Hindawi
Mobile Information Systems
Volume 2022, Article ID 4213335, 10 pages
https://doi.org/10.1155/2022/4213335

Research Article

A Video Classification Method Based on Spatiotemporal Detail Attention and Feature Fusion

Xuchao Gong and Zongmin Li

School of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China

Correspondence should be addressed to Xuchao Gong; gongxuchao2010@163.com

Received 28 February 2022; Revised 15 April 2022; Accepted 20 April 2022; Published 28 April 2022

Academic Editor: Mian Ahmad Jan

With the explosive growth of Internet video data, demands for accurate large-scale video classification and management are increasing. In the real-world deployment, the balance between effectiveness and timeliness should be fully considered. Generally, the video classification algorithm equipped with time segment network is used in industrial deployment, and the frame extraction feature is used to classify video actions However, the issue of semantic deviation will be raised due to coarse feature description. In this paper, we propose a novel method, called image dense feature and internal significant detail description, to enhance the generalization and discrimination of feature description. Specifically, the location information layer of space-time geometric relationship is added to effectively engrave the local features of convolution layer. Moreover, the multimodal feature graph network is introduced to effectively improve the generalization ability of feature fusion. Extensive experiments show that the proposed method can effectively improve the results on two commonly used benchmarks (kinetics 400 and kinetics 600).

1. Introduction

Video classification is the task of automatically identifying the category of input video. The key challenges are not only to serialize and understand the spatiotemporal relationship in the image but also to abstract the most important information and predict the categories with the extracted representation. In the last decade, video classification network has achieved great success, and a series of excellent methods [1–5] have been proposed. At the same time, many benchmark datasets have been built [7, 15].

In the process of image recognition, the two spatial dimensions of width and height are processed through convolution operation. Many experiments show that [8, 9] image has the characteristics of isotropy such as translation invariance under the first-order approximation. Corresponding to the video, the object motion contained in it has the same characteristics of consistent direction [10]. Therefore, it is effective to adopt the method of frame extraction for video classification [1–4], which is one commonly used solution in industry.

The methods of dual stream video classification [1, 3, 11, 12] are favored by the industry in specific practical applications. When implemented, this kind of methods model RGB and optical flow data, respectively, and connect features for video classification. However, they are based on the complete features of video frames [1–3, 11, 12] and ignore the importance of features in different spatiotemporal positions.

As shown in Figure 1, during penguin diving, the main body penguin will have limb changes in action with the progress of time, and the spatial correlation between the motion scene and action will also be reflected. The correlation is shown by the red-dashed arrow in Figure 2; there will be the correlation of actions and the consistency of background, whether between the adjacent grids of the current frame or between different frames adjacent to time.

On the whole, the dual stream video classification methods have several important problems:

(1) Frame level features describe global information, lack of key region feature mining, and insufficient

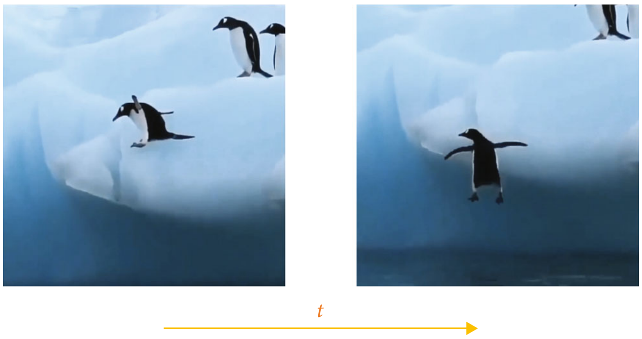

Figure 1: Video subject motion spatial-temporal information example.

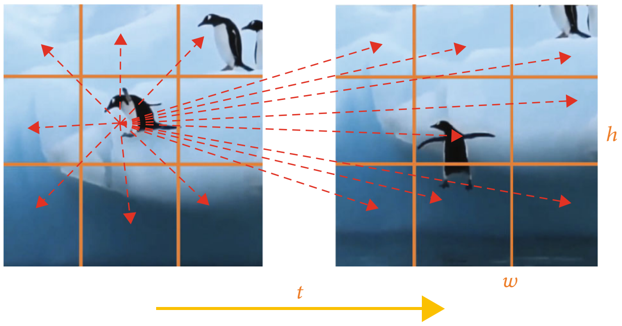

Figure 2: Video subject motion spatial-temporal information example.

modeling for the proportion of action subject relative image

(2) Optical flow and image features are often learned separately, and the lack of interaction between features leads to the inability to fuse the two effectively. Therefore, there is still a lot of room for improvement in the method

In order to solve these problems, this paper proposes a video classification method based on image dense features and internal salient detail description. The proposed method has the following advantages:

(1) We propose to find the key motion features. By combining spatiotemporal location coding feature, the key motion areas can be found more easily

(2) With the help of multimodal feature fusion, which through cross-attention mechanism, RGB features and optical flow features can fully interact with each other

2. Related Work

At present, video classification includes methods based on spatiotemporal representation and video stream recognition. In the methods based on spatiotemporal representation, actions can be expressed as the changes of spatiotemporal objects in time and space. And the key feature can be filtered and captured by spatiotemporal directional modeling [13, 14]. In order to take the spatiotemporal relationship in feature extraction into account, 3D convolution [6, 15, 16] expands the 2D spatial image model [17–20] to spatial-temporal domain. Among them, I3D network [15] uses dual stream CNN with expanded 3D convolution on RGB and optical flow sequences, which has achieved good results in video classification. Some other methods focus on serialization timing modeling. The common method is CNN+LSTM [1, 12, 21–23], and they use CNN to extract frame features and LSTM to integrate temporal features. Considering that the separate processing of spatiotemporal information can effectively improve the computational performance, there are some schemes to decompose the convolution into separate two-dimensional space and one-dimensional time filters [24–27]. Spatiotemporal joint feature filtering and fusion, including separable spatiotemporal feature modeling, can quickly grasp the main features and effectively model the key information of video actions. It is also a popular implementation and deployment method in the industry. However, its disadvantages are also obvious; frame level features describe the global information, lack of key area feature mining, and insufficient modeling for the relatively small proportion of the action subject's relative image.

In the early stage, there were more classic manual feature design methods based on optical flow, such as flow histogram [28], motion boundary histogram [29], and trajectory graph [30]. Before the popularization of deep learning, these methods have shown good results in motion recognition. In the context of deep neural network, the dual stream method [11] effectively uses optical flow to obtain the key information of action by considering optical flow as another video input mode. This idea has also been effectively verified in many other methods [1, 12, 24], and some progress has been made. In the TSN [1] method, the video is divided into multiple segments, and the sample of RGB frame image and optical flow image is randomly selected in different time periods to extract the information for video action recognition. However, since the two features of optical flow and image are often learned separately and the lack of interaction between the features, the two sources cannot be fused effectively, so there is still a lot of room for improvement in the method.

With the development of visual attention methods, they are widely used in video understanding tasks. For video summarization, in order to optimize the recognition effect, a dynamic and static visual attention method is prosed [31], and a global&local multihead attention method is adopted in [32]. GSE-GCN [33] proses a Granularity-Scalable Ego-Graph Convolutional Network for obtaining a more satisfying summary. And [34] uses static and motion features with the parallel attention to improve video summarization results. For video classification, ViS4mer [35] uses a multiscale temporal S4 decoder for subsequent long-range temporal inference. MViT [36] proses window attention and pooling attention operations for calculating local information and aggregating them. In order to evaluate the movements of infants in the video, [37] uses the spatiotemporal attention selection mechanism. For solving the differences between features, BA-Transformer [38] applies different attention operations to different feature channel groups.

Recently, the video classification methods based on transformer has made great progress. Video Transformer Net [39] adds the time attention on the basis of the pretrained VIT model [40], which has produced good performance. TimeSformer [41] studied five structures in spatiotemporal modeling and proposed a spatiotemporal attention mechanism based on factorization. The experimental results show that the algorithm achieves a good tradeoff between speed and performance. Based on the picture classification structure, Video Swin Transformer [42] adds the time dimension, and good results are achieved. ViViT [43] discussed four different ways to realize spatiotemporal attention on the basis of VIT [40]. In order to reduce the amount of computation of the model, tokshift transformer [44] proposes a pure convolution free video classification algorithm. Multiscale ViT [45] is a multiscale visual transformer for video classification; for reducing the amount of computation, spatiotemporal modeling is carried out through the attention mechanism, and this method has achieved good results.

Because the algorithm in this paper is to improve the methods based TSN [1], which are commonly used in industry, the video classification methods based on transformer are not compared in this paper.

3. Video Classification Based on Salient Detail Description of Video Image

This paper proposes a video classification method based on salient detail description of video images (VCM-SDD). The flow of the proposed method is shown in Figure 3, where our VCM-SDD includes two serial feature fusion architectures. The first one is spatiotemporal consistency modeling, which operates RGB image and optical flow image separately. We can find the significant details of the time domain and the spatial domain under the single mode. The second one is multimodal feature fusion, which achieves feature interaction by using RGB image features and optical flow image features as query and key/value items to each other. The fusion of these two architectures is used for final classification.

3.1. Spatiotemporal Consistency Modeling. Many experiments show that [8, 9] images have isotropic characteristics such as translation invariance under the first-order approximation. At the same time, the image feature based on nonoverlapping grid can describe the characteristics of moving subject in more detail. No matter between adjacent grids of the current frame or between frames with similar time, there will be action correlation and consistency of action occurrence background. Therefore, it is feasible to mine the salient details of actions under the condition of spatiotemporal consistency modeling. We represent a video as $V = R^{N*t*c*h*w}$; that is, a video is first decomposed into n segments from the whole video. The video length of each subsegment is T, C represents the number of image channels, and h and w represent the width and height of the image, respectively. Image appearance features (RGB) and motion optical flow features are extracted from each subsegment of the video.

In the RGB image grid feature extraction, the last convolution layer is selected to obtain a two-dimensional image feature description:

$$F_{\text{rgb\_conv}} = \text{Conv}_n(\text{Conv}_{n-1}(\cdots(\text{Conv}_1(I_{\text{RGB}})))), \tag{1}$$

where n represents the serial number of the convolution layer, conv$_n$ represents the convolution feature map of the n $_{\text{th}}$ layer in the network, I_{RGB} is the input image, and $f_{\text{rgb\_Conv}}$ represents the n $_{th}$ layer convolution feature of the extracted RGB image. The final output dimension is $n * n * D$, where n represents the resolution of the feature map and D represents the number of feature map channels.

In the optical flow image grid feature extraction, the same network as the image feature extraction is selected, and is specifically expressed as

$$F_{\text{flow\_conv}} = \text{Conv}_n(\text{Conv}_{n-1}(\cdots(\text{Conv}_1(I_{\text{flow}})))), \tag{2}$$

where n represents the serial number of the convolution layer, conv$_n$ represents the convolution feature mapping of the n $_{th}$ layer in the network, I_{flow} is the input image, and $I_{\text{flow\_Conv}}$ represents the n $_{th}$ layer convolution feature of the extracted optical flow image, and the final output dimension is $n * n * D$, where n represents the resolution of the feature map and D represents the number of feature map channels.

In order to better integrate the relative position information of visual features, the relative position information according to the grid geometry is added. The bounding box of the region can be expressed as (x, y, W, H), where x , y, W, and H represent the central coordinates of the grid and its width and height; therefore, for any two grids grid$_i$ and grid$_j$, their geometric relationship as a 4-dimensional vector is expressed:

$$\text{geom\_img}(i, j) = \left(\log \left(\frac{|x_i - x_j|}{w_i} \right)^{-1}, \log \left(\frac{|y_i - y_j|}{h_i} \right)^{-1}, \log \left(\frac{|w_j|}{w_i} \right)^{-1}, \log \left(\frac{|h_j|}{h_i} \right)^{-1} \right) \tag{3}$$

Considering $w_i = w_j$ and $h_i = h_j$ in the grid feature, the last two terms of the geometric relationship vector in the above formula can be removed, so it can be simplified as

$$\text{geom\_img}(i, j) = \left(\log \left(\frac{|x_i - x_j|}{w_i} \right)^{-1}, \log \left(\frac{|y_i - y_j|}{h_i} \right)^{-1} \right). \tag{4}$$

It can be seen that the farther the distance, the smaller the corresponding geometric relationship value. The relationship between different grids is shown in Figure 4. Considering that the interframe features are adjacent in time

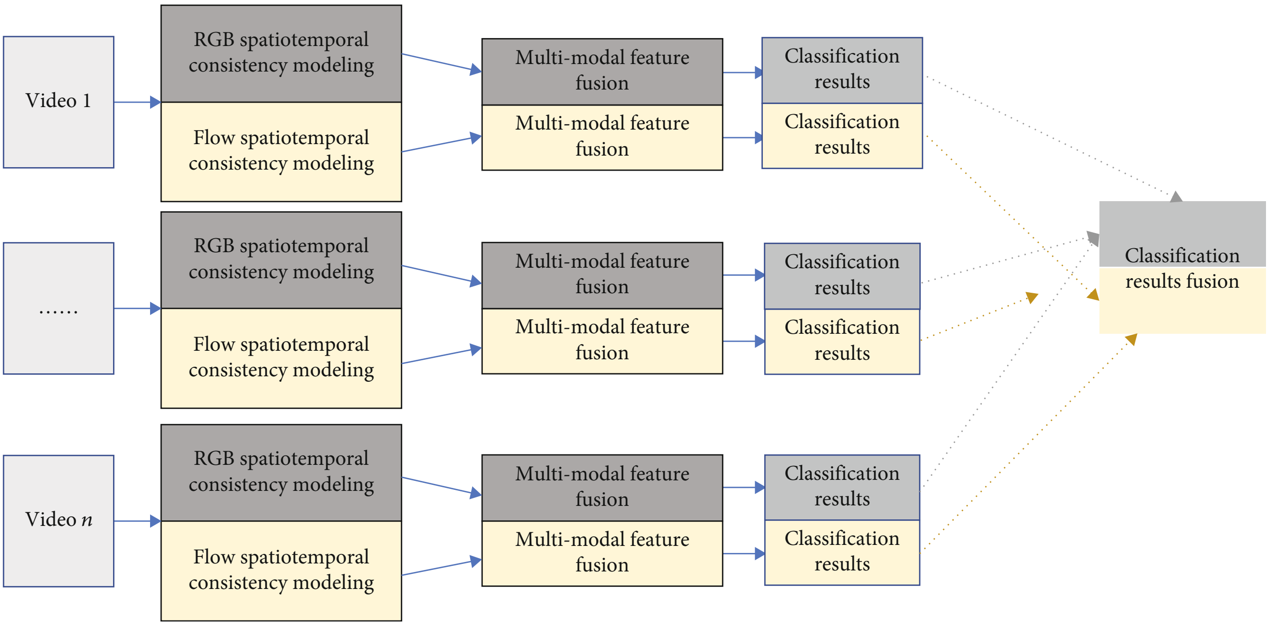
Video 1
RGB spatiotemporal consistency modeling
Flow spatiotemporal consistency modeling
Multi-modal feature fusion
Multi-modal feature fusion
Classification results
Classification results
......
RGB spatiotemporal consistency modeling
Flow spatiotemporal consistency modeling
Multi-modal feature fusion
Multi-modal feature fusion
Classification results
Classification results
Video n
RGB spatiotemporal consistency modeling
Flow spatiotemporal consistency modeling
Multi-modal feature fusion
Multi-modal feature fusion
Classification results
Classification results
Classification results fusion

Figure 3: The framework of the proposed method.

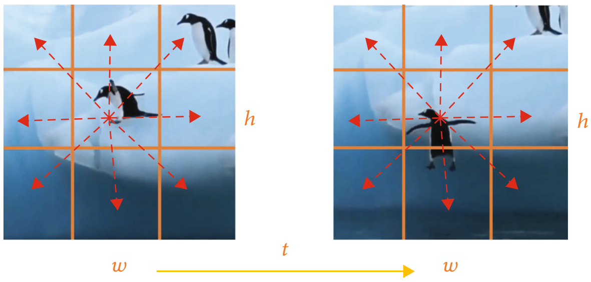

Figure 4: Geometric relationship measurement between grids.

sequence, their expression will be closer. We propose a geometric measure of time relationship:

$$\text{geom\_time}(i, j) = \log\left(\frac{|t_i - t_j|}{T}\right)^{-1}. \qquad (5)$$

In the above formula, t_i and t_j, respectively, represent the time of the current feature, and T is the number of frames extracted from the current video clip. It can be seen from the above formula that the closer the distance, the greater the value; the geometric relationship measurement after spatial-temporal combination can be finally expressed as

$$\text{geom\_all}(i, j) = \left(\log\left(\frac{|x_i - x_j|}{w_i} + e\right)^{-1}, \log\left(\frac{|y_i - y_j|}{h_i} + e\right)^{-1}, \log\left(\frac{|t_i - t_j|}{T} + e\right)^{-1}\right). \qquad (6)$$

In order to prevent abnormal logarithm operation of geometric relationship in the same spatial position or time position, we add an offset term e after each feature. Next, we build the feature similarity matrix according to the spatial-temporal geometric relationship:

$$\text{sim}(i, j) = \text{RELU}\left(\text{geom\_all}(i, j)W_p\right)W_g. \qquad (7)$$

In the above formula, W_p is a learnable feature mapping matrix, and its dimension is $3 * 32$ in the experiment. W_g is a learnable similarity mapping matrix, and the dimension is set to $32 * 1$ during the experiment. The specific method is to expand the geometric relationship vector to the high-dimensional space through the mapping matrix W_p and carry out the similarity mapping W_g after RELU activation to obtain the similarity eigenvalue.

The similarity eigenvalue proposed in this paper can be regarded as a spatiotemporal constrained cross attention mechanism. The consistency and difference between regions can be measured to realize the feature fusion between regions. To avoid introducing semantic noise, we create a geometric alignment graph $G = (V, E)$. The grid features extracted with time are represented as independent node V. In the dimensions of time, width, and height, when the distance between nodes is less than or equal to 2, the grid nodes will be connected to form edge set E. According to the above rules, we construct an undirected graph. According to the geometric position relationship information of equation (7), the weight matrix W can be obtained and normalized:

$$s_{i,j} = \frac{e^{\text{sim}(i,j)}}{\sum_{j\in A(v_i)} e^{\text{sim}(i,j)}}. \qquad (8)$$

$S_{i,j}$ represent the normalized similarity between node i and node j. V_i represents the characteristic node of the grid, and $A(V_i)$ represents all neighbor nodes adjacent to node V_i. After the above similarity operation, the feature output

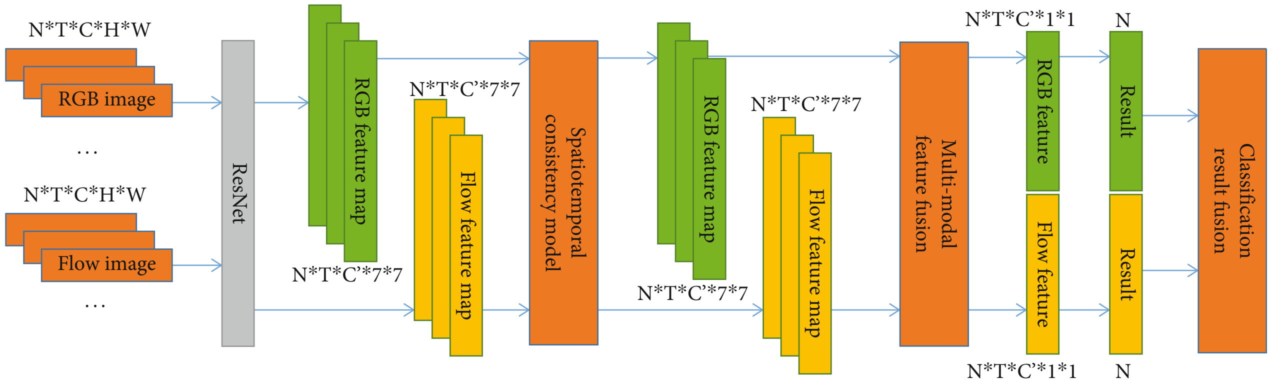

Figure 5: The structure of the model in this paper.

Table 1: Experimental parameter setting and hardware selection.

	lr	Epochs	Batch size	Optimizer	GPU	Platform
Kinetics 400	0.008	100	128	Adamw	TeslaV100∗8	Pytorch1.6
Kinetics 600	0.012	100	128	Adamw	TeslaV100∗8	Pytorch1.6

after RGB spatial-temporal fusion can be expressed as

$$F_{\text{fusion}} = \sum_{j\in A(V_i)} s_{i,j} V_j. \tag{9}$$

RGB image grid features and optical flow image grid features can obtain their respective spatial-temporal fusion features through the above operations. This operation can extract the significant visual information in the image. It is helpful to improve the effect of video classification. The following experiments also prove the effectiveness of the proposed algorithm. The specific features after the fusion of RGB image and optical flow image are expressed as follows:

$$\begin{aligned} F_{\text{rgb\_fusion}} &= \sum_{j\in A(V_i)} s_{i,j} F_{\text{rgb\_conv}}, \\ F_{\text{flow\_fusion}} &= \sum_{j\in A(V_i)} s_{i,j} F_{\text{flow\_conv}}. \end{aligned} \tag{10}$$

3.2. Multimodal Feature Fusion. This paper is an improvement of TSN [1] and TSM [2], which belongs to the segmented video classification method commonly used in industry. The two features of optical flow and image are studied separately, and there is a lack of interaction between the features, which makes it impossible to fuse them effectively. In order to solve this problem, this paper carries out feature fusion modeling by means of cross attention mechanism and used as retrieval items and value/key items to each other. At this stage, RGB image grid features and optical flow image grid features alternately act as retrieval items and numerical items. For RGB image features, to fuse with optical flow features, the modeling method of cross attention can be expressed as

$$F_{\text{att\_rgb}} = \operatorname{softmax}\left(\frac{F_{\text{rgb\_fusion}} * F_{\text{flow\_fusion}}}{\sqrt{\left\| F_{\text{rgb\_fusion}} * F_{\text{flow\_fusion}} \right\|}}\right) F_{\text{flow\_fusion}}. \tag{11}$$

For optical flow image features, to fuse with RGB features, cross attention can be expressed as

$$F_{\text{att\_flow}} = \operatorname{softmax}\left(\frac{F_{\text{rgb\_fusion}} * F_{\text{flow\_fusion}}}{\sqrt{\left\| F_{\text{rgb\_fusion}} * F_{\text{flow\_fusion}} \right\|}}\right) F_{\text{rgb\_fusion}}. \tag{12}$$

Considering the above two equations, $f_{\text{rgb\_fusion}}$ and $f_{\text{flow\_fusion}}$ feature dimension increases the amount of calculation. This will lead to long reasoning and training time. In this paper, the method similar to that in transformer [46] is used to disassemble and group the features. The features with larger dimensions are divided into multiple parts for calculation according to the method of multihead calculation.

3.3. Classification Result Fusion. In TSN [1], in the video result prediction stage, all segment recognition networks share model parameters. The learned model performs frame level evaluation like a normal image network. The details are as follows:

$$R_{\text{clip}} = \text{AVG}\left[M\left(\text{clip}_{\text{rgb}_i}\right), M\left(\text{clip}_{\text{flow}_i}\right)\right], \quad \text{where} \quad i = 1, \cdots, n, \tag{13}$$

Table 2: Exploration of spatial-temporal consistency and multimodal fusion of kinetics 400.

	Top1	Top5
VCM-SDD No_STC No_MFF	71.7	90.9
VCM-SDD STC No_MFF	73.9	91.4
VCM-SDD No_STC MFF	78.5	93.6
VCM-SDD STC MFF	80.1	94.4
TSN [1]	71.3	91.5
TSM [2]	75.1	91.8

Table 3: The comparison between this algorithm and other methods on the kinetics 400.

	Top1	Top5	GFLOPs
I3D [15]	72.1	90.3	108
Two-stream I3D [15]	75.7	92.0	216
S3D-G [26]	77.2	93.0	—
Nonlocal R50 [47]	76.5	92.6	—
Nonlocal R101 [47]	77.7	93.3	—
R(2 + 1)D Flow [25]	67.5	87.2	152
STC [48]	68.7	88.5	—
ARTNet [49]	69.2	88.3	23.5
S3D [26]	69.4	89.1	66.4
ECO [50]	70.0	89.4	216
R(2 + 1)D [25]	73.9	90.9	152
TSN [1]	71.3	91.5	33
TSM [2]	75.1	91.8	65
SlowFast 16 * 8, R101 [3]	78.9	93.5	213
SlowFast 16 * 8, R101+NL [3]	79.8	93.9	234
VCM-SDD 8 * 6, R101_NP	77.4	93.1	46.8
VCM-SDD 8 * 6, R101	78.5	93.5	46.8
VCM-SDD 16 * 6, R101_NP	79.3	93.9	46.8
VCM-SDD 16 * 6, R101	**80.1**	**94.4**	**46.8**

where $\text{clip}_{\text{rgb}i}$ and $\text{clip}_{\text{flow}i}$ represent the result of RGB image and optical flow frame extraction in segment i of the video. Here, one frame is extracted from each video, M represents the result of single-stage model reasoning, and R_{clip} represents the final classification result. The method of multisegment averaging is adopted.

Different from the above fusion results, here the multimodal fusion feature is used:

$$R_{\text{clip}} = \text{AVG}\begin{bmatrix} M\left(\text{avg\_}F\left(F_{\text{att\_rgb}}\right)\right), \\ M\left(\text{avg\_}F\left(F_{\text{att\_rgb}}\right)\right) \end{bmatrix}, \quad \text{where} \quad i = 1, \cdots, n, \tag{14}$$

where avg_F means that the extracted multiframe image features are averaged. The reason why the average value is used here is that there is action continuity in multiple frames. Using the average value to comprehensively measure the characteristic value can effectively reflect the significance detail description, $F_{\text{att\_RGB}}$ and $f_{\text{att\_Flow}}$ represent the fusion features obtained by taking RGB image and optical flow image as retrieval items, respectively.

4. Experiments

4.1. Model and Parameter Setting. The model in this paper is shown in Figure 5. Firstly, image features and optical flow features are extracted according to ResNet101. Then, the spatiotemporal information is added through spatiotemporal consistency modeling. And obtain modal interaction information through multimodal feature fusion. Video classification is carried out after feature enhancement. The number in Figure 5 shows the shape information of the feature.

In order to explore the impact of different video segments on the classification results, 8 and 16 clips of a video are collected evenly along the time axis during the test. For each segment, 6 frames of images are evenly extracted for feature extraction, and finally, the results are fused in the way of average of the results described in Section 3.3.

Table 1 shows the parameter settings and hardware information in the experiment. When extracting RGB image and optical flow features, similar to TSN and TSM methods, firstly scale the images to 256 ∗ 256, get 224 ∗ 224 images by center cropping. In order to facilitate comparison, this paper selects Top1 and Top5 evaluation indicators commonly used in video classification.

4.2. Experiment Dataset. In the experiment, in order to evaluate the effectiveness of the algorithm proposed in this paper, we evaluated it on kinetics dataset [15]. Kinetics is a large-scale and high-quality YouTube video URL dataset, which contains many human action markers. The dataset was released by DeepMind to help the research of machine learning on video understanding. It contains two different versions according to different categories. The kinetics 400 contains about 260 K video clips, including 240 k training data and 20 K verification data, covering 400 types of human actions, with at least 400 video clips for each type of action. Each clip is about 10 seconds long and is marked with an action category. All clips are manually annotated in multiple rounds, and each clip comes from a separate YouTube video. These actions include a wide range of human object interaction actions, such as playing musical instruments, and human-human interaction actions, such as shaking hands and hugging. Kinetics 600 contains 420 K YouTube videos, including 392 k training data and 30 K verification data, with a total of 600 categories. Each category has at least 600 videos, and each video lasts about 10 seconds. At the same time, kinetics is also the basic dataset of the international human action classification competition organized by ActivityNet.

4.3. Comparison and Discussion. In order to prove the effectiveness of the algorithm, experiments are carried out on whether to add spatiotemporal consistency and multimodal feature fusion. Table 2 shows the corresponding experimental results. The experiment is based on the pretraining model of ResNet101 on ImageNet as the feature extraction model

Table 4: The comparison between this algorithm and other methods on the kinetics 600.

	Top1	Top5	GFLOPs
I3D [15]	71.9	90.1	108
StNet-IRv2 RGB [51]	79.0	—	—
AttentionNAS [5]	79.8	94.4	—
LGD-3D R101 [52]	81.5	95.6	—
SlowFast 16 $*$ 8, R101 [3]	81.1	95.1	213
SlowFast 16 $*$ 8, R101+NL [3]	81.8	95.1	234
TSN [1]	71.7	90.6	33
TSM [2]	75.6	92.1	65
VCM-SDD 8 × 6, R101_NP	79.6	94.3	46.8
VCM-SDD 8 × 6, R101	80.4	94.7	46.8
VCM-SDD 16 × 6, R101_NP	81.3	94.9	46.8
VCM-SDD 16 × 6, R101	81.9	95.3	46.8

of RGB image and optical flow image. The video is divided into 16 segments, and 6 frames are extracted from each segment.

VCM-SDD No_STC No_MFF means that do not include the spatial-temporal consistency and multimodal feature fusion in model training and inference, from which we can see that the effect is the worst.

VCM-SDD STC No_MFF means that spatial-temporal consistency is added in training and reasoning, but there is no multimodal fusion. Compared with the experimental items that are not added, Top1 and Top5 are increased from 71.7 and 90.9 to 73.9 and 91.4, respectively, with absolute values of 2.2 and 0.5. Experiments show the effectiveness of the spatiotemporal consistency algorithm.

VCM-SDD No_STC MFF indicates that there is no spatial-temporal consistency in training and reasoning, but multimodal fusion is added. Compared with the experimental items with spatial-temporal consistency, Top1 and Top5 are increased from 73.9 and 91.4 to 78.5 and 93.6, respectively, with absolute values of 4.6 and 2.2. It can be seen that multimodal fusion is more critical to the improvement of classification performance.

VCM-SDD STC MFF indicates that spatial-temporal consistency and multimodal feature fusion are added in the experiment. The accuracy rates of Top1 and Top5 are 80.1 and 94.4, respectively, which is the highest in the whole ablation experiment, which proves the effectiveness of the algorithm proposed in this paper.

Compared with the benchmark algorithms TSN and TSM, VCM-SDD STC MFF has made significant progress in kinetics 400. This is mainly due to spatial-temporal consistency and multimodal feature fusion. If without these two operations, the result of VCM-SDD No_STC No_MFF is similar to TSN but worse than TSM; this is mainly because the temporal and spatial correlations of features are not considered. When the spatial-temporal consistency operation is added, the temporal and spatial relationship between features is strengthened, and the VCM-SDD STC No_MFF result is better than TSN and slightly worse than TSM. When multimodal fusion is added, the cross modeling is carried out between different features. The generalization performance is strengthened, and the VCM-SDD No_STC MFF result is better than TSN and TSM. It shows that modal interaction plays a positive role in video classification algorithm. Compared with the separate operation, the video classification effect has been further improved after combining the two operations. It is proved that the operation of spatial-temporal consistency and multimodal fusion is effective. As shown in Table 3, the experiment in kinematics 600 also proves the effectiveness of the proposed algorithm.

4.4. Comparison of Experimental Results of Different Methods. Table 3 shows the comparison results of different methods on the kinetics 400; R101_NP is the result of not loading the pretraining model. It can be seen that the effect of dividing into 16 segments is better. Without ImageNet pretraining, the accuracy of Top1 divided into 16 segments is 79.3, and the accuracy of Top1 divided into 8 segments is 77.4, an increase of 1.9 percentage points. Top5 also has an improvement of 0.8 points. In the case of ImageNet pretraining, the accuracy of Top1 divided into 16 segments is 80.1, the accuracy of Top1 divided into 8 segments is 78.5, an increase of 1.6 percentage points, and the accuracy of top5 is also increased by 0.9 points.

This algorithm divides the video into 16 $*$ 6. With ImageNet pretraining, Top1 and Top5 are 80.1 and 94.4, respectively, which are 8.8 and 3.9 percentage points higher than TSN algorithm, 5.0 and 2.6 percentage points higher than TSM algorithm, and 0.3 and 0.5 percentage points higher than SlowFast algorithm with better performance. It can be seen from the experimental results that the combination of spatial-temporal consistency and multimodal fusion has a certain improvement in the image-based two-way recognition method.

In order to compare the amount of calculation between different methods, this paper compares the GFLOPs of each algorithm. It can be seen that compared with the baseline TSN and TSM, the amount of computation of this algorithm is between the two algorithms, and the computational efficiency of this algorithm meets the deployment requirements. Compared with SlowFast, the amount of calculation of this algorithm is significantly reduced.

Table 4 shows the comparison results of different methods on the kinetics 600 dataset. ResNet101 is also used as the backbone here. In order to explore the influence of the number of different video segments on the classification results, the video is also divided into 8 segments and 16 segments. It can be seen that the effect of being divided into 16 paragraphs is also better. Without ImageNet pretraining, the accuracy of Top1 divided into 16 segments was 81.3, and the accuracy of Top1 divided into 8 segments was 79.6, an increase of 1.7 percentage points; top5 also has an increase of 0.7 points. With ImageNet pretraining, the accuracy of Top1 divided into 16 segments was 81.9, and the accuracy of Top1 divided into 8 segments was 80.4, an increase of 1.5 percentage points; top5 also has an increase of 0.4 points. This algorithm divides the video into 16 $*$ 6. With ImageNet pretraining, Top1 and Top5 are 81.9 and 95.1, respectively, which are 10.2 and 4.5 percentage points higher than TSN algorithm, 5.3 and 3.0 percentage points higher than TSM algorithm, and 0.1 and 0.2 percentage points higher than SlowFast algorithm with better performance. From the experimental results in kinetics 600, it can be seen that the combination of spatial-temporal consistency and multimodal fusion has been improved in the image-based two-stream recognition method.

5. Conclusion

This paper presents a method to describe image dense features and internal salient details. It is used to enhance the generalization and distinguishability of feature description and improve the effect of video classification. In this paper, the location information layer of spatial-temporal geometric relationship is added to effectively carve the local features of convolution layer and enhance the ability of visual representation and detail description of local features of grid. The multimodal feature graph network interaction modeling mechanism is introduced to effectively improve the generalization ability of feature fusion. The results on the two datasets verify the effectiveness of the proposed method. At the same time, the proposed algorithm in this paper still has room for improvement. Firstly, this paper only models the grid features, while we find that the bounding box features of the moving subject have better expression performance. Secondly, we only fuse different modal features in the later stage of modeling. In future study, we will consider integrating modal fusion into the whole modeling process.

Data Availability

The data included in this paper are available without any restriction.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 61379106) and National Key R&D Plan of China (2019YF0301800).

References

[1] L. Wang, Y. Xiong, Z. Wang et al., “Temporal segment networks: towards good practices for deep action recognition,” in *European conference on computer vision*, pp. 20–36, Amsterdam, Netherlands, 2016.

[2] J. Lin, C. Gan, and S. Han, “Tsm: temporal shift module for efficient video understanding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, Seoul, Korea, 2019.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “Slowfast networks for video recognition,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, Seoul, Korea, 2019.

[4] D. Kondratyuk, L. Yuan, Y. Li et al., “Movinets: Mobile video networks for efficient video recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16020–16030, Kuala Lumpur, Malaysia, 2021.

[5] X. Wang, X. Xiong, M. Neumann et al., “Attentionnas: Spatiotemporal attention cell search for video classification,” in *European Conference on Computer Vision*, pp. 449–465, Glasgow, US, 2020.

[6] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, “Eco: Efficient convolution operators for tracking,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6638–6646, Honolulu, USA, 2017.

[7] R. Goyal, S. Ebrahimi Kahou, V. Michalski et al., “The something something video database for learning and evaluating visual common sense,” in *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, Venice, Italy, 2017.

[8] D. L. Ruderman, “The statistics of natural images,” *Computation in neural systems*, vol. 5, no. 4, pp. 517–548, 1994.

[9] J. Huang and D. Mumford, “Statistics of natural images and models,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 541–547, Miami, USA, 1999.

[10] E. H. Adelson and J. R. Bergen, “Spatiotemporal energy models for the perception of motion,” *Josa A*, vol. 2, no. 2, pp. 284–299, 1985.

[11] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Twenty-Eighth Annual Conference on Neural Information Processing Systems Conference*, pp. 27–37, Montreal, Canada, 2014.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, Las Vegas, USA, 2016.

[13] A. Klaser, M. Marszałek, and C. Schmid, “A spatio-temporal descriptor based on 3d-gradients,” in *BMVC 19th British Machine Vision Conference*, pp. 275-276, Leeds, UK, 2008.

[14] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*, pp. 65–72, Beijing, China, 2005.

[15] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, Honolulu, USA, 2017.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, Santiago, Chile, 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 25–31, Sierra Nevada, Spain, 2012.

[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, Santiago, Chile, May 2014.

[19] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, Boston, USA, 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, USA, 2016.

[21] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510–1517, 2018.

[22] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision*, pp. 803–818, Munich, Germany, 2018.

[23] J. Donahue, L. Anne Hendricks, S. Guadarrama et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, Boston, USA, 2015.

[24] R. Christoph and F. A. Pinz, "Spatiotemporal residual networks for video action recognition," in *Advances in neural information processing systems*, pp. 3468–3476, mit, Morgan Kaufmann, 2016.

[25] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking CNN models for audio classification," 2020, https://arxiv.org/abs/2007.11154.

[26] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision*, pp. 305–321, Munich, Germany, 2018.

[27] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, Venice, Italy, 2017.

[28] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, USA, 2008.

[29] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*, pp. 428–441, Graz, Austria, 2006.

[30] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558, Sydney, Australia, 2013.

[31] H. Ahmad, H. U. Khan, S. Ali, S. I. Rahman, F. Wahid, and H. Khattak, "Effective video summarization approach based on visual attention," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 71, no. 1, pp. 1427–1442, 2022.

[32] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Combining Global and Local Attention with Positional Encoding for Video Summarization," in *IEEE International Symposium on Multimedia (ISM)*, pp. 226–234, Naples, Italy, 2021.

[33] G. Wu, J. Lin, and C. T. Silva, "IntentVizor: Towards Generic Query Guided Interactive Video Summarization Using Slow-Fast Graph Convolutional Networks," 2021, https://arxiv.org/abs/2109.14834.

[34] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Supervised video summarization via multiple feature sets with parallel attention," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Shenzhen, China, 2021.

[35] M. Mohaiminul Islam and G. Bertasius, "Long movie clip classification with state-space video models," 2022, https://arxiv.org/abs/2204.01692.

[36] Y. Li, C. Y. Wu, H. Fan et al., "Improved multiscale vision transformers for classification and detection," 2021, https://arxiv.org/abs/2112.01526.

[37] B. Nguyen-Thai, V. Le, C. Morgan, N. Badawi, T. Tran, and S. Venkatesh, "A spatio-temporal attention-based model for infant movement assessment from videos," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3911–3920, 2021.

[38] L. Chi, Z. Yuan, Y. Mu, and C. Wang, "Non-local neural networks with grouped bilinear attentional transforms," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11804–11813, Seattle, USA, 2020.

[39] D. Neimark, O. Bar, M. Zohar et al., "Video transformer network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3163–3172, Montreal, Canada, 2021.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, pp. 1–22, Xian, China, 2021.

[41] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding," in *The Thirty-ninth International Conference on Machine Learning*, pp. 1–3, Vienna, Austria, 2021.

[42] Z. Liu, J. Ning, Y. Cao et al., "Video Swin Transformer," 2021, https://arxiv.org/abs/2106.13230.

[43] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViVit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, Montreal, Canada, 2021.

[44] H. Zhang, Y. Hao, and C. W. Ngo, "Token shift transformer for video classification," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 917–925, Chengdu, China, 2021.

[45] H. Fan, B. Xiong, K. Mangalam et al., "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, Montreal, Canada, 2021.

[46] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Advances in neural information processing systems*, pp. 30–41, Long beach, USA, 2017.

[47] J. Xu, L. Zhang, W. Zuo, D. Zhang, and X. Feng, "Patch group based nonlocal self-similarity prior learning for image denoising," in *Proceedings of the IEEE international conference on computer vision*, pp. 244–252, Santiago, Chile, 2015.

[48] A. Diba, M. Fayyaz, V. Sharma et al., "Spatio-temporal channel correlation networks for action classification," in *Proceedings of the European Conference on Computer Vision*, pp. 284–299, Munich, Germany, 2018.

[49] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1430–1439, Salt Lake City, USA, 2018.

[50] M. Zolfaghari, K. Singh, and T. Brox, "Eco: efficient convolutional network for online video understanding," in *Proceedings of the European conference on computer vision*, pp. 695–712, Munich, Germany, 2018.

[51] D. He, F. Li, Q. Zhao, X. Long, Y. Fu, and S. Wen, "Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition," 2018, https://arxiv.org/abs/1806.10319.

[52] Z. Qiu, T. Yao, C. W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12056–12065, Long Beach, USA, 2019.