance and speed of the network with different input frames and variants. The input frame number was respectively set to 250 and 64. The evaluated network structures include baseline I3D (BS-For and BT-For) and the proposed methods BDNet (BDSNet and BDTNet). Table 3 lists the recognition accuracies and speed on Split-1 of the UCF101 obtained by BS-For, BT-For, BDSNet, and BDTNet. With the input frame number 250, the time and accuracies respectively were 0.247 s, 0.233 s, 94.8%, and 94.7% for the Baseline (BS-For and BT-For), and 0.243 s, 0.235 s, 94.5%, 95.9% for the proposed method (BDSNet and BDTNet), respectively. When the input frame number is 64, the counterpart result was respectively 0.076 s, 0.069 s, 95.2%, 94.5%, 0.076 s, 0.069 s, 95.3%, and 94.8%. When the input frame number is the same, the recognition speed for all variants is proximate. However, the SpaNet takes longer than the TempNet and is longer at 0.014 s, 0.007 s, 0.008 s, and 0.007 s, respectively. The reason is that RGB images have three channels and need to process more information than two-channel optical flow images.

For the baseline and the proposed method, when inputting 64 frames, the spatial stream network obtains better recognition accuracies, 95.2% and 95.3%, which is improved by 0.4% and 0.8% compared with 250 frames. A reasonable explanation is that RGB images have fewer background changes and more redundancy, and many redundant images hinder the correct recognition of actions. For the temporal stream network, when the input is 250 frames, it achieves better accuracy of 94.7% and 95.9%, with an increase of 0.2%

Table 2 The recognition accuracy of the different branch networks (%) and three fusion methods accuracy (%) of different integration of the forward and backward networks on the Split-1 of the UCF101 dataset

Models	Accuracy	Max	Sum	Weigh
BDSNet-F	95.2	95.0	95.1	95.3
BDSNet-B	94.8			
BDTNet-F	94.7	95.1	95.1	95.2
BDTNet-B	94.5			

Table 3 The recognition accuracy (%) and speed (s/video) of all variants under different input on the Split-1 of the UCF101 dataset

Method	Frames	Time	Accuracy	Fusion
BDSNet-F	250	0.247	94.8	97.7
BDTNet-F	250	0.233	94.7	
BDSNet-F	64	0.076	95.2	97.2
BDTNet-F	64	0.069	94.5	
BDSNet	250	0.243	94.5	97.7
BDTNet	250	0.235	95.9	
BDSNet	64	0.076	95.3	97.4
BDTNet	64	0.069	94.8	
BDSNet	64	0.076	95.3	97.8
BDTNet	250	0.235	95.9	

and 1.1%, respectively. The possible reason is that the optical flow image carries less information, so the more optical flow images are input, the greater the probability of obtaining different information, which will promote the recognition performance of the temporal stream network. For the final fusion result of the Baseline and the proposed method, when the input is 250 frames, the recognition result reaches 97.7% and 97.7%, respectively. When the input is 64 frames, the recognition results are 97.2% and 97.4%, respectively. Therefore, we set the input to 64 and 250 for the spatial and temporal stream network, achieving an accuracy of 97.8%. At the same time, the required computing time is reduced by 0.19 s compared with Baseline. We use this configuration in subsequent experiments.

Although a single I3D network has achieved excellent recognition, the fusion of spatial and temporal features can further improve the recognition results. Here, to verify the superiority of the dynamic fusion strategy, we designed an experiment on the three splits of UCF101 using fixed-weight (1 and 1.5 for the two-stream) and dynamic-weight fusion (BDNet). The dynamic weight is computed according to the proportion of optical flow images with rich motion information expressed as the number of grayscales in the image in a video. The calculation process can be described by Eq. (4). In addition, the example of the grayscales of flow images is shown in Fig. 6, from which it can be found that flow images with much motion information have more grayscales that can distinguish the importance of spatial and temporal features. The dynamic-weight fusion can be described by Eq. (5), and the experimental results are reported in Table 4. As can be seen, even using the fixed-weight fusion, the proposed method achieves comparable recognition results compared with the method extracting features along a single direction but uses fewer frames to improve the recognition speed. Dynamic-weight fusion recognition effect is higher than fixed-weight fusion, improving by 0.1%, 0.2%, and 0.1%, respectively. To compare the recognition effect, we visualized the recognition results of the proposed method and the base model I3D, as shown in Figs. 4 and 5. It can be found that the proposed method can recognize some actions that I3D cannot recognize, and the recognition result is better than I3D. The experimental results show delicate improvement and one possible reason we ignore performing additional image augmentation during testing. However, in action recognition systems, high-quality images are crucial for the success of the recognition system, while low-quality images may degrade the recognition system's performance [49].

Table 4 The recognition (%) of different fusions on the three splits of the UCF101 dataset

Method	Split-1	Split-2	Split-3
BS-For + BDTNet + fixed	97.8	97.8	98.1
BDSNet + BDTNet + fixed	97.8	97.7	98.1
BS-For + BT-For + dynamic	97.9	97.8	98.1
BDSNet + BDTNet + dynamic	98.0	97.9	98.2

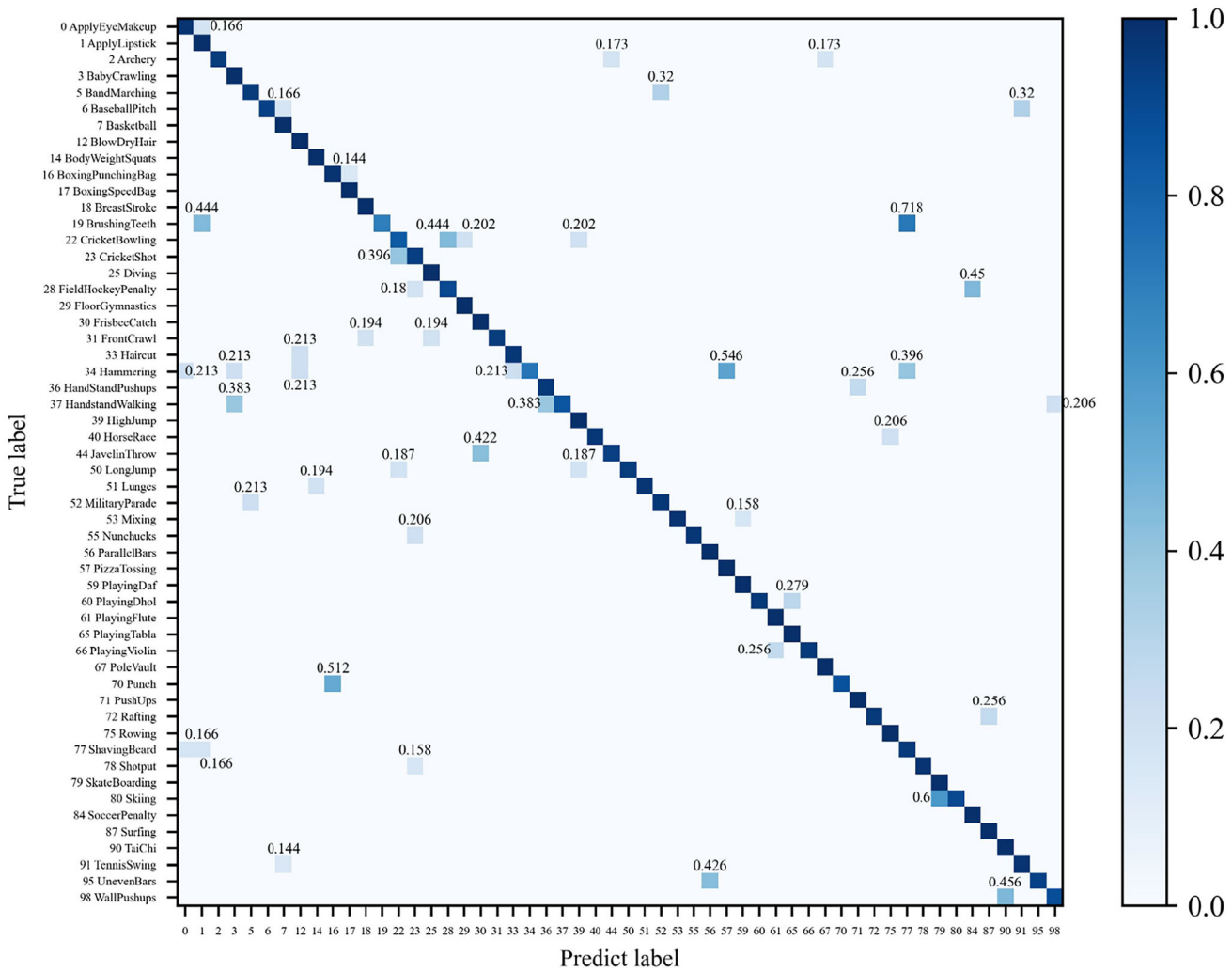


Fig. 4 The confusion matrix of the recognition result adopted the proposed method on Split-1 of the UCF101. Here, we only visualized the action categories that were identified wrong, and all probability values are expressed in percentage (% is omitted)

Furthermore, we evaluated the effectiveness of our method BDNet with I3D and Temporal Segment Networks (TSN) [22] on Split-1 of UCF101. The results are reported in Table 5 and show that the SpaNet takes less time than the I3D, and the TempNet takes more time, the total time of BDNet is shorter 0.239 s and 0.103 s than that of I3D and TSN, respectively. Compared with the I3D, we can use fewer frames (64 frames for SpaNet) to model the long-range video due to using the bi-directional sequence but obtaining a comparable accuracy. I3D adopts 250 frames as input to cover the whole video, which will take more time to achieve high-precision recognition. Compared with I3D and TSN, our BDNet achieves similar or better performance with lower FLOPs, including SpaNet and TempNet of I3D (with 48.9% and 48.9% fewer FLOPs) and TSN (with 5.48% and 15.9% fewer FLOPs). Our method also has fewer parameters than TSN (with 42.2% and 42.4% less Param). The BDNet requires more memory, but the increase in memory is still acceptable, especially in light of the improvement in recognition accuracy and efficiency of BDNet.

4.4 Comparison with the state of the art

Finally, to further illustrate the effectiveness of the proposed BDNet, we compared it with the state-of-the-art methods, including methods based on hand-crafted features and deep learning, in terms of action recognition over all three splits of UCF101 and HMDB51. The recognition results of different models are listed in Table 6. Most deep learning methods perform better than traditional hand-craft feature methods. In addition, comparing the proposed method with methods based on deep learning, as can be seen, the BDNet (98.0%, 81.4%) achieves a higher recognition accuracy than most methods based on a two-stream network using 2D convolution, including Two Stream (88%, 59.4% [23]), TSN (94.2%, 69.4% [22]), TSN + As (94.6%, 71.2% [25]), VideoMAE

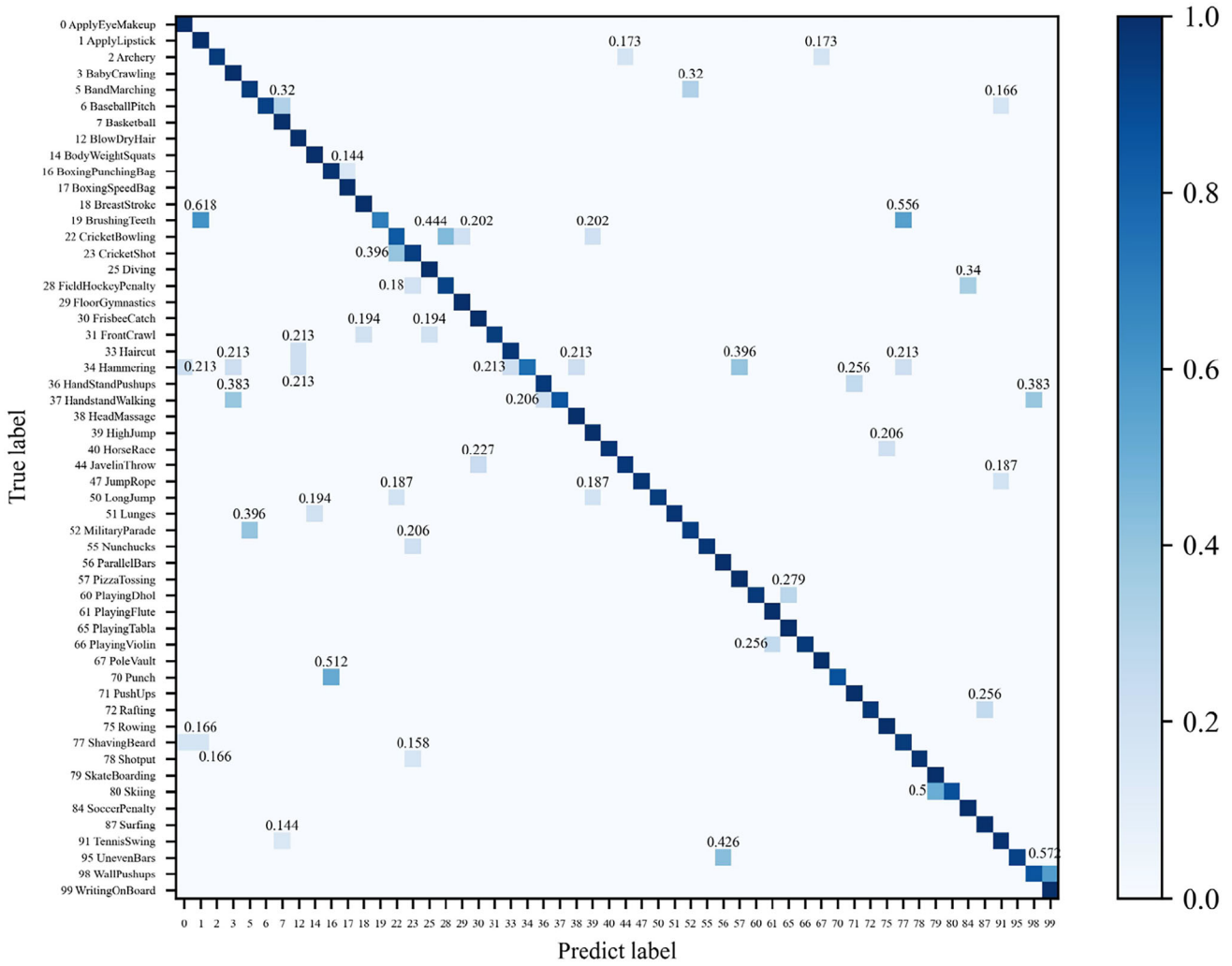


Fig. 5 The confusion matrix of the recognition result adopted the I3D on Split-1 of the UCF101

Table 5 The average recognition speed (s), Param(M) and FLOPs (G)of the BDNet, I3D and TSN on Split-1

Method	BDNet	I3D	TSN
SpaNet	0.076	0.247	0.165
TempNet	0.235	0.233	0.178
Fusion	< 0.001	< 0.001	< 0.001
Total	0.241	0.480	0.344
SpaNet Param(M)	24.68	12.34	42.71
TempNet Param(M)	24.63	12.32	42.73
SpaNet FLOPs(G)	223.01	436.33	235.94
TempNet FLOPs(G)	205.40	401.91	244.20
Accuracy (%)	98.0	97.8	94.1

(96.1%, 73.3% [53]) and BQN (97.6%, 77.6% [54]). Compared with the 2D CNN baseline TSN in [22], the BDNet improves recognition accuracy by 3.8% and 12.0%. Additionally, the proposed method outperforms VideoMAE [53] 1.9% and 8.1% on UCF101 and HMDB51, respectively. The proposed method outperforms the VideoMAE. Obviously, comparing the recognition approaches with 2D convolution, adopting 3D convolution to extract spatiotemporal features along the temporal dimension can avoid missing abundant temporal information. Therefore, the method based on 3D convolution has a better recognition effect. Although there are problems that require more time and computing resources for 3D CNNs, these problems will be solved with the development of computer technology (Fig. 6).

From Table 6, we also see that the recognition accuracy of the proposed method (98.0%, 81.4%) was higher than that of the LSTMs, consisting of VideoLSTM (89.2%, 56.4% [14]), BubbleNET (97.6%, 82.6% [13]). Therefore, adopting a bi-directional structure to enlarge the perception scope of the network to simulate the decision-making process of

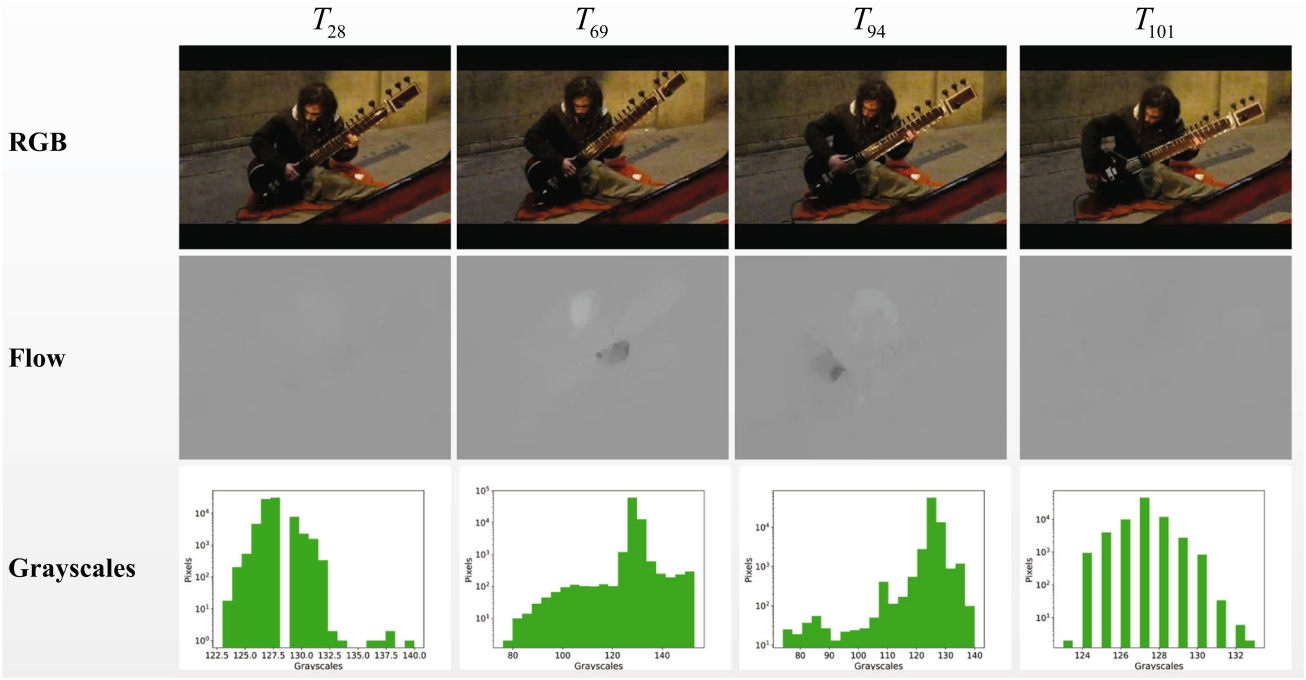


Fig. 6 An example of the RGB, optical flow images, and grayscale statistics. T_i is the index of the image frame in action, "Playing the Sitar." The first row is RGB images, the second is optical flow images, and the third is the grayscale statistical of the latter

Table 6 Comparison of the action recognition performances with the state-of-the-art methods on the UCF101 and HMDB51 datasets

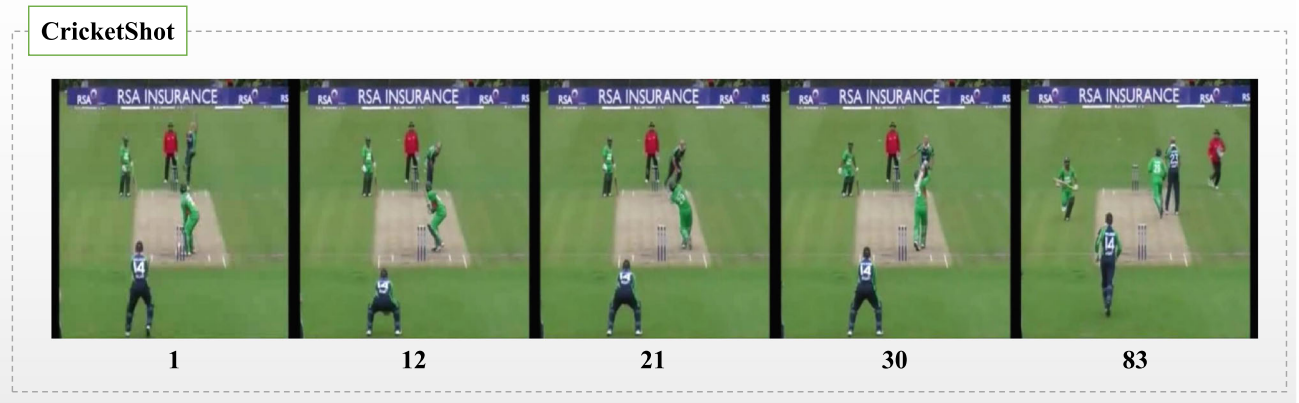
Method	UCF101 (%)	HMDB51 (%)
iDT + FV [31]	85.9	57.2
iDT + HSV [34]	87.9	61.1
Two Stream [23]	88.0	59.4
FsTCN [50]	88.1	59.1
C3D [26]	85.2	51.6
VideoLSTM [14]	89.2	56.4
Fusion [51]	92.5	65.4
TSN [22]	94.2	69.4
ST-ResNet + iDT [52]	94.6	70.3
STM-ResNet [52]	94.2	68.9
I3D + Kinetics [24]	97.8	80.2
TSN + AS [25]	94.6	71.2
I3D RGB + DMC-Net [36]	96.5	77.8
VideoMAE [53]	96.1	73.3
BubbleNET [13]	97.6	82.6
BQN [54]	97.6	77.6
VidTr [55]	96.7	74.4
Ours	98.0	81.4

bi-directional LSTM can improve the ability of CNNs to recognize actions. Furthermore, the proposed method also achieved a better or comparable performance than C3D (85.2%, 51.6% [26]), I3D (97.8%, 80.2% [24]), and I3D RGB + DMC-Net (96.5%, 77.8% [36]). Compared with I3D, the BDNet improves the recognition accuracy by 0.2% and 1.2% on the UCF101 and HMDB51, respectively. The reasonable explanation is that the intra-class differences in actions are small. As shown in Fig. 7, the three actions with similar backgrounds and motion modes result in recognition confusion. Compared with the actions of CricketBowling, CricketShot occurs in a moment and is prone to miss these key-frames in hundreds of frames of videos. In addition, adding target detection may help improve the recognition performance of the network due to the balls are small and often ignored in the video. Although the accuracy improves slightly, the proposed method BDNet, using a bi-directional structure, and adding dynamic fusion, achieves a comparative recognition accuracy and improves the recognition efficiency.

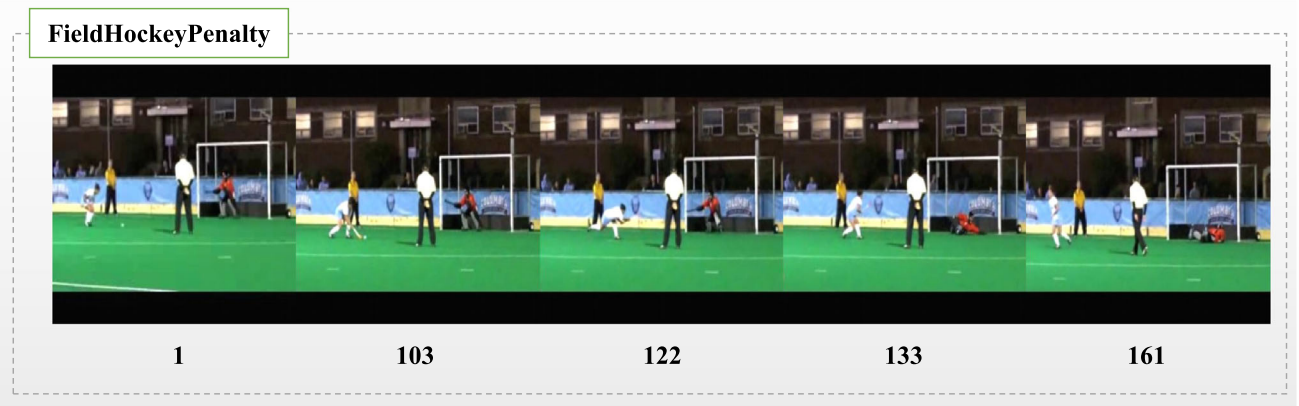
5 Conclusion

This paper proposes a bi-directional network (BDNet) to achieve effective action recognition capable of enlarging the range of input video and considering longer-range context information. In the BDNet, a bi-directional network structure is adopted to integrate the features along both the forward and backward video sequence, which can simulate the decision-making process of a bi-directional LSTM network for time series data. BDNet can obtain the spatial and temporal features that represent the appearance and motion information

(a) CricketBowling



(b) CricketShot



(c) FiledHockeyPenalty

Fig. 7 Three examples qualitatively explain why the actions are difficult to recognize in the UCF101 dataset. The number represents the frame's index in the video for each action. The figure's intervals are inconsistent because the distinguishing action happens at different time segments

of actions, respectively. Subsequently, a dynamic-weight fusion strategy assigns each video's spatial and temporal characteristics with different levels of importance. Specifically, the weight is computed by the proportion of optical flow images with rich motion information in optical flow images for each video, which further improves the adaptability of BDNet. Experiments conducted on the UCF101 and HMDB51 datasets demonstrate the promising capabilities of the proposed method in action recognition.

In the experiments, we found that the fusion of spatial and temporal features can significantly improve the recognition results of a single network. However, there are still circumstances where the recognition of a single network is correct, but the fusion results are wrong. The reasonable explanation is that there is little difference in inter-class for the appearance or motion of actions, and incorrect recognition easily dissembles the correct recognition of another network. Thus, the appropriate classifier will be further studied to address this problem in future work.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61973065, Grant U20A20197, and Grant 61973063, by the Joint fund of Science & Technology Department of Liaoning Province and State Key Laboratory of Robotics, China under Grant 2020-KF-12-02, by Liaoning Key Research and Development Project 2020JH2/10100040, by the Foundation of National Key Laboratory OEIP-O-202005, and by the Fundamental Research Funds for the Central Universities under Grant N182608004.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

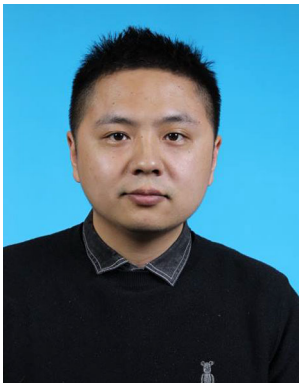
Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Shah, C., White, R.W.: Task intelligence for search and recommendation. Synth. Lect. Synth. Lect. Inf. Concepts Retr. Serv. **13**(3), 1–8 (2021)
2. Nguyen, P.-A., Ngo, C.-W.: Interactive search vs. automatic search: an extensive study on video retrieval. ACM Transact. Multimedia Comput. Commun. Appl. (TOMM) **17**(2), 1–24 (2021)
3. Zarmehi, N., Amini, A., Marvasti, F.: Low rank and sparse decomposition for image and video applications. IEEE Trans. Circuits Syst. Video Technol. **30**(7), 2046–2056 (2019)
4. Chen, J.: Intelligent recommendation system of dance art video resources based on the wireless network, Secur. Commun. Netw., 2021, (2021).
5. Khan, M.A., Sharif, M., Akram, T., Raza, M., Saba, T., Rehman, A.: Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition. Appl. Soft Comput. **87**, 105986 (2020)
6. Liao, Z., Hu, H., Liu, Y.: Action recognition with multiple relative descriptors of trajectories. Neural Process. Lett. **51**(1), 287–302 (2020)
7. Leng, C., Ding, Q., Wu, C., Chen, A.: Augmented two stream network for robust action recognition adaptive to various action videos. J. Vis. Commun. Image Represent. **81**, 103344 (2021)
8. Jiang, M., Pan, N., Kong, J.: Spatial-temporal saliency action mask attention network for action recognition. J. Visual Commun. Image Represent. **71**, 102846 (2020)
9. Goyal, G., Noceti, N., Odone, F.: Cross-view action recognition with small-scale datasets. Image Vision Comput. **120**, 104403 (2022)
10. Abdelbaky, A., Aly, S.: Two-stream spatiotemporal feature fusion for human action recognition. Vis. Comput. **37**(7), 1821–1835 (2021)
11. Berlin, S.J., John, M.: Spiking neural network based on joint entropy of optical flow features for human action recognition. Vis. Comput. **38**(1), 223–237 (2022)
12. Liu, C., Ying, J., Yang, H., Hu, X., Liu, J.: Improved human action recognition approach based on two-stream convolutional neural network model. Vis. Comput. **37**, 1327–1341 (2021)
13. LO, B. I., HC, M. V. and Schwartz, W. R.: Bubblenet: a disperse recurrent structure to recognize activities, in 2020 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2216–2220 (2020).
14. Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.: VideoLSTM convolves, attends and flows for action recognition. Comput. Vis. Image Underst. **166**, 41–50 (2018)
15. Chen, J., Samuel, R.D.J., Poovendran, P.: LSTM with bio inspired algorithm for action recognition in sports videos. Image Vision Comput. **112**, 104214 (2021)
16. Yao, X., Zhang, J., Chen, R., Zhang, D. and Zeng, Y.: Weakly supervised graph learning for action recognition in untrimmed video, Visual Comput., pp. 1–15, (2022).
17. Fang, Z., Zhang, X., Cao, T., Zheng, Y., Sun, M.: Spatial-temporal slowfast graph convolutional network for skeleton-based action recognition. IET Comput. Vision **16**(3), 205–217 (2022)
18. Sun, N., Leng, L., Liu, J., Han, G.: Multi-stream slowfast graph convolutional networks for skeleton-based action recognition. Image Vision Comput. **109**, 104141–104216 (2021)
19. Han, C., Zhang, L., Tang, Y., Huang, W., Min, F., He, J.: Human activity recognition using wearable sensors by heterogeneous convolutional neural networks. Expert Syst. Appl. **198**, 116764 (2022)
20. Tang, Y., Zhang, L., Min, F., He, J.: Multiscale deep feature learning for human activity recognition using wearable sensors. IEEE Trans. Industr. Electron. **70**(2), 2106–2116 (2022)
21. Huang, W., Zhang, L., Wang, S., Wu, H., Song, A.: Deep ensemble learning for human activity recognition using wearable sensors via filter activation. ACM Transact. Embed. Comput. Syst. **22**(1), 1–23 (2022)
22. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L. Temporal segment networks: towards good practices for deep action recognition, European Conference on Computer Vision, pp. 20–36, (2016).
23. Simonyan, K. and Zisserman, A.: Two-stream convolutional networks for action recognition in videos, Adv. Neural Inform. Process Syst., 27, (2014).
24. Carreira, J. and Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308, (2017).
25. Dong, W., Zhang, Z., Tan, T.: Attention-aware sampling via deep reinforcement learning for action recognition. Proc. AAAI Conf. Artif. Intell. **33**(01), 8247–8254 (2019)
26. Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M.: Learning spatiotemporal features with 3d convolutional networks,

Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497, (2015).
27. 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 **444**, 319–331 (2021)
28. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W.: Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access **6**, 1155–1166 (2017)
29. Ullah, W., Ullah, A., Haq, I.U., Muhammad, K., Sajjad, M., Baik, S.W.: CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimed. Tools Appl. **80**, 16979–16995 (2021)
30. Singh, B., Marks, T. K. Jones, M. Tuzel, O. and Shao, M.: A multi-stream bi-directional recurrent neural network for fine-grained action detection, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1961–1970 (2016).
31. Wang, H. and Schmid C.: Action recognition with improved trajectories, Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558, (2013).
32. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput. Vis. Image Underst. **150**, 109–125 (2016)
33. Li, D., Jahan, H., Huang, X., Feng, Z.: Human action recognition method based on historical point cloud trajectory characteristics. Vis. Comput. **38**(8), 2971–2979 (2022)
34. Klaser, A., Marszałek, M. and Schmid, C.: A spatio-temporal descriptor based on 3D-gradients," in BMVC 2008–19th British Machine Vision Conference. British Machine Vision Association, pp. 1–10 (2008).
35. Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B.: Learning realistic human actions from movies," in 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8 (2008).
36. Shou, Z., Lin, X., Kalantidis, Y., Sevilla-Lara, L., Rohrbach, M., Chang, S.-F. and Yan, Z. DMC-Net: generating discriminative motion cues for fast compressed video action recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1268–1277, (2019).
37. Long, X., Gan, C., Melo, G., Liu, X., Li, Y., Li, F. and Wen, S.: Multimodal keyless attention fusion for video classification, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, (2018).
38. Nagrani, A. Sun, C., Ross, D., Sukthankar, R., Schmid, C. and Zisserman, A.: Speech2action: cross-modal supervision for action recognition," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10317–10326, (2020).
39. Li, C., Zhong, Q., Xie, D. and Pu, S. Collaborative spatiotemporal feature learning for video action recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7872–7881, (2019).
40. Wang, P., Li, W., Wan, J., Ogunbona, P. and Liu, X. Cooperative training of deep aggregation networks for RGB-D action recognition, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32(1), (2018).
41. Min, X., Zhai, G., Zhou, J., Farias, M.C., Bovik, A.C.: Study of subjective and objective quality assessment of audio-visual signals. IEEE Trans. Image Process. **29**, 6054–6068 (2020)
42. Girdhar, R., Carreira, J., Doersch, C. and Zisserman, A.: Video action transformer network, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253 (2019).
43. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. and Hu, H.: Video swin transformer, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3202–3211 (2022).
44. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C. and Schmid, C.: Multiview transformers for video recognition, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3333–3343 (2022).
45. Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., Chiaberge, M.: Action transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recognit. **124**, 108487 (2022)
46. Wang, X., Miao, Z., Zhang, R. and Hao, S.: I3D-LSTM: a new model for human action recognition," in IOP Conference Series: Materials Science and Engineering, vol. 569, no. 3. IOP Publishing, p. 032035 (2019).
47. Ma, J., Li, Z., Cheng, J.C., Ding, Y., Lin, C., Xu, Z.: Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. Sci. Total Environ. **705**, 135771 (2020)
48. Zach, C., Pock, T., and Bischof, H.: A duality based approach for realtime tv-l 1 optical flow," in Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29. Springer, pp. 214–223 (2007).
49. Zhai, G., Min, X.: Perceptual image quality assessment: a survey. Science China Inf. Sci. **63**, 1–52 (2020)
50. Sun, L., Jia, K., Yeung, D.-Y. and Shi, B. E. Human action recognition using factorized spatiotemporal convolutional networks," Proceedings of the IEEE International Conference on Computer Vision, pp. 4597–4605, (2015).
51. Feichtenhofer, C., Pinz, A. and Zisserman, A.: Convolutional two-stream network fusion for video action recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941, (2016).
52. Feichtenhofer, C., Pinz, A. and Wildes, R. P.: Spatiotemporal multiplier networks for video action recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4768–4777, (2017).
53. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. Adv. Neural Inform. Process. Syst. **35**, 10078–10093 (2022)
54. Huang, G. and Bors, A. G.: Busy-quiet video disentangling for video classification, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1341–1350 (2022).
55. Zhang, Y., Li, X., Liu, C., Shuai, B., Zhu, Y., Brattoli, B., Chen, H., Marsic, I. and Tighe, J.: "VidTr: video transformer without convolutions," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 13577–13587 (2021).



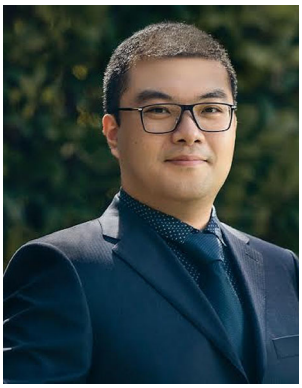
Chuanjiang Leng received B.E. degree from Beijing University of Chemical Technology in 2017 and received M.S. degree from Northeastern University in 2019. He is currently pursuing Ph.D. degree in the Faculty of Robot Science and Engineering, Northeastern University. His research interest is action recognition, deep learning and image segmentation.

Qichuan Ding received his B.E. degree in information and computing science from Shandong University, Jinan, China, in 2007 and Ph.D. degree from University of Chinese Academy of Sciences, China, in 2015. He is an Associate Professor with Northeastern University, Shenyang, China. His main research interests include sEMG signal processing, wearable robots and pattern recognition.

Chengdong Wu received B.E. degree from Shenyang Jianzhu University in 1983 and received B.S. degree from Tsinghua University in 1988. He received Ph.D. degree from Northeastern University in 1994. He is currently a full professor with the Faculty of Robot Science and Engineering, Northeastern University. His interest is robot vision, computer vision and medical image processing.

Ange Chen received B.E. degree from Beijing University of Chemical Technology in 2017 and received master's degree from Northeastern University in 2019. He is currently a Ph.D. candidate of the Faculty of Robot Science and Engineering, Northeastern University. His interest is human pose estimation and medical image processing.

Huan Wang received B.E. degree and M.S. degree from Northeastern University in 2018 and 2020. He is currently pursuing Ph.D. degree in the Faculty of Robot Science and Engineering, Northeastern University. His research interest is image description, multi-scale methods and sparse representations of images, spatial and frequency image filtering, and image super-resolution.



Hao Wu received Ph.D. degree from Engineering Faculty, University of Sydney. Currently, he is working on the University of Sydney. His interest is robot vision, computer vision and image matching.