Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-11403-z

3D convolutional networks with multi-layer-pooling selection fusion for video classification

Zheng-ping Hu[1,2] · Rui-xue Zhang[1] · Yue Qiu[1] · Meng-yao Zhao[1] · Zhe Sun[1,2]

Received: 30 October 2020 / Revised: 27 July 2021 / Accepted: 2 August 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
C3D has been widely used for video representation and understanding. However, it is performed on spatio-temporal contexts in a global view, which often weakens its capacity of learning local representation. To alleviate this problem, a concise and novel multi-layer feature fusion network with the cooperation of local and global views is introduced. For the current network, the global view branch is used to learn the core video semantics, while the local view branch is used to capture the contextual local semantics. Unlike traditional C3D model, the global view branch can only provide the big view branch with the most activated video features from a broader 3D receptive field. Via adding such shallow-view contexts, the local view branch can learn more robust and discriminative spatio-temporal representations for video classification. Thus we propose 3D convolutional networks with multi-layer-pooling selection fusion for video classification, the integrated deep global feature is combined with the information originated from shallow layer of local feature extraction networks, through the space-time pyramid pooling, adaptive pooling and attention pooling three different pooling units, different time–space feature information is obtained, and finally cascaded and used for classification. Experiments on the UCF-101 and HMDB-51 datasets achieve correct classification rate 95.0% and 72.2% respectively. The results show that the proposed 3D convolutional networks with multi-layer-pooling selection fusion has better classification performance.

Keywords Video classification · Mid-level semantics · Pooling operator · Convolutional neural network · Multi-scale feature fusion

✉ Zheng-ping Hu
hzp_ysu@163.com

[1] School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei, China

[2] Hebei Key Laboratory of Information Transmission & Signal Processing, Qinhuangdao, HeBei, China

Published online: 13 August 2021

1 Introduction

Video classification has been a hot research topic in computer vision, owing to its wide application in many fields such as video surveillance, video analysis, video retrieval and video annotation. In the past years, considerable progress has been made for video classification based on different learning model. However, it is difficult to improve the performance of previous methods from a single view feature due to the video background clutters, various inter-class changes and fine-grained intra-class differences may cause classification confusions and misclassification.

In recent years, great progress has been made in understanding human behavior in video activities. Although the Improved Dense trajectory IDT [28] algorithm is the best in the manually based video classification, it lacks the ability to classify and recognize complex videos due to its large computation. Thanks to the significant advancements in computational capabilities of GPU and the availability of large amount of supervised datasets, deep learning not only has gained a lot of attention but also achieved great success on video-based classification. There have been extensive researches that utilize various deep network architectures to explore more effective videos feature extractor. Based on different types of network architectures, three major research directions are explored: 2D CNNs, 3D CNNs and hybrid networks. As the most popular 2D network architecture, two-stream CNN achieves good performance by extracting spatial and temporal information separately from RGB images and optical flows. However, 2D CNNs couldn't capture spatial–temporal information simultaneously. More importantly, it demonstrates that the 3D ConvNets are better at exploiting the spatiotemporal information of video actions than 2D ConvNets and unsupervised methods. 3D CNNs overcome this weakness by introducing the 1D temporal convolution operator, although 3D CNNs need large-scale datasets in training and will lead to too many model parameters as well as too heavy computation cost, they may bring more satisfying classification results. Deep 3-dimensional convolutional networks (3D ConvNets) trained on large scale video datasets have achieved promising results on video classification. Recently Hybrid networks [1, 13, 18, 31, 32] combine such as 3D CNNs and GNN(Graph Neural Networks), C3D and STIPs Attention Model,C3D and object-level semantics are used to improve their performance.

However, for the video representation building based on 3D ConvNets, there still remains some challenges to be solved and one of them is the vulnerability to the large intra-class variability of content. As the intra-class variation in static images is not as large as that in video actions. Early approach has reflected the limitations of data argumentation techniques such as spatial and temporal jittering and different video sample rate. So to alleviate the problem, we need better models to learn robust video representation for video classification. Such as in ref [4] proposed to combine the deep spatiotemporal features extracted by 3D convolutions and a spatiotemporal pyramid pooling (STPP) layer to learn the video representation. Common methods for temporal feature aggregation include simple averaging or maximum pooling as well as more sophisticated pooling techniques. However, these techniques may be suboptimal. Indeed, simple techniques such as average or maximum pooling may become inaccurate for global feature. It has been shown that integrating VLAD as a differentiable module in a neural network can improve the aggregated representation for the task of vision [2]. And this has motivated us to integrate the C3D architecture with the median and last layer by replacing the last layers in C3D with different pooling operator.

Our main contributes can be summarized as follows:

(1) We propose a novel pooling strategy to perform video classification with the cooperation of med-level and last semantics information. To the best of our knowledge, our model is among the first to utilize different layer and different pooling operator together for realizing video classification.
(2) By carefully designing the median last node and different pooling module in the C3D network, our proposed method can jointly enjoy the merits of global representative video content and discriminative refinement video information.
(3) We evaluate the proposed framework in two standard video datasets, UCF-101 and HMDB-51. Extensive experimental results demonstrate the effectiveness of proposed method against state-of-the-art methods.

The rest of this paper is organized as follows: In Sect. 2, some related work on video classification methods based on deep learning is discussed. Then we present details of the proposed approach in Sect. 3. After that, Experimental results and comparisons are provided and the performance of our approach on two public benchmark datasets is evaluated in Sect. 4. Finally, In Sect. 5, conclusion is given for the video classification model.

2 Related works

Most early methods mainly depend on a number of hand-crafted feature descriptors. The typical ones are the improved dense trajectory (IDT), motion energy image and motion history image, histogram of oriented gradients (HOG). However, hand-crafted approaches generally capture the local information and cannot learn the discriminative features automatically from videos; therefore, they may lack the ability to achieve superior performance for complex or similar video. Different from early methods, the powerful modeling ability of deep convolutional neural network has gained a lot of attention and achieved great success on video classification.

Video classification has attracted intensive research interests in recent years. In this section, we review the previous work related to our method from feature extraction and aggregation, feature extraction include three aspects of time model-based 2D CNN, two-stream network and spatio-temporal network model C3D.

(1) **Time model-based 2D CNN.** Karpathy et al. [17] divided the whole video into multiple fixed-length segments, each frame image was extracted separately, and a variety of fusion methods were designed to fuse all segment features, but the classification accuracy isn't satisfying. From the perspective of utilizing the temporal relationship between video frames, the literature [6] introduced the LRCN (Long-term Recurrent Convolutional Networks). 2D CNN is used to extract spatial features from different frames, and then the LSTM (Long-Short Term Memory) time model is used to encode the relationship between the spatial features of the sequence representation to model the time information, but the LSTM performs non-adaptive the local time relationship and it is harder to train. From the perspective of human behavior understanding, Du et al. [8] proposed RPAN (Recurrent Pose-Attention Network), and introduced the attitude attention mechanism to assist video classification based on Conv-LSTM [5]. To mimic the feedforward and feedback connections that mimic the brain, Shi et al. [22] proposed a deep network shuttle Net. Unlike traditional RNNs, all GRUs in shuttle Net are loop-connected, sharing information between multiple paths, and finally adopting note that the mechanism chooses the best information flow path. In order to better describe the characteristics of time domain

information, Lin et al. [19] and others proposed TSM time channel shift network, which stores and partially exchanges the feature information extracted from adjacent video frames through 2D CNN to achieve the effect of channel shift. Therefore, abundant time-domain information is obtained through channel shift model, and offline classification mode and online classification mode are formed. To adaptively judge the temporal and spatial modeling of distinguishing features, Sudhakaran et al. [24] proposed a GSN network model. The input features are divided into two features, one is used for channel shift to extract temporal information, and another is used as spatial information and finally fused with temporal features. Through training, the network could use gating unit to adaptively select temporal and spatial modeling.

(2) **Two-stream network based on multiple input modes.** Two-stream CNN framework is the most representative architecture in 2D CNNs, where RGB images and optical flows are fed separately into a spatial stream network and a temporal stream network to extract the appearance and motion information. Based on this idea, Simonyan et al. thought that RGB input cannot effectively present motion information to the CNN, and proposed to use dense optical streams to represent the temporal components of the video [23], use RGB and stack optical flow frames, respectively. As for appearance and motion information, the two-stream combination significantly improves the accuracy of video classification, indicating the importance of temporal motion information for video classification. In ref [15], three independent CNN models were used to extract spatial, temporal and audio features on static frames, stacked optical flow images, audio spectra were used to fuse three features. However, these methods usually take global features as the starting point and ignore the local details of behavior changes, which may lead to the algorithm being sensitive to external environment such as occlusion and illumination changes. To solve above problems, Zhu et al. [34] identified different actions by fusing local global spatial features and temporal features of video frames, and weighted the global spatial features with visual attention. From the perspective of video sparse sampling, Wang et al. [30] proposed TSN (Temporal segment networks). The sparse sampling strategy was used to decompose the entire video data and RGB difference and warped optical flow fields were included. Different from the above time stream networks, Carreira and Zisserman proposed I3D to expand the 2D network in the two-stream network to 3D [3]. Pre-training was performed on the Kinetics dataset, and excellent classification results are performed on UCF101 dataset.

It is known that optical flow detection can extract the context information of video frames, which leading optical flow information is a good classification feature of video representation. However, the method relies on the traditional optical flow algorithm to model the time information. This two-stage method has high calculation cost and high storage requirement. So the use of neural networks to obtain optical flow characteristics is popular, such as flownet2.0 proposed by Ilg et al. [14] and the enhanced version of flownet [7]. Overall, the speed cost is reduced and the performance is greatly improved. Input RGB and moving optical flow image information into CNN network, and output the result through information fusion. Recently, Piergiovanni and Ryoo [20] considered using the convolutional flow representation layer to learn the motion representation information, and stacking the representation layer to form FCF(flow-conv-flow), so as to optimize of network speed and performance.

(3) **Spatio-temporal network**, which uses stacked RGB video frames as input to learn temporal and spatial features from consecutive video frames. Tran et al. [25] proposed that the depth 3D CNN network (C3D) extracts spatio-temporal features from video, and performs space–time convolution and pooling in all layers to improve the accuracy of classification. The disadvantage is that the computational cost is high. To improve the 3D

convolution, Qiu et al. [21] proposed P3D, which approximates the original $3\times3\times3$ convolution with a $1\times3\times3$ spatial direction convolution and a $3\times1\times1$ time direction convolution. P3D optimizes C3D in terms of parameter number and running speed. Introducing a novel multi-scale deep alternative neural network, literature [29] combined the advantages of convolutional neural networks and recurrent neural networks to collect rich video context information. Different from the above frame sampling strategy, Varol et al. [27] proposed a long-term temporal convolutions (LTC) network to increase the time range of video frames to improve accuracy. Ref [16] firstly preserved the overall time dynamics presented in the original video by selecting partial frames as alternative videos, and used 3D CNN to learn the overall temporal characteristics of the alternate video. Taking long-term content into account and enabling fast per-video processing at the same time, ECO in reference [35] proposed a network architecture which is based on merging long-term content already in the network rather than in a post-hoc fusion. Although spatio-temporal networks can solve the problem of space–time modeling, the classification accuracy is lower than that of two-stream networks. In order to maintain the performance of 3D convolution and reduce the computational complexity, existed study includes two directions: (1) Improved C3D for reducing the computational complexity, for example, Tran et al. [26] proposed a channel separable network which only contains 1 * 1 * 1 ordinary 3D convolution and k * k * k deep convolution. 1 * 1 * 1 ordinary 3D convolution is used to deal with channel interaction, and k* k * k convolution is used to deal with local interaction. (2) Integrated C3D architecture with the proposed layer processing scheme. Video features are typically extracted from individual frames or short video clips. The remaining question is: how to aggregate video features over the entire and potentially long video? One way to achieve this is to employ recurrent neural networks. Other methods capture only the distribution of features in the video, not explicitly modeling their temporal ordering. The simplest form of this approach is the average or maximum pooling of video features over time. Here we extend this work by aggregating different output layer and different pooling operator.

The spatio-temporal network above can learn features through three-dimensional convolution distributed over the entire network, and the operation efficiency is high. However, as the depth of the network deepens, the features extracted by the model is abstract, so that the model finally ignores the median details existed in the data. At the same time, some features extracted from median layers may contain information that the depth features do not have. Thus, we present a multi-level feature fusion 3D convolution network to extract the multi-level features of video frame segments to fully take advantage of important visual cues, and finally fuse the results of different feature levels. Experimental results on UCF-101 and HMDB-51 databases show that the proposed method can achieve better performance.

3 The proposed approach

3.1 Network architecture

The deep neural network model of multi-level feature fusion designed in this paper is shown in Fig. 1. Firstly, continuous 32-frame video frames are used in the video, and the characteristics of the input data are extracted by using the 3D ResNeXt-101 network, and median layer is obtained through the pooling unit. In addition, the feature is compressed

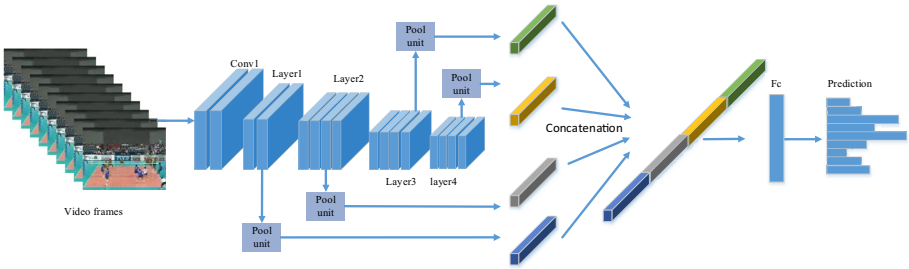


Fig. 1 Schematic of the 3D convolutional networks with multi-layer-pooling selection fusion. Different from C3D architecture follows with classification, we adopt different layer output with different operation fusion to mine abstract convolutional hybrid features, which can be fused with the features from different views to make better use of the complementary characteristics of C3D. The result of classification is obtained finally by multiple types of features after concatenation

and formed into a vector, and the bottom feature is finally cascaded with the advanced feature and classified by the fully connected layer.

3.2 Pooling unit

The pooling unit mainly compresses the feature map of each layer into feature vectors. Here, three different pooling schemes are used, which include space–time pyramid pooling, attention pooling, and adaptive pooling. The following three pooling schemes are introduced.

3.2.1 Space–time pyramid pooling

The original space pyramid model [10] is firstly used for object detection, because it can generate fixed output features with arbitrary input. Multiple corresponding regions can correspond to objects of different scales, thus improving the robustness of the network to object scaling. In our model, the spatial pyramid is extended to the Spatio-temporal Pyramid Pooling (STPP), and each level of the pyramid could use different window sizes and moving steps as a pooling unit. Through the pyramid pooling, feature maps are extracted from different angles, and then the extracted features are aggregated, which can enhance the robustness of video understanding and improve the classification performance. The space–time pyramid pooling unit is shown in Fig. 2, to reduce the number of parameters, we uses a two-level space–time pyramid pooling. Where N is the number of feature channels, and the features obtained from the network by the pooling unit are respectively divided into $8\times N$ and $1\times N$, and then extended to a $9\times N$ vector are sent to the fully connected layer.

The maximum pooling layer is used in the space–time pyramid. The corresponding convolution kernel K_i , step size S_i and filling quantity P_i in each level pyramid are respectively shown in Eqs. (1), (2) and (3). Among them,h_i represents the size of the feature matrix in the dimension, n represents the n-th layer space–time pyramid, ceil means rounded up, and floor means rounded down.

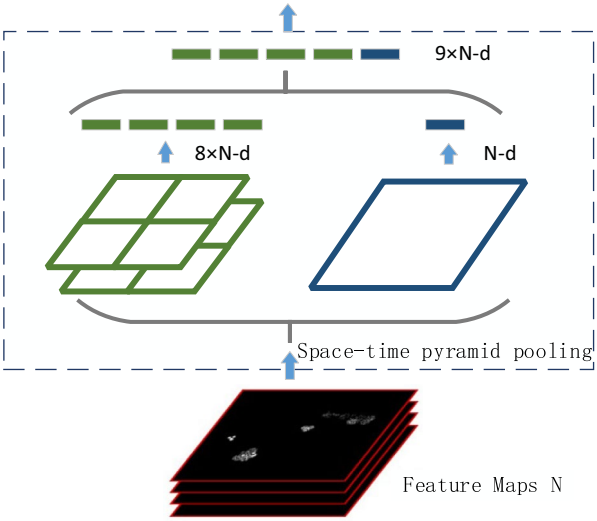


Fig. 2 Schematic of the space-time pyramid pooling unit. In the Spatio-temporal Pyramid Pooling (STPP), and each level of the pyramid could use different window sizes and moving steps as a pooling unit. Through the pyramid pooling, feature maps are extracted from different angles, and then the extracted features are aggregated

$$K_i = \lceil \frac{h_i}{n} \rceil = ceil\left(\frac{h_i}{n}\right) \qquad i = 1, 2, 3 \tag{1}$$

$$S_i = \lceil \frac{h_i}{n} \rceil = floor\left(\frac{h_i}{n}\right) \qquad i = 1, 2, 3 \tag{2}$$

$$P_i = \lfloor \frac{K_i * n - h_i + 1}{2} \rfloor = floor\left(\frac{K_i * n - h_i + 1}{2}\right) \qquad i = 1, 2, 3 \tag{3}$$

3.2.2 Adaptive pooling

The particularity of adaptive pooling is that the size of the output tensor is given. The adaptive pooling is used to convert the feature matrix of c×t×h×w into a matrix of c×1×1×1 size along the space–time dimension. Each three-dimensional feature channel is further transformed into a real number with a global receptive field by a compression operation. And the output dimension matches the input feature channel number, which represents the global distribution of the response on the feature channel, and the layer close to the input can also obtain the global receptive field. Finally, according to the input tensor input_size and the output tensor output_size, the parameters of the adaptive pooling layer can be obtained by formula (4), where the floor is rounded down.

$$stride = floor(input\_size / output\_size) kernel = input\_size - (output\_size - 1) * stridepadding = 0 \tag{4}$$

3.2.3 Attention pooling

SeNet (Squeeze and Excitation Network) [12] adopts a novel feature calibration strategy to automatically acquire the importance of each feature channel through network learning,

and then explicitly models the interdependence between feature channels. Inspired by SeNet, we use an improved attention-pooling unit (Se-pooling), as shown in Fig. 3. Based on the adaptive pooling to compress the spatio-temporal dimensional features, a Bottleneck structure is constructed by two fully connected layers to model the correlation between the channels. First, it reduces the feature dimension to 1/16 of the input, and then goes back to the initial dimension through the fully connected layer. Moreover, adding the fully connected layer can establish a complex correlation between the channels. The output of the fully connected layer is regarded as the weight of each channel. Finally, the weighting is combined with the global information of each channel by a multiplication operation.

3.3 Classification network

In order to verify the applicability of the proposed method, experiments were performed on ResNet-101 [11], ResNeXt-101 [9] and DenseNet-121 [33], respectively. All networks above adopt the VGG/ResNet repetitive layer strategy to increase the classification accuracy by increasing the width and depth of the network, and reducing the complexity of the network. The basic unit structure is shown in Fig. 4, where F denotes the number of feature maps, and each basic unit includes a Batch-Normalization layer and a ReLU layer.

However, compared to the traditional ResNet, ResNeXt-101 adds the first layer and the third layer convolution of size $1\times1\times1$ to the network, controls the number of convolution

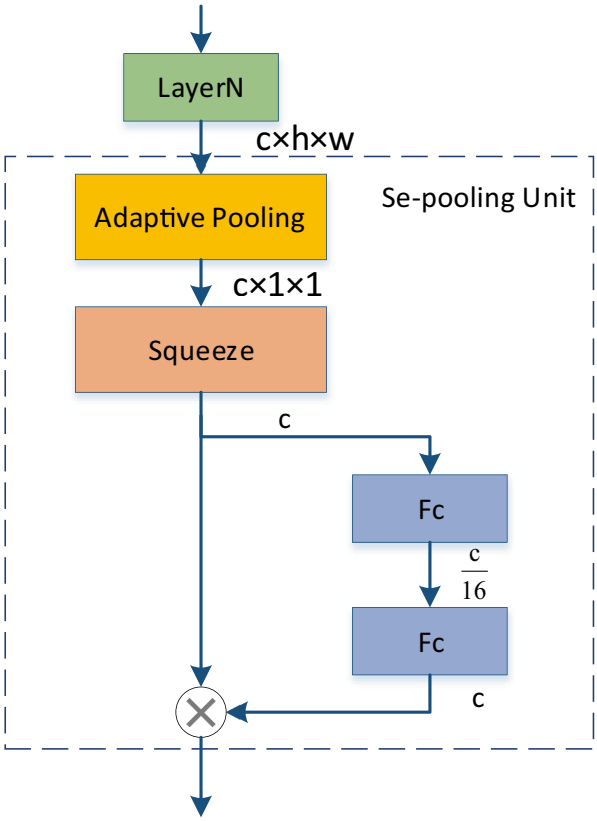
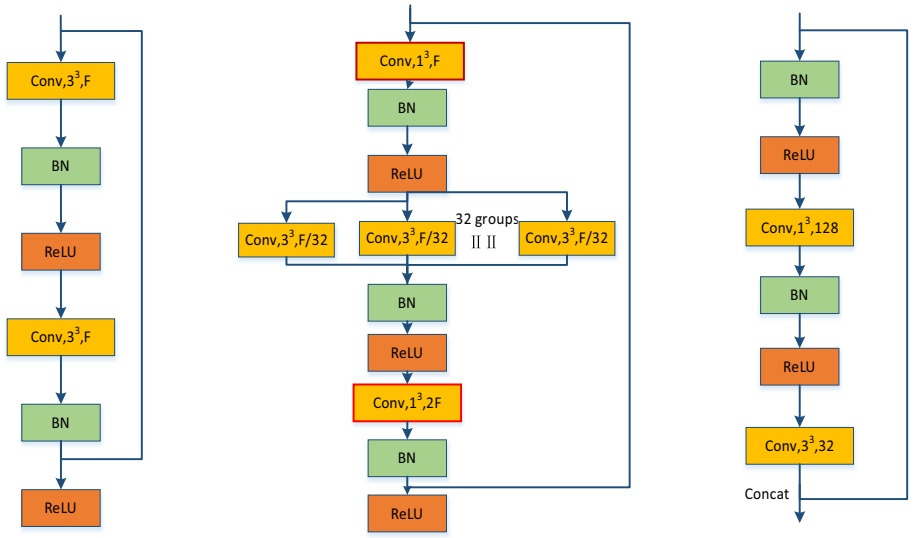


Fig. 3 Schematic of the Se-pooling unit. Se-pooling unit adopts a novel feature calibration strategy to automatically acquire the importance of each feature channel through network learning, and then explicitly models the interdependence between feature channels. Based on the adaptive pooling to compress the spatio-temporal dimensional features, a Bottleneck structure is constructed by two fully connected layers to model the correlation between the channels



(a) ResNet **(b)** ResNeXt **(c)** DenseNet

Fig. 4 Unit structure of (**a**) ResNet-101, (**b**) ResNeXt-101and (**c**) DenseNet-121, respectively. All networks here adopt the VGG/ResNet repetitive layer strategy by increasing the width and depth of the network, and reducing the complexity of the network. The basic unit structure includes a Batch-Normalization layer and a ReLU layer

kernels in the middle layer, and reduces the number of parameters of the network. In the middle layer convolution, the idea of packet convolution is used [22], which divides the feature mapping into groups, which reduces the difficulty of training on the network and improves the network performance. And the number of convolution groups introduced in the experiment is 32 groups.

DenseNet connects all layers in the network, so that each layer in the network is associated with all previous layers. The unit module of DenseNet is shown in Fig. 4. Although it is similar to the unit module component unit of ResNet, the actual sequence and parameters are quite different. Through feature reuse and bypass settings, DenseNet not only greatly reduces the number of parameters of the network, but also alleviates the problem of gradient disappearance to a certain extent, and maximizes the information flow between all layers in the network. At the same time, some features extracted from median layers may still be directly used by deeper layers. The specific network architecture is shown in Fig. 5. For feature multiplexing, cascading operations on feature dimensions are used in cross-layer joins rather than pixel-by-pixel add operations.

4 Experiments and results

This section conducted an experimental evaluation to verify the validity of the model. First, the data set and experiment settings are explained clearly and secondly the results of the model are analyzed. At last, the performance of the method and the most advanced method in video classification are compared.

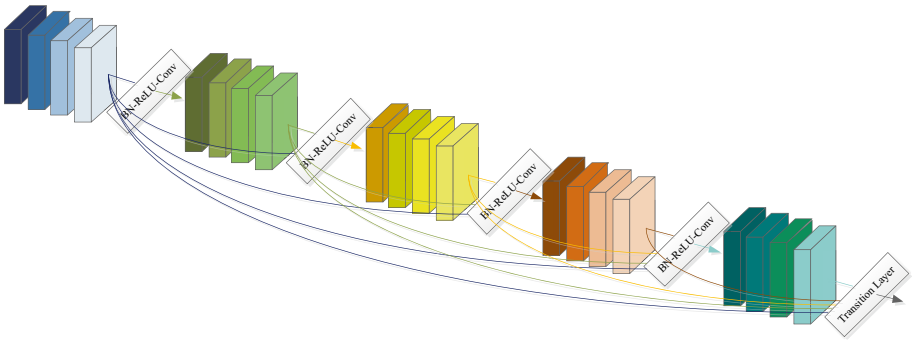


Fig. 5 Schematic of the DenseNet structure. DenseNet connects all layers in the network, so that each layer in the network is associated with all previous layers. Although it is similar to the unit module component unit of ResNet, the actual sequence and parameters are quite different

4.1 Datasets

We evaluate the compressed video classification method on the UCF-101 and HMDB-51 datasets. The UCF-101 dataset consists of 101 categories, including 13,320 video clips. The average duration of each video is about 7 s. And the video includes human-object interaction, human motion, human-to-human interaction, instrumental performance, and sports. However, the HMDB-51 dataset consists of 51 categories, including 6849 video clips. The video content is similar to UCF-101, including facial movements, human movements, body movements, and interaction with objects. Both datasets are divided into three training/test set partitions (70% training and 30% testing) with different contents. In addition, the final classification result of proposed is the average performance of the three test divisions.

4.2 Experiments setting

During the training process, the time points in the video are selected from the training data by uniform sampling, and then 32 consecutive video segments are generated around the selected time point. If the video is shorter than 32 frames, multiple cycles are performed as needed. Next, randomly select a spatial position from the four corners or the center of the video frame, and select the spatial scale of the sample for multi-scale cropping, and adjust the sample to 112×112 size in position, scale and aspect ratio. Each sample is horizontally flipped with a 50% probability. The final input data size is 3 channels $\times$ 32 frames $\times$ 112 pixels $\times$ 112 pixels. The network is trained with the stochastic gradient descent method with momentum. The video frame and video tag are inputted into the network, and the pre-training weights on the Kinetics dataset are fine-tuned, the learning rate is set to 0.001, and the verification loss is saturated and divided by 10. In addition, the momentum is set to 0.9 and the weight attenuation is set to 1e-5. In the test phase, the input segment is generated using a sliding window approach, and the final result is the feature average of all video segments (ie, video-level features).

Table 1 Comparison of classification performance of different pooling units plus different layers

Different methods	UCF-101/%	HMDB-51/%	different layers selection
ResNet-101	87.15	59.23	Last layer
ResNet-101 + STPP	87.67	61.49	Last layer
ResNet-101 + Adaptive Pooling	88.13	63.08	Last layer
ResNet-101 + Se-pooling	88.21	63.34	Last layer
ResNet-101	88.26	62.59	Last two layer
ResNet-101 + STPP	88.51	63.85	Last two layer
ResNet-101 + Adaptive Pooling	89.06	64.77	Last two layer
ResNet-101 + Se-pooling	88.90	64.85	Last two layer
ResNet-101	89.74	65.86	Last three layer
ResNet-101 + STPP	89.93	65.08	Last three layer
ResNet-101 + Adaptive Pooling	90.80	66.64	Last three layer
ResNet-101 + Se-pooling	89.90	66.25	Last three layer
ResNet-101	89.67	65.14	Last four layer
ResNet-101 + STPP	89.90	65.09	Last four layer
ResNet-101 + Adaptive Pooling	90.24	65.85	Last four layer
ResNet-101 + Se-pooling	88.86	66.39	Last four layer

4.3 Results and analysis

4.3.1 Investigations of the parameters

First, we compare the performance of the three pooling units proposed for proposed method, and the experimental verification is carried out on the split1 of UCF-101 and HMDB-51. At this time, the main classification network uses ResNet-101 with different pooling unit and different layers selection, and the corresponding accuracy are shown in Table 1.

It can be seen in Table 1 that the addition of the pooling unit has a certain improvement in the classification. Selection different pooling unit and different layers all will lead to different scores. When the selection layers number add to three, the proposed method will obtain best performance. The pyramid pooling is too large due to the multi-level pooling structure, which makes network training difficult, so it is less effective on the smaller HMDB-51 dataset. Compared to the three pooling units, the adaptive pooling improves about 1% on both datasets obtains the highest accuracy, which is better than the other two pooling units, so adaptively Pooling and three layers selection is the final pooling unit.

Then, in order to further verify the applicability of the proposed method and compare the performance of different network extraction features, the comparison networks (ResNet, DenseNet, and ResNeXt) are compared on the basis of adaptive pooling unit and three layers selection, on split1 of UCF-101 and HMDB-51. The accuracy rates are shown in Table 2.

Table 2 Performance comparison of different classification networks

Network model	UCF-101/%	HMDB-51/%
ResNet-101	90.80	66.64
DenseNet-121	90.53	62.13
ResNeXt-101	92.28	67.82

Table 3 Comparisons of frame segments under different lengths

Network model	UCF-101/%	HMDB-51/%	GFLOPs
ResNeXt-101(16f)	91.49	65.79	9.68
ResNeXt-101(32f)	92.28	67.82	19.36
ResNeXt-101(64f)	95.08	74.10	38.71

Finally, in most video classification tasks, the video is processed into a frame segment input network, and the length of the frame segment also has a great influence on the classification. The accuracy of different length frame fragments and GFLOPs (Giga Floating Point Operations) pairs are shown in Table 3. When the video length is long, as it provides more information for classification, we could obtain better performance while the computation loads add naturally.

4.3.2 Experimental results

The accuracies of different division model based on multi-layer feature fusion in the three classification methods of UCF-101 and HMDB-51 datasets are shown in Table 4. And the final classification result is the average of the accuracy on the two datasets introduced above.

4.4 Comparison of different methods

In order to further analyze the performance in video classification, we compare the proposed model with some existing mainstream classification algorithms. Table 5 shows the results of the algorithms on the UCF-101 and HMDB-51 datasets. All accuracies are the average of the three division methods, and the dimension represents the dimensions of the convolution kernel.

Here, we fine-tune the network weights pre-trained on the 64-frame Kinetics dataset. The accuracy increases to 95.0% and 72.2% with the input still 32 frames, indicating the importance of pre-training weights in video classification tasks. It can be seen from Table 5 that compared with other video classification methods, the video classification method implemented in this paper can achieve better classification effect. Compared with iDT, iDT is the best traditional method for classification except for deep learning, and the accuracy of this method is 8.6% higher than iDT. Compared with other deep learning methods, such as Two-stream CNN, TSN, C3D, pseudo-3D residual network P3D, STMN + iDT, STIPs Attention Model the results of this paper are superior to most methods in both datasets, which proves the effectiveness of the proposed method in this paper.

Table 4 Classification accuracies on the UCF-101 and HMDB-51 datasets

Division method	UCF-101/%	HMDB-51/%
Split01	92.28	67.82
Split02	93.03	66.32
Split03	92.07	66.06
Avg	92.46	66.73

Table 5 Comparisons of our model and existing mainstream algorithms

Method	Dimension	UCF-101/%	HMDB-51/%
iDT [28]	-	86.4	61.7
Two-stream CNN	2D	88.0	59.4
TSN [17]	2D	94.2	69.4
C3D [8]	3D	82.3	-
P3D [5]	3D	88.6	-
Trajectory Pooling [14]	2D	92.1	65.6
STMN + iDT [18] (TIP 2019]	2D	94.5	70.2
STIPs Attention Model [31] (2020)	3D	94.8	71.5
Ours	3D	92.5	66.7
Ours(64f-pretrain)	3D	**95.0**	**72.2**

5 Conclusions

In this paper, we propose a 3D convolutional networks with multi-level-pooling selection fusion for improving the video classification performance, which is based on different CNNs pooling operator and different layer-level feature to focus on the informative information of video. The reason is that as the depth of the network deepens, some features extracted from median layers may contain information that the depth features do not have. To select more representative video representation features, we compare the three pooling units proposed in this paper, select adaptive pooling unit to obtain different levels of feature information, and finally cascade and classify different levels of hybrid features. Experimental results have verified the effectiveness of our method on capturing the descriptive and discriminative information in different layers. In the future, we will extend our work to two aspects. First, we will reduce more irrelevant points for accurate location of salient regions. Second, hybrid 3D CNNs will be applied to explore their capability on videos.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grants 61771420 and and 62001413, the Natural Science Foundation of Hebei Province under Grants F2020203064, as well as the China Postdoctoral Science Foundation Grant 2018M641674 and the Doctoral Foundation of Yanshan University under Grants BL18033. In this paper, we utilize the public video database and thank the provider of the databases.

References

1. Ali A, Zhu Y, Chen Q et al (2019) Leveraging spatio-temporal patterns for predicting citywide traffic crowd flows using deep hybrid neural networks, in: Proc Intern Conf Parallel Distributed Syst 125–132
2. Arandjelovic R, Gronat P, Torii A et al (2018) NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Trans Pattern Anal Mach Intell 40:1437–1451
3. Carreira J, Zisserman A (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, in: 2017 IEEE Conf Comp Vision Pattern Recogn 4724–4733
4. Cheng C, Lv P, Su B (2018) Spatiotemporal pyramid pooling in 3D convolutional neural networks for action recognition, in: Intern Conf Image Process 3468–3472
5. Donahue J, Hendricks L, Guadarrama S et al (2015) Long-term recurrent convolutional networks for visual recognition and description, in: IEEE Conf Computer Vision Pattern Recogn 2625–2634
6. Donahue J, Hendricks L, Rohrbach M et al (2017) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Trans Pattern Analysis Machine Intell (39):677–691
7. Dosovitskiy A, Fischer P, Ilg E et al (2015) FlowNet: learning optical flow with convolutional networks, in: IEEE Intern Confe Comp Vision 2758–2766

8. Du W, Wang Y, Qiao Y (2017) RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos. IEEE Intern Conf Comp Vision 2017:3745–3754
9. Hara K, Kataoka H, Satoh Y (2018) Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?, in. IEEE/CVF Conf Comp Vision Pattern Recogn 2018:6546–6555
10. He K, Zhang X, Ren S et al (2015) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans Pattern Anal Mach Intell 37:1904–1916
11. He K, Zhang X, Ren S et al (2016) Deep Residual Learning for Image Recognition, in. IEEE Conf Comp Vision Pattern Recogn 770–778
12. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks, in. IEEE/CVF Conf Comp Vision Pattern Recogn 7132–7141
13. Hu Y, Gao J, Xu C (2020) Learning Dual-Pooling Graph Neural Networks for Few-shot Video Classification, IEEE Trans Multimedia (Early Access). https://doi.org/10.1109/TMM.2020.3039329
14. Ilg E, Mayer N, Saikia T et al (2017) FlowNet 2.0: evolution of optical flow estimation with deep networks, 2017 IEEE Conf Comp Vision Pattern Recogn 1647–1655
15. Jiang Y, Wu Z, Tang J et al (2018) Modeling Multimodal Clues in a Hybrid Deep Learning Framework for Video Classification. IEEE Trans Multimedia 20:3137–3147
16. Jing L, Yang X, Tian Y (2018) Video you only look once: Overall temporal convolutions for action recognition. J Vis Commun Image Represent 52:58–65
17. Karpathy A, Toderici G, Shetty S (2014) Large-Scale Video Classification with Convolutional Neural Networks, in: IEEE Conf Comp Vision Pattern Recogn 1725–1732
18. Li C, Zhang B, Chen C et al (2019) Deep manifold structure transfer for action recognition. IEEE Trans Image Process 28:4646–4658
19. Lin J, Gan C, Han D (2019) TSM: Temporal Shift Module for Efficient Video Understanding, in: Intern Conf Comp Vision 7082–7092
20. Piergiovanni A, Ryoo M (2019) Representation Flow for Action Recognition, in. IEEE/CVF Conf Comp Vision Pattern Recogn 9937–9945
21. Qiu Z, Yao T, Mei T (2017) Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, in. IEEE Intern Conf Comp Vision 5534–5542
22. Shi Y, Tian Y, Wang Y et al (2017) Learning Long-Term Dependencies for Action Recognition with a Biologically-Inspired Deep Network, in: IEEE Inter Conf Comp Vision 716–725
23. Simonyan K, Zisserman A (2014) Two-stream convolutional networksfor action recognition in videos. Proc 27th Intern Confer Neural Inform Process Syst 568–576
24. Sudhakaran S, Escalera S, Lanz O (2020) Gate-Shift Networks for Video Action Recognition, in. IEEE/CVF Conf Comp Vision Pattern Recogn 1099–1108
25. Tran D, Bourdev L, Fergus R et al (2015) Learning Spatiotemporal Features with 3D Convolutional Networks, in. IEEE Intern Conf Comp Vision 4489–4497
26. Tran D, Wang H, Feiszli M et al (2019) Video Classification With Channel-Separated Convolutional Networks, in. IEEE/CVF Intern Conf Comp Vision 25551–5560
27. Varol G, Laptev I, Schmid C (2018) Long-Term Temporal Convolutions for Action Recognition. IEEE Trans Pattern Anal Mach Intell 40:1510–1517
28. Wang H, Schmid C (2013) Action recognition with improved trajectories, in: Intern Conf Comp Vision 3551–3558
29. Wang J, Wang W, Gao W (2018) Multiscale Deep Alternative Neural Network for Large-Scale Video Classification IEEE Transact Multimedia 20:2578–2592
30. Wang L, Xiong Y, Wang Z et al (2020) Temporal segment networks: towards good practices for deep action recognition, in: Euro Conf Comp Vision 20–36
31. Wu H, Ma X, Li Y (2020) Convolutional Networks With Channel and STIPs Attention Model for Action Recognition in Videos. IEEE Trans Multimedia 22:2293–2306
32. Zhang J, Mei K, Zheng Y et al (2019) Exploiting Mid-Level Semantics for Large-Scale Complex Video Classification. IEEE Trans Multimedia 21:2518–2530
33. Zhao S, Liu Y, Han Y et al (2018) Pooling the Convolutional Layers in Deep ConvNets for Video Action Recognition. IEEE Trans Circuits Syst Video Technol 28:1839–1849
34. Zhu S, Fang Z, Wang Y et al (2019) Multimodal activity recognition with local block CNN and attention-based spatial weighted CNN. J Vis Commun Image Represent 60:38–43
35. Zolfaghari M, Singh K, Brox T (2018) Eco: Efficient convolutional network for online video understanding, Proc Euro Con Comp Vision 713–730

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.