Human action recognition method based on Motion Excitation and Temporal Aggregation module

Qing Ye a,∗, Zexian Tan b, a, Yongmei Zhang a

a School of Information Science and Technology, North China University of Technology, Beijing 100144, China
b System Operation Center, Bank of Beijing, Beijing 100033, China

A R T I C L E   I N F O
Keywords:
Human action recognition
Motion Excitation and Temporal Aggregation module
Temporal relational sampling
Spatiotemporal two stream network
Temporal modeling
Cross modality pre-training

A B S T R A C T
Aiming at the problem of low modeling efficiency and feature loss of temporal modeling in human action recognition, we propose a human action recognition method based on Motion Excitation and Temporal Aggregation module (META). The method can capture multi-state and multi-scale temporal information to achieve effective motion excitation. Firstly, temporal relational sampling is performed on video frames; Secondly, META is proposed to capture multi-state and multi-scale temporal information. META is composed of Multi-scale Motion Excitation module (MME) and Squeeze and Excitation Temporal Aggregation module (SETA). MME captures the feature level temporal difference by transforming the features into the temporal channel, which directly establishes the relationship between features and temporal channel, and solves the problem of low modeling efficiency. SETA transforms the local convolution into a set of sub-convolutions. Multiple sub-convolutions form hierarchies to extract features together and share the results of the upper convolutional layer, which increases the final temporal receptive field and solves the problem of feature loss. Moreover, the optical flow features are extracted through Cross modality pre-training to improve the utilization of temporal information. Finally, the result of human action recognition is carried out by combining spatiotemporal two stream features. Experimental results show that the accuracy of this method in UCF101 and HMDB-51 is 96.0% and 71.2% respectively, which is higher than other studies in the same period.

1. Introduction

Human action recognition [1] is one of the research hotspots in computer vision, and it is widely used in many fields such as intelligent monitoring, behavior analysis and human-computer interaction. Early action recognition mainly uses artificial design to mark interest points or artificial design features to represent actions. However, due to the problem of high complexity of human actions and small differences in similar actions, the accuracy of action recognition needs to be improved [2, 3]. In recent years, with the development of deep learning in the field of image recognition, deep learning has been widely used in human action recognition.

With the development of deep learning, people realize the importance of temporal modeling for action recognition. Therefore, two stream network [4] using optical flow features also appears. The two stream network extracts RGB features and motion features in stacked optical flow respectively. Finally, the human action recognition is determined by fusing RGB features and optical flow features. However, this method only operates a single frame in space or adjacent frames in a short time. It ignores the correlation of each link of human action in a long time. After that, a variety of two stream networks such as TSN [5], TRN [6] and TeaNet [7] appear. Some of them start from long-term temporal modeling, and some start from enhancing optical flow feature expression. The above methods fully reflect the importance of temporal modeling for obtaining action features.

This paper believes that short-term temporal modeling and long-term temporal modeling are both important for action recognition, because they are complementary in temporal modeling. Therefore, this paper proposes a human action recognition method based on Motion Excitation and Temporal Aggregation (META) module. The focus of this method is to solve the problems of low modeling efficiency and feature loss in temporal modeling. This method constructs a spatiotemporal two stream network based on TRN framework, and uses temporal relational sampling to strengthen the relationship between frames in tem-
poral modeling. Moreover, we propose META for temporal modeling. META is composed of Multi-scale Motion Excitation module (MME) and Squeeze and Excitation Temporal Aggregation module (SETA). MME makes it possible to extract RGB features and feature level temporal difference at the same time. MME integrates short-term temporal modeling into spatiotemporal feature learning to solve the problem of low modeling efficiency. SETA forms a hierarchical structure through multiple sub convolutions to extract effective features of different temporal scales. Therefore, the final temporal receptive field is increased, and the problem of feature loss is solved. To further improve the modeling efficiency, we use cross modality pre-trained network to extract the features of the optical flow image, which improves the utilization of the optical flow features.

In summary, human action recognition method based on Motion Excitation and Temporal Aggregation (META) module is proposed in this paper, which is used to solve the problems of low efficiency and feature loss in temporal modeling. Among them, this paper proposes MME module for short-term temporal modeling and SETA module for long-term temporal modeling. Above modules capture the feature level time difference from the feature level. It is more efficient than traditional convolution to obtain features. In this paper, human action recognition based on Motion Excitation and Temporal Aggregation module is studied. The main contents include the following five chapters.

The Section 1 summarizes the development of human action recognition, and leads to the proposed human action recognition algorithm based on Motion Excitation and Temporal Aggregation module. The Section 2 introduces the development of human action recognition in this research direction from the perspectives of long-time temporal modeling and short-time temporal modeling. The Section 3 elaborates the algorithm principle of this method in detail. The Section 4 introduces the experimental results of this paper. The Section 5 generalizes the methods and problems of this paper.

2. Related work

Short term temporal modeling is a hot topic of action recognition in recent years. One of the main methods includes two stream CNN [4] and its evolution [5, 6], which respectively model in time and space, extract RGB features and optical flow features, and fuse them in the later stage. Some choose to use RGB difference instead of optical flow as motion representation [5]. But this is just another video mode to process RGB differently and train a separate network to integrate with RGB. It is essentially a two-stream network. Others main methods include 3D-CNN [8] and others try to reduce the computational cost of 3D convolution by decomposing 3D convolution into 2D convolution and 1D time convolution, such as R (2 + 1) D [9], S3D [10], P3D [11] and CT Net [12]. According to this research route, some researches focus on designing better temporal modules for efficient action recognition, which are inserted into 2D CNN. For example, TSM [13], TIN [14], TEINet [15], TANet [16] and TEA [17]. Among them, TEA network proposes Motion Excitation module to obtain these short-term moving characteristics described by the feature level temporal difference, so as to realize the short-term temporal modeling. By referring to the idea of TEA network, Action-Net [18] and TDN [19] network are proposed on CVPR in 2021. Action-Net proposed its own Motion Excitation module by subtracting the spatial features of the three adjacent moments. Referring to the idea of Motion Excitation, TDN network puts forward the Short-term temporal difference module by fusing the characteristic level time difference in the segment. In addition, some methods try to design a multi branch architecture based on two stream network. The branches use specially designed temporal modules or input two RGB frames sampled in different FPS to capture RGB features and motion features, such as STM [20] and Slowfast [21].

In the long-term temporal modeling, RNN [22] or Long-term temporal convolutions [23] are used to stack more frames to simulate long-term modeling. With the emergence of TSN network [5], its sparse sampling and aggregation strategy has been proved to be an effective long-term modeling framework, and has performed well on multiple datasets. However, the disadvantage of TSN is that the network only performs temporal fusion in the final stage, which leads to its failure to capture more precise time features. StNet [24] proposed a local and global module to model time information hierarchically. TEA [17] network constructs Multiple Temporal Aggregation module through hierarchical stacking sub convolution, which is equivalent to short-term temporal modeling. Based on TEINet [15], TDN [19] network proposes long-term temporal difference module to simulate long-term temporal modeling.

Different from the above methods, the META we proposed deals with the problems of short-term temporal modeling and long-term temporal modeling through its internal two sub modules. MME solves the problem of low modeling efficiency of short-term temporal modeling. SETA solves the problem of feature loss of long-term temporal modeling. Moreover, META integrates short-term temporal modeling and long-term temporal modeling in one module, which greatly reduces the complexity of the network.

3. Method

The human action recognition network based on META proposed in this paper is shown in Fig. 1. Firstly, we sparsely sample video to get RGB video frame, and extract optical flow of video. Then, we use temporal relational sampling to form RGB video frame segment groups and optical flow image segment groups. Secondly, RGB video frame segment groups and optical flow image segment groups respectively enter into the spatial feature extraction network based on META and the temporal feature extraction network. After the RGB image is processed by spatial feature network, the action features described by RGB feature and feature-level temporal difference are obtained. The optical flow image is processed by temporal feature extraction network to obtain optical flow features. The features of each video segment in time and space are fused to get the spatial and temporal features for the whole video. According to the weights, the Spatiotemporal two stream fusion is carried out, and classification is carried out after fusion.

3.1. Temporal relational sampling based on TRN

Temporal Relation Network (TRN) in video frames sampling can strengthen the relationship between frames in temporal modeling. Fig. 2 is the diagram of TRN algorithm. Firstly, we evenly sample N frames from RGB video frames or optical flow frames. Considering that the experiment is verified on UCF101 [25] dataset and HMDB-51 [26] dataset, combined with the total number of frames, after many experiments, we can obtain better action recognition results when the number of sampling frames is 8. The frame image is recorded as \( v = \{ F_1, F_2, F_3, ..., F_8 \} \). We randomly select different numbers of frame segments to compose 2, 3,..., N frame segment groups. This sampling method will not cause action recognition confusion due to insufficient use of correlation information. Similarly, there will be no increase in the amount of data. It is conducive to the prediction of action recognition, and can strengthen the relationship between frames in temporal modeling.

3.2. Action recognition network based on META

In this paper, META is proposed, which can be used in most convolutional neural networks. META is composed of Multi-scale Motion Excitation module (MME) and Squeeze and Excitation Temporal Aggregation module (SETA). MME is integrated into the ResNet101’s bottleneck layer. SETA is used to take the place of 3x3 convolution layers in original residual path. Through the superposition of META modules, the action recognition network is constructed. The action recognition
network based on META module is shown in Fig. 3. RGB video gets T frame images through sparse sampling. The image is sent to ResNet which is composed of META as the backbone network. After convolution, MME and SETA, the spatial action features including RGB features and feature-level temporal difference are obtained. Then the features are classified to get the recognition results.

In terms of short-term temporal modeling, most methods use optical flow or 3D convolution ((2 + 1)D convolution). 3D convolution ((2+1)D convolution) can effectively extract the temporal information between adjacent frames. However, the complex network parameters lead to low modeling efficiency.

In long-term temporal modeling, traditional methods mainly include two categories: 1) 2D convolution is performed on each frame image to extract features, and then multi frame feature images are pooled to fuse the results, such as TSN [5]. This method is simple to implement, but this simple global pooling operation will lead to the loss of features. 2) The local 3D/(2 + 1)D convolution operation is used to process the local time window. By repeatedly superimposing local convolution in the network, the long-term time relationship is indirectly simulated. However, the complex network parameters lead to low modeling efficiency.

In view of the above problems, the META proposed in this paper can solve the above two kinds of problems. MME is used to solve the
problem of low modeling efficiency. SETA is used to solve the problem of feature loss. The action feature extracted by MME is different from that of optical flow. MME deals with the change of feature level between adjacent frames, while optical flow deals with the change of pixel level between two adjacent frames. Moreover, MME can establish the relationship between RGB features and temporal information without training solely. The idea of SETA comes from Res2Net [27]. Although (2 + 1)D convolution is also selected in SETA, it uses a group of sub convolutions instead of one-dimensional temporal convolution. One hierarchical structure is formed by sub convolutions, and there are connections between adjacent sub convolutions. The sub convolutions can exchange information with adjacent frames for many times when features are sent to the module. The equivalent time receptive field is multiplied to simulate long-term temporal modeling. Moreover, the shared features between the convolutions are squeezed to enhance the transfer of useful features and suppress the transfer of invalid features.

3.2.1. Multi-scale Motion Excitation module (MME)

The motion feature is obtained by comparing the displacement of two adjacent frames. In most previous methods, optical flow is usually used to describe motion features. Most methods separate the spatiotemporal feature learning, and only fuse in the last few layers. In contrast, in the MME proposed in this paper, temporal modeling is expanded from pixel level to feature level, and modeling is carried out on spatial channels, so as to combine spatiotemporal modeling into one framework. This improves the effectiveness of short-term temporal modeling.

The network structure of MME is shown in Fig. 4. When a feature X is given, with a size of [N, T, C, H, W], a 1x1 2D convolution is used to reduce the dimension of the feature and reduce the number of unnecessary channels.

\[
X' = \text{conv}_{1x1,2D}(X), \quad X' \in \mathbb{R}^{N \times T \times C / 1 \times H \times W}
\]  

In equation (1), * represents convolution operation, X' is the feature after channels reduction, and r = 16 represents reduction ratio.

Feature level motion description can be approximately expressed as the difference between two adjacent frames at time \(t - 1\) and t in the T dimension. Therefore, we first transform the channel of the feature, and then use the transformed feature to calculate the motion feature, instead of directly subtracting the original feature. Direct subtraction may lead to the replacement of the eigenvalue. The equation (2) is as follows:

\[
M(t - 1) = \text{conv}_{3x3,2D}(X(t)) - X(t) - X(t - 1),
\]

\[
M(t - 1) \in \mathbb{R}^{N \times T \times C / 1 \times H \times W}, \quad 2 \leq t \leq T - 1
\]  

The feature level motion description M(t − 1) at time \((t - 1)\) can be represented by the feature level time difference between adjacent frames at time \((t - 1)\) and t in the T dimension. \(\text{conv}_{3x3,2D}\) is a 3x3 2D convolution used to transform each channel.

In order to enhance the feature level motion description, we extend the adjacent frames of T dimension from \((t - 1)\) and t to \((t - 1)\), t, and \((t + 1)\) in the same way, we can get the difference at time \(t\) and \((t + 1)\) in the T dimension. We do two channel transformations for \(X'(t + 1)\), and then subtract from \(X'(t)\) after one channel transformation to prevent the loss of eigenvalues. The equation (3) is as follows:

\[
M(t) = \text{conv}_{3x3,2D}(X(t + 1)) - \text{conv}_{3x3,2D}(X(t)),
\]

\[
M(t) \in \mathbb{R}^{N \times T \times C / 1 \times H \times W}, \quad 1 \leq t \leq T - 1
\]  

The feature level motion description M(t) at time t can be represented by the feature level time difference between adjacent frames at time t and \((t + 1)\) in the T dimension. \(\text{conv}_{3x3,2D}\) is a 3x3 2D convolution used to transform each channel.

By superimposing the motion matrices \(M^{-1}\) and \(M\), we can get the feature level motion description M between the adjacent three frames. Then we use the global average pool layer to summarize the spatial information, in which the detailed spatial information is not important because of the goal of stimulating motion sensitive channels.

After pooling, we use a 1x1 2D convolution to reduce the channel number of M to \([N, T, C, 1, 1]\). The final motion feature weight A can be obtained by sigmoid operation to activate M'.

Finally, we need to use the motion feature weight A to activate the motion features in the input features. The motion feature weight A is multiplied by the channel of the input feature to enhance the motion feature. Considering that the static features in input features will be suppressed, however, the static features also play a certain role in action recognition. We use the idea of residual to retain the static features.

\[
X^o = A \odot X + X, \quad X^o \in \mathbb{R}^{N \times T \times C \times H \times W}
\]  

In equation (4), \(X^o\) represents MME module’s output, which retains both static features and enhanced motion features. \(\odot\) is dark channel multiplication.

3.2.2. Squeeze and Excitation Temporal Aggregation module (SETA)

Local temporal convolution is usually used to deal with adjacent frames to obtain local temporal features in previous action recognition algorithms. To obtain long-term temporal features, we can only stack a huge amount of local temporal convolution structures in the depth network. The effect of this method is not obvious, because the optimization message delivered from the long interval frame has been greatly weakened and the effectiveness has been greatly weakened. In order to solve the problem of feature loss, Squeeze and Excitation Temporal Aggregation module (SETA) is proposed, which can be effectively applied to
long-term temporal modeling. SETA refers to the method of Res2Net sub convolution, which designs the sub convolution into a hierarchical structure without introducing other parameters, so that the next sub convolution can use the features of the previous sub convolution. This kind of operation is equivalent to expanding the receptive field of time dimension. Moreover, we squeeze the shared features between the convolutions to enhance the transfer of useful features and suppress the transfer of invalid features.

As shown in Fig. 5, given an input feature, along the channel dimension, we divide the feature into four segments, and \([N,T,C/4,H,W]\) represents each segment’s shape. The local volume is integrated into several sub convolutions, and the last three segments are processed by time sub convolution and space sub convolution in turn.

\[
\begin{align*}
X'_1 &= X_1, \\
X'_2 &= \text{conv}_{\text{spatial}} \ast (\text{conv}_{\text{temporal}} \ast X_2). \\
X'_3 &= \text{conv}_{\text{spatial}} \ast [\text{conv}_{\text{temporal}} \ast (X_3 + F_{\text{se}}(X'_2))]. \\
X'_4 &= \text{conv}_{\text{spatial}} \ast [\text{conv}_{\text{temporal}} \ast (X_4 + F_{\text{se}}(X'_3))].
\end{align*}
\]  
(5)

In equation (5), \(X'_i \ (i = 1, 2, 3, 4)\) represents the output of the \(i\) segment, \(\text{conv}_{\text{spatial}}\) represents the spatial convolution, \(\text{conv}_{\text{temporal}}\) represents the temporal convolution, and \(F_{\text{se}}\) represents the squeezing excitation method. In the last three segments, we add residual connection to transform the parallel architecture into a hierarchical cascade structure and expand the receptive field of time dimension. Before the residual connection is transmitted to the next segment, the output of the previous segment is squeezed to enhance the transmission of useful features and suppress useless features. The specific operation refers to the idea of SENet [28].

Firstly, the output of each segment is globally averaged pooled with extrusion operation, and its spatial size feature is compressed to make its shape change to \([N,T,C/4,1,1]\), and the global features are concentrated in \(N, T\) and \(C\) dimensions. Then, the adaptive weight learning is carried out through two fully connected layers, and a new feature map with size of \([N,T,C/4,1,1]\) is obtained after learning. The process equation (6) is as follows:

\[
S = F_{\text{se}}(U, W) = \sigma \left( W_2 \delta \left(W_1 U \right) \right)
\]  
(6)

In equation (6), \(S\) is the feature image containing importance information, \(F_{\text{se}}\) is the weighted excitation function, \(U\) is the feature image before passing through the two fully connected layers, \(W_1\) and \(W_2\) are the weight (dimension reduction parameter) of the first and second fully connected layers respectively, \(W\) is the general name of \(W_1\) and \(W_2\), \(\sigma\) is the sigmoid activation function, and \(\delta\) is the ReLU activation function.

The weighted feature graph is given different weights on each channel of different dimensions to express the importance of feature information, enhance the transmission of useful features and suppress useless features.

Different segment has different receptive field in this module. By connecting the information of the previous segment in series, the last segment’s equivalent receptive field is expanded three times. At the end of the module, multiple outputs are combined through a cascading strategy.

\[
X^0 = [X^0_1, X^0_2, X^0_3, X^0_4]
\]  
(7)

In equation (7), with the increase of receptive field, \(X^0\) contains space-time representation of different time ranges, \(X^0 \in \mathbb{R}^{N \times T \times C \times H \times W}\), which is better than the previous local time convolution method.

With the increase of temporal receptive field, SETA can more easily establish the information between long-term actions. Compared with multiple local time convolution stacking method, SETA has less computational complexity, can reduce the redundant features generated by long-term modeling, and enhance the useful action features.

In Fig. 6, the left is the original image, the middle is the Grad-CAM visualization image of ResNet101 network, and the right is the Grad-CAM visualization image of ResNet101 with SETA added. It can be seen...
clearly that before adding SETA, the network pays attention to the background and characters (RGB color covers a larger area, and the color of human’s actions is lighter), and does not pay attention to the action itself, which will produce a lot of redundant features. After adding SETA, the focus of network attention is shifted to the action itself (RGB color coverage becomes smaller, the background area is reduced, and the color of the human’s action becomes darker), which improves the transmission of useful features, suppresses useless features (background, etc.), and enhances the network’s capture of motion features.

3.3. Acquisition of optical flow image

Some principles of optical flow calculation adopted in this paper are as follows.

Assuming \((x, y)\) is a pixel in the video frame image, the brightness of the pixel at time \(t + \Delta t\) is \(E(x + \Delta x, y + \Delta y, t + \Delta t)\). The changes of pixels on the \(x\) and \(y\) axes are \(\Delta x\) and \(\Delta y\) respectively, and the time interval of pixel changes is \(\Delta t\). In equation (8) and equation (9), the horizontal displacement component of the optical flow of the pixel is represented by \(u\), and the vertical displacement component is represented by \(v\):

\[
u = \frac{dx}{dt} \tag{8}
\]

\[
v = \frac{dy}{dt} \tag{9}
\]

When the time interval \(\Delta t\) is approximately 0, the brightness of the pixel at time \(t\) is approximately unchanged, and the following equation (10) is obtained:

\[
E(x, y, t) = E(x + \Delta x, y + \Delta y, t + \Delta t) \tag{10}
\]

When the brightness of the pixel changes, the following equation (11) can be obtained:

\[
E(x + \Delta x, y + \Delta y, t + \Delta t) = E(x, y, t) + \frac{\partial E}{\partial x} \Delta x + \frac{\partial E}{\partial y} \Delta y + \frac{\partial E}{\partial t} \Delta t + \epsilon \tag{11}
\]

When the time interval is approximately 0, the second-order infinitesimal term in the equation can be ignored, and the velocity change and direction change of each point at a certain moment can be obtained, which is expressed by \(\nabla E_{gw}\), as shown in equation (12):

\[
-\frac{\partial E}{\partial t} = \frac{\partial E}{\partial x} \frac{dx}{dt} + \frac{\partial E}{\partial y} \frac{dy}{dt} \nabla E_{gw} \tag{12}
\]

By adopting the above method, this paper obtains the optical flow image corresponding to the RGB image, as shown in Fig. 7, where Fig. 7 (a) represents the original RGB image, the action is figure skating and high jump, and Fig. 7 (b) represents the optical flow image corresponding to each figure in Fig. 7 (a). It can be seen that optical flow image mainly presents motion change information, and RGB image contains rich spatial information.

3.4. Spatiotemporal fusion and classification

After the RGB video frame is extracted by the spatial feature extraction network based on META, the composite features of motion feature and RGB feature are generated. The temporal feature extraction network of optical flow frame is trained by cross modality pre-training to generate pixel level temporal features. The features generated by different video frames in space are fused together according to the mean value to generate spatial features, and the temporal features are obtained in the same way. The fusion ratio was tested on UCF101 and HMDB-51 datasets. In order to obtain the best spatiotemporal fusion ratio of different datasets, we have carried out the linear change experiment of the fusion ratio of spatial features and temporal features, and the experimental accuracy is shown in Fig. 8.
As can be seen from Fig. 8, before the time and space fusion ratio is 1:1, the fusion accuracy increases with the increase of the optical flow fusion ratio. After the time and space fusion ratio is 2:1, the fusion accuracy begins to decrease with the increase of the optical flow fusion ratio, and the highest fusion accuracy is about 1.5. Therefore, the temporal and spatial fusion ratio of UCF101 dataset is 1.5.

It can be seen from Fig. 9 that before the time-space fusion ratio is 1:1:3, the fusion accuracy will also increase with the increase of the optical flow fusion ratio. When the ratio of time to space fusion is 2:1, with the increase of the optical flow fusion ratio, the fusion accuracy begins to decrease, and the maximum fusion accuracy rate is about 2.0. Therefore, the time-space fusion ratio of HMDB-S1 dataset is 2.0. Thus, HMDB-51 dataset is more sequential than UCF101 dataset.

This paper studies the multi-classification problem in action recognition, so Softmax classifier is selected. In addition, Softmax classifier combined with cross entropy loss function can effectively avoid gradient dispersion. The Softmax classifier classifies the features, and then each action will generate a probability label corresponding to it. The logistic regression model can be extended to obtain the Softmax regression model, which is very common in multi-classification problems. For training set \( \{ (x^{(1)}, y^{(1)}) \ldots (x^{(m)}, y^{(m)}) \} \), there are \( y^{(i)} \in \{1, 2, 3, \ldots, K\} \). The training set can be divided into \( K \) classes. The probabilities \( P(y = j | x) \), \( j = (1, 2, \ldots, k) \) of each class are trained by each input \( X \) in the training set, and they correspond to each other by one. If the \( K \) estimated probability values of function \( h_{\theta}(x) \) are to use a \( K \)-dimensional vector group whose sum of vector elements is 1. As shown in equation (13):

\[
h_{\theta}(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1 | x^{(i)}, \theta) \\
p(y^{(i)} = 2 | x^{(i)}, \theta) \\
\vdots \\
p(y^{(i)} = k | x^{(i)}, \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\
\vdots \\
e^{\theta_K^T x^{(i)}} \end{bmatrix}
\]

(13)

Where \( \theta_1, \theta_2, \ldots, \theta_K \in \mathbb{R}^{n+1} \) is the parameter of the model.

In this paper, the selection of softmax as the classifier is determined by the loss function selected in this paper, and the cross entropy loss function is selected in this paper. The combination of softmax classifier and cross entropy loss function can effectively avoid the occurrence of gradient dispersion. At the same time, softmax is very suitable for multi-classification of images. After softmax operation, we can get the normalized value of each type of data. According to the normalized value, softmax can give the index with the maximum value.

As shown in Fig. 10, the verification method of this paper is cross verification method. The specific steps are as follows: firstly, the data set is divided into \( K \) disjoint data subsets. Then, randomly select a data subset in the \( K \) data subsets as the verification set, and the remaining \( K-1 \) data subsets as the training set. Train the model on the selected training set, calculate the current verification probability on the verification set, and repeat the above steps \( K \) times. Finally, the obtained verification probability is averaged to obtain the verification probability of this method on the data set.

This paper proposes a human action recognition method based on META, which uses a variety of temporal modeling methods to obtain multi-state and multi-scale temporal information. Firstly, temporal relational sampling is used to strengthen the temporal relationship between video frames. We propose MME module to solve the problem of low efficiency in short-term temporal modeling. SETA module is proposed to solve the problem of feature loss in long-term temporal modeling. Finally, a cross-modality pre-trained model is used to improve the utilization of optical flow features.

4. Experimental results

4.1. Experimental datasets

UCF101 and HMDB-51 contain not only many basic actions in life, but also many sports. They are two challenging and authoritative public datasets, so our experiments are carried out on these two datasets.

Fig. 11 shows an action example of the part in the HMDB-51 dataset. HMDB-51 contains 51 types of actions, a total of 6849 videos, each action contains at least 51 videos, with a resolution of 320 * 240. The data set comes from YouTube, Google Video, etc., with a total size of 2G. The main movements include: smile, laugh, chew, talk, fence, hug, kick, kiss and shake hands.

Fig. 12 shows an action example of the part in the UCF101 dataset. UCF101 is a real action video action recognition dataset, collected from YouTube. 13320 videos from 101 action categories are provided. UCF101 provides the greatest diversity in motion, and has great changes in camera motion, object appearance and posture, object scale, viewpoint, cluttered background, lighting conditions, etc. The videos in 101 action categories are divided into 25 groups, each group can contain 4-7 videos of an action. It includes five kinds of actions: human and object interaction, simple body action, human and human interaction, playing musical instruments and sports.

The above two public data sets are selected because they are still challenging in the field of human action recognition and are widely used public data sets. During the experiment, this paper uses the model pre trained by ImageNet. Compared with the kinetic pre training model, it is easier to obtain and more widely used.

4.2. Ablation experiment

In order to verify the improvement effect of MME and SETA in META on human action recognition, we combined MME and SETA with ResNet101 respectively, and conducted ablation experiments on HMDB-51 dataset without affecting the ablation results.

As shown in Fig. 13 (a) means that a 1D time channel convolution is inserted after the first 2D space channel convolution of the standard ResNet block. Fig. 13 (b) means inserting SETA in ResNet block instead of 1D time channel convolution and 3D space channel convolution of 3 * 3, that is, using the layer-level sub convolution group structure in the SETA to extract the time and space channel features simultaneously.
The ablation experiment uses human action recognition methods based on META proposed in this paper to test HMDB51 dataset on spatial RGB. The experimental results are shown in Table 1 and the four methods in Table 1 correspond to a, b, c and d in Fig. 13 respectively.

It can be seen from Table 1 that adding SETA makes the network 0.7% higher than (2 + 1)D ResNet. Compared with (2 + 1)D ResNet, SETA uses hierarchical structure to expand the equivalent receptive field of time dimension in each block, so as to construct a network with long-term temporal modeling ability, which improves the accuracy of human action recognition from the perspective of long-term temporal modeling. From the point of view of short-term temporal modeling, adding MME is 1% higher than (2 + 1)D ResNet. Compared with (2 + 1)D ResNet, MME can capture the feature time difference of action to describe the movement, and make the network pay attention to the time information reflecting the actual action. META integrates the advantages of SETA and MME in temporal modeling, and pays more attention to the information of human action in temporal dimension, which is 1.4% higher than (2 + 1)D ResNet.

4.3. Experimental results

In this paper, a human action recognition method based on META is implemented under the Pytorch framework, and the combination of META and ResNet101 uses the combined network to extract spatial features. Considering the complexity of optical flow feature parameters, this paper uses cross-modal pre-trained BN perception network for temporal modeling of optical flow. The spatiotemporal features of HMDB-51 dataset are fused according to the ratio of time: space equal to 2:1, and the spatiotemporal features of UCF101 dataset are fused according to the ratio of time: space equal to 1.5:1. The experimental environment is as follows: the system is Win10, the CPU is i9-9900KF, the GPU is RTX2080Ti, and the memory is 32 GB. Under the above hardware conditions, through a large number of experiments, the proposed human action recognition method based on temporal modeling has a high accuracy in human action recognition.
modality pre-training for optical flow, but directly uses ImageNet’s pre-training model. In the reproduction of TEA, we did not use the pre-training model of Kinetics as in the original paper. The above method makes the pre-training conditions the same as other methods, so that the results can be compared.

It can be seen from Table 3 that our method is 14.0% higher than the best traditional method IDT; It is 7.6% higher than the two-stream three-dimensional convolutional network T3D; Under the same backbone network and pre-training conditions, our method is 11.8% higher than the common two-stream network after fusion. Under the same pre-training model, our method has higher accuracy than TEA. Compared with TEA pre-trained with two datasets, our method is slightly backward, but saves the pre-training time. It can be seen that the META proposed in this paper, combined with ResNet101, improves the utilization of temporal features and improves the accuracy of action recognition.

In order to further evaluate the performance of this model, this paper supplementary verifies the human action recognition method based on motion excitation and time aggregation module on HMDB51 data set. We calculated the sensitivity, precision and F1-score of the model, which are shown in Table 4. However, there are few studies on human action recognition with the above measurement parameters, so it is not compared with other studies.

### 4.3.2. Experimental results in UCF101

During the training of UCF101, we use ImageNet pre-trained weights for spatial feature extraction network and perform cross modality pre-training for temporal feature network. The experimental results are shown in Table 5. The accuracy of TRN method in UCF101 dataset is the result of this experiment, which is based on the reference method under the above experimental conditions. In the reproduction, in order to compare the results with this method, ResNet101 is used in the spatial feature extraction network, and the pre-training is carried out on ImageNet, and the spatiotemporal fusion ratio is 2:1. Under the same conditions, compared with TRN, the accuracy of RGB channel is increased by 1.4% by adding META module. After cross modal pre-training, the accuracy of optical flow channel is improved by 15.7%, and the fusion is 10.6% higher than TRN. It can be seen that the proposed META module can extract the information of action in time dimension to the greatest extent compared with ResNet101 network, the accuracy of action recognition is improved.

Fig. 14 shows the change of value accuracy and loss when the human action recognition network based on temporal modeling is trained on HMDB51. Fig. 14 (a) corresponds to the META-RGB method in Table 2, and Fig. 14 (b) corresponds to the META-Flow method in Table 2.

Our method is also compared with the study under the same experimental conditions. The experimental results are shown in Table 3. The proposed META improves the accuracy of ResNet101 significantly, and the human action recognition network based on temporal modeling is better than other studies.

The accuracy rates of TSN and TEA (pre-train: ImageNet) in Table 3 on HMDB51 dataset are the results of this experiment under the above experimental conditions. In the reproduction, TSN does not use cross modality pre-training for optical flow, but directly uses ImageNet’s pre-training model. In the reproduction of TEA, we did not use the pre-training model of Kinetics as in the original paper. The above method makes the pre-training conditions the same as other methods, so that the results can be compared.

It can be seen from Table 3 that our method is 14.0% higher than the best traditional method IDT; It is 7.6% higher than the two-stream three-dimensional convolutional network T3D; Under the same backbone network and pre-training conditions, our method is 11.8% higher than the common two-stream network after fusion. Under the same pre-training model, our method has higher accuracy than TEA. Compared with TEA pre-trained with two datasets, our method is slightly backward, but saves the pre-training time. It can be seen that the META proposed in this paper, combined with ResNet101, improves the utilization of temporal features and improves the accuracy of action recognition.

In order to further evaluate the performance of this model, this paper supplementary verifies the human action recognition method based on motion excitation and time aggregation module on HMDB51 data set. We calculated the sensitivity, precision and F1-score of the model, which are shown in Table 4. However, there are few studies on human action recognition with the above measurement parameters, so it is not compared with other studies.

### 4.3.1. Experimental results in HMDB51

When training in HMDB51, we use ImageNet pre-trained weights for spatial feature extraction network to pre-train the temporal feature network. The experimental results are shown in Table 2. The accuracy of TRN method on HMDB51 dataset is the result of this experiment, which is based on the reference method under the above experimental conditions. In the reproduction, in order to compare the results with this method, ResNet101 is used in the spatial feature extraction network, and the pre-training is carried out on ImageNet, and the spatiotemporal fusion ratio is 2:1. Under the same conditions, compared with TRN, the accuracy of RGB channel is increased by 1.4% by adding META module. After cross modal pre-training, the accuracy of optical flow channel is improved by 15.7%, and the fusion is 10.6% higher than TRN. It can be seen that the proposed META module can extract the information of action in time dimension to the greatest extent compared with ResNet101 network, the accuracy of action recognition is improved.

Fig. 14 shows the change of value accuracy and loss when the human action recognition network based on temporal modeling is trained on HMDB51. Fig. 14 (a) corresponds to the META-RGB method in Table 2, and Fig. 14 (b) corresponds to the META-Flow method in Table 2.

Our method is also compared with the study under the same experimental conditions. The experimental results are shown in Table 3. The proposed META improves the accuracy of ResNet101 significantly, and the human action recognition network based on temporal modeling is better than other studies.

The accuracy rates of TSN and TEA (pre-train: ImageNet) in Table 3 on HMDB51 dataset are the results of this experiment under the above experimental conditions. In the reproduction, TSN does not use cross modality pre-training for optical flow, but directly uses ImageNet’s pre-training model. In the reproduction of TEA, we did not use the pre-training model of Kinetics as in the original paper. The above method makes the pre-training conditions the same as other methods, so that the results can be compared.

It can be seen from Table 3 that our method is 14.0% higher than the best traditional method IDT; It is 7.6% higher than the two-stream three-dimensional convolutional network T3D; Under the same backbone network and pre-training conditions, our method is 11.8% higher than the common two-stream network after fusion. Under the same pre-training model, our method has higher accuracy than TEA. Compared with TEA pre-trained with two datasets, our method is slightly backward, but saves the pre-training time. It can be seen that the META proposed in this paper, combined with ResNet101, improves the utilization of temporal features and improves the accuracy of action recognition.

In order to further evaluate the performance of this model, this paper supplementary verifies the human action recognition method based on motion excitation and time aggregation module on HMDB51 data set. We calculated the sensitivity, precision and F1-score of the model, which are shown in Table 4. However, there are few studies on human action recognition with the above measurement parameters, so it is not compared with other studies.

### 4.3.1. Experimental results in HMDB51

When training in HMDB51, we use ImageNet pre-trained weights for spatial feature extraction network to pre-train the temporal feature network. The experimental results are shown in Table 2. The accuracy of TRN method on HMDB51 dataset is the result of this experiment, which is based on the reference method under the above experimental conditions. In the reproduction, in order to compare the results with this method, ResNet101 is used in the spatial feature extraction network, and the pre-training is carried out on ImageNet, and the spatiotemporal fusion ratio is 2:1. Under the same conditions, compared with TRN, the accuracy of RGB channel is increased by 1.4% by adding META module. After cross modal pre-training, the accuracy of optical flow channel is improved by 15.7%, and the fusion is 10.6% higher than TRN. It can be seen that the proposed META module can extract the information of action in time dimension to the greatest extent compared with ResNet101 network, the accuracy of action recognition is improved.

Fig. 14 shows the change of value accuracy and loss when the human action recognition network based on temporal modeling is trained on HMDB51. Fig. 14 (a) corresponds to the META-RGB method in Table 2, and Fig. 14 (b) corresponds to the META-Flow method in Table 2.

Our method is also compared with the study under the same experimental conditions. The experimental results are shown in Table 3. The proposed META improves the accuracy of ResNet101 significantly, and the human action recognition network based on temporal modeling is better than other studies.

The accuracy rates of TSN and TEA (pre-train: ImageNet) in Table 3 on HMDB51 dataset are the results of this experiment under the above experimental conditions. In the reproduction, TSN does not use cross modality pre-training for optical flow, but directly uses ImageNet’s pre-training model. In the reproduction of TEA, we did not use the pre-training model of Kinetics as in the original paper. The above method makes the pre-training conditions the same as other methods, so that the results can be compared.

It can be seen from Table 3 that our method is 14.0% higher than the best traditional method IDT; It is 7.6% higher than the two-stream three-dimensional convolutional network T3D; Under the same backbone network and pre-training conditions, our method is 11.8% higher than the common two-stream network after fusion. Under the same pre-training model, our method has higher accuracy than TEA. Compared with TEA pre-trained with two datasets, our method is slightly backward, but saves the pre-training time. It can be seen that the META proposed in this paper, combined with ResNet101, improves the utilization of temporal features and improves the accuracy of action recognition.

In order to further evaluate the performance of this model, this paper supplementary verifies the human action recognition method based on motion excitation and time aggregation module on HMDB51 data set. We calculated the sensitivity, precision and F1-score of the model, which are shown in Table 4. However, there are few studies on human action recognition with the above measurement parameters, so it is not compared with other studies.
Table 3. Comparison results with other studies on HMD851 dataset.

| Method               | Backbone       | Pre-train       | Top-1 (%)        |
|----------------------|----------------|-----------------|------------------|
| IDT [3]              | None           | None            | 57.2%            |
| Two-Stream-Fusion [4] | ResNet101      | ImageNet        | 59.4%            |
| T3D+TSN [29]         | T3D-169        | ImageNet        | 63.5%            |
| Res3D [30]           | 3D-ResNet      | Sports-1M       | 54.9%            |
| TSN [5]              | BN-Inception   | ImageNet        | 69.4%            |
| TEA [17]             | ResNet50       | ImageNet + Kinetics | 73.3%        |
| Reference [31]       | GoogLeNet      | ImageNet        | 51.8%            |
| TSN + Bi-LSTM [32]   | DenseNet169    | ImageNet        | 70.1%            |
| META-RGB (Ours)      | ResNet101/BN-Inception | ImageNet     | 55.1%            |
| META-Flow (Ours)     | ResNet101      | ImageNet        | 62.0%            |
| META-RGB + Flow (Ours)| ResNet101/BN-Inception | ImageNet     | 71.2%            |

Table 4. Evaluating parameters of the model.

| Method   | Backbone | Precision  | Sensitivity | F1-score |
|----------|----------|------------|-------------|----------|
| META     | ResNet101| 71.45%     | 70.83%      | 71.16%   |

Table 5. Experimental results on UCF101 dataset.

| Method            | Backbone       | Pre-train       | Top-1 (%)        |
|-------------------|----------------|-----------------|------------------|
| TRN-RGB [6]       | ResNet101      | ImageNet        | 85.6%            |
| TRN-Flow [6]      | ResNet101      | ImageNet        | 80.1%            |
| TRN-RGB + Flow [6]| ResNet101      | ImageNet        | 91.9%            |
| META-RGB (Ours)   | ResNet101/BN-Inception | ImageNet     | 86.2%            |
| META-Flow (Ours)  | ResNet101/BN-Inception | ImageNet     | 92.3%            |
| META-RGB + Flow (Ours)| ResNet101/BN-Inception | ImageNet     | 96.0%            |

UCF101. Fig. 15 (a) corresponds to the META-RGB method in Table 5, and Fig. 15 (b) corresponds to the META-Flow method in Table 5.

Our method is also compared with the study under the same experimental conditions. The experimental results are shown in Table 6. The proposed META module improves the accuracy of resnet101 significantly, and the human action recognition network based on temporal modeling is better than other studies.

The accuracy rates of TEA (pre-train: ImageNet) in UCF101 dataset in Table 6 are the results of this experiment under the above experimental conditions. In the reproduction of TEA, it is consistent with Table 1, and does not use the pre-training model of Kinetics in the original paper. Where fusion is involved in the table, RGB and optical flow are fused at a ratio of 1:1.5. The above method makes the pre-training conditions the same as other methods, so that the results can be compared.

As can be seen from Table 6, our method is 11.1% higher than the best traditional method IDT; For the same two stream network, our accuracy is 8% and 1.8% higher than that of the original two stream and TSN, respectively; Our method is even 3.6% higher than ECO12F, which uses 3D convolution network and larger pre-training dataset; Under the same pre-training conditions, our method after fusion is 1.8% higher than TSN + Bi-LSTM; Under the same pre-training model, our method has higher accuracy than TEA. Compared with TEA pre-trained with two datasets, our method is slightly backward, but saves the pre-training time. This is consistent with the results obtained on HMDB-51 dataset.

In order to get the time needed to recognize an action using the method in this paper, we input a video with 30 frames per second and a total length of 10 seconds for action recognition. Our method can identify accurately in only 1.73 seconds, which fully shows the efficiency of this method.

In order to further evaluate the performance of this model, this paper supplementary verifies the human action recognition method based on motion excitation and time aggregation module on UCF101 data set. We calculated the sensitivity, precision and F1-score of the model, which are shown in the Table 7. However, there are few studies on human action recognition with the above measurement parameters, so it is not compared with other studies.

5. Conclusion

In this paper, we propose a human action recognition method based on temporal modeling. The focus of this method is to solve the problems of low modeling efficiency and feature loss in temporal modeling. This method constructs a spatiotemporal two stream network based on TRN framework, and uses temporal relational sampling to strengthen the relationship between frames in temporal modeling. Moreover, we propose META for temporal modeling. META is composed of Multi-scale Motion Excitation module (MME) and Squeeze and Excitation Temporal Aggregation module (SETA). MME integrates short-term temporal modeling...
Table 6. Comparison results with other studies on UCF101 dataset.

| Method          | Backbone             | Pre-train    | Top-1 (%)  |
|-----------------|----------------------|--------------|------------|
| IDT [3]         | None                 | None         | 85.9%      |
| Two-Stream-Fusion [4] | ResNet101            | ImageNet     | 88.0%      |
| TSN [5]         | BN-Inception         | ImageNet     | 94.2%      |
| ECO12 [33]      | BN-Inception-3D-ResNet18 | Kinetics    | 92.4%      |
| Spatiotemporal-3DCNN [34] | VGG-M-2048          | ImageNet     | 91.7%      |
| TEA [17]        | ResNet50             | ImageNet+Kinetics | 96.9%   |
| Reference [35]  | ResNet34             | ImageNet     | 91.5%      |
| StNet [21]      | ResNet50             | ImageNet+Kinetics | 93.5%   |
| Reference [29]  | GoogLeNet            | ImageNet     | 93.5%      |
| TSN+Bi-LSTM [30] | DenseNet169          | ImageNet     | 94.2%      |
| Temporal Squeeze Network [36] | TeSNet            | ImageNet     | 95.2%      |
| META-RGB (Ours) | ResNet101/BN-Inception | ImageNet     | 86.2%      |
| META-Flow (Ours) |                      |              | 92.3%      |
| META-RGB+Flow (Ours) |                    |              | 96.0%      |

Table 7. Evaluating parameters of the model.

| Method  | Backbone   | Precision | Sensitivity | F1-score |
|---------|------------|-----------|-------------|----------|
| META    | ResNet101  | 95.73%    | 96.85%      | 96.28%   |

into the whole spatiotemporal feature learning to solve the problem of low modeling efficiency. SETA forms a hierarchical structure through multiple sub convolutions to extract effective features of different temporal scales. Therefore, the final temporal receptive field is increased, and the problem of feature loss is solved. In order to further improve the modeling efficiency, we use the cross modality pre-trained network to extract the features of the optical flow image, which improves the utilization of the optical flow features. This method solves the problems of low efficiency and feature loss of temporal modeling from the perspectives of short-term temporal modeling and long-term temporal modeling. In this paper, HMDB-51 and UCF101 datasets were tested, and the accuracy rates were 71.2% and 96.0% respectively, which were higher than other research methods. Due to the limitation of experimental conditions, larger datasets were not used for pre-training. In the following research, if we use the pre-training of ImageNet and kinetics datasets, the recognition accuracy can be further improved.

Declarations

Author contribution statement

Qing Ye, Zexian Tan: Conceived and designed the experiments; wrote the paper.

Zexian Tan: performed the experiments; analyzed and interpreted the data; wrote the paper.

Qing Ye, Yongmei Zhang: contributed reagents, materials, analysis tools or data; wrote the paper.

Funding statement

Professor Qing Ye was supported by Technology Project of Beijing Municipal Education Commission [SQKM201810009002], National Natural Science Foundation of China [61371143 & 61806008], Natural Science Foundation of Beijing Municipality [4192022], Science and Technology Development Center, Ministry of Education [2018A03029], National Key R&D Program of China [2020YFB1600702].

Data availability statement

Data will be made available on request.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.
References

[1] H.I. Luo, K. Tong, F.S. Kong, The progress of human action recognition in videos based on deep learning: a review, Acta Electron. Sin. 47 (5) (2019) 1162–1173.
[2] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2–3) (2005) 107–123.
[3] H. Wang, C. Schmid, Action recognition with improved trajectories, in: 2013 IEEE International Conference on Computer Vision, IEEE, 2014, pp. 3551–3558.
[4] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1933–1941.
[5] L. Wang, Y. Xiong, Z. Wang, et al., Temporal segment networks: towards good practices for deep action recognition, in: Computer Vision - 14th European Conference (ECCV 2016), vol. 9912, 2016, pp. 20–36.
[6] B. Zhou, A. Andonian, A. Torralba, Temporal relational reasoning in videos, in: 15th European Conference on Computer Vision, ECCV, vol. 1205, 2017, pp. 831–846.
[7] G.X. Huang, G.B. Adrien, Learning spatio-temporal representations with temporal squeeze pooling, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2103–2107.
[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489–4497.
[9] T. Du, H. Wang, L. Torresani, et al., A closer look at spatiotemporal convolutions for action recognition, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6450–6459.
[10] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy, Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification, in: 18th European Conference on Computer Vision (ECCV 2018), vol. 12129, 2018, pp. 318–335.
[11] Z. Qu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3D residual networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5534–5542.
[12] X. Li, X. Li, L. Wang, J. Wang, Y. Qiao, CT-net: channel tensorization network for video classification, in: 2021 International Conference on Learning Representations (ICLR), 2021, pp. 670–688.
[13] J. Lin, C. Gan, S. Han, TSM: temporal shift module for efficient video understanding, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 7082–7092.
[14] H. Shao, S. Qian, Y. Liu, Temporal interlacing network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 11966–11973.
[15] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Tong Lu, TEINet: towards an efficient architecture for video recognition, in: AAAI, 2020, pp. 11669–11676.
[16] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, Tong Lu, TAM: temporal adaptive module for video recognition, in: ICCV, 2020, pp. 13688–13698.
[17] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, L. Wang, TEA: temporal excitation and aggregation for action recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 906–915.
[18] Z. Wang, Q. She, A. Smolic, ACTION-net: multipath excitation for action recognition, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13209–13218.
[19] L. Wang, Z. Tong, B. Ji, G. Wu, TDN: temporal difference networks for efficient action recognition, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1895–1904.
[20] B. Jiang, M. Wang, W. Gan, W. Wu, J. Yan, STM: SpatioTemporal and motion encoding for action recognition, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2000–2009.
[21] C. Feichtenhofer, H. Fan, J. Malik, K. He, SlowFast networks for video recognition, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6201–6210.
[22] Y.H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4695–4702.
[23] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1510–1517.
[24] L.D. He, C.Z. Zhou, C. Gan, F. Li, X. Liu, D.Y. Li, L. Wang, S.L. Wen, Siblet: local and global spatial-temporal modeling for action recognition, in: AAAI, 2019, pp. 8401–8408.
[25] S. Khurram, R.Z. Amir, S. Mubarak, UCF101: a dataset of 101 human action classes from videos in the wild, in: 2012 Center for Research in Computer Vision, 2012, pp. 1154–1162.
[26] H. Kurhane, H. Huang, E. Garrote, T. Poggio, T. Serre, HMDR: a large video database for human motion recognition, in: 2011 International Conference on Computer Vision, 2011, pp. 2556–2563.
[27] S. Gao, M.M. Cheng, K. Zhao, et al., Res2Net: a new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. (99) (2019) 1.
[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[29] A. Diba, M. Fayyaz, V. Sharma, et al., Temporal 3D ConvNet: new architecture and transfer learning for video classification, https://arxiv.org/abs/1711.08206, 2019.
[30] D. Tran, J. Roy, Z. Shou, et al., ConvNet architecture search for spatiotemporal feature learning, https://arxiv.org/abs/1708.05038v1, 2018.
[31] S. Chen, W. Wei, B. He, S. Chen, J. Liu, Human motion recognition method based on improved deep convolutional neural network, Appl. Res. Comput. 36 (03) (2019) 945–949, 953.
[32] Qian Li, Wenzhu Yang, Xiangyang Chen, Dongdong Yuan, Yuxia Wang, Human action recognition model based on tight coupling temporal and spatial double-flow convolutional neural network, J. Comput. Appl. 40 (11) (2020) 3178–3183.
[33] M. Zolfaghari, K. Singh, T. Brox, ECO: efficient convolutional network for online video understanding, in: The European Conference on Computer Vision, 2018, pp. 695–712.
[34] Y. Tianming, C. Zhi, Y. Wenjing, Spatial and temporal double flow person action recognition model based on video deep learning, J. Comput. Appl. 38 (3) (2018) 895–899, 915.
[35] Shu-ping Lv, Yi Huang, Ying-ying Wang, Research on human motion recognition based on two stream convolutional neural network, Exp. Technol. Manag. 38 (08) (2021) 144–148.
[36] Guoxi Huang, Adrian G. Bors, Learning spatio-temporal representations with temporal squeeze pooling, https://arxiv.org/pdf/2005.04685v1.pdf.