International Journal of Automation and Computing
www.ijac.net

18(5), October 2021, 718-730
DOI: 10.1007/s11633-021-1289-9

STRNet: Triple-stream Spatiotemporal Relation Network for Action Recognition

Zhi-Wei Xu[1,2] Xiao-Jun Wu[1,2] Josef Kittler[3]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China

[2] Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Wuxi 214122, China

[3] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

Abstract: Learning comprehensive spatiotemporal features is crucial for human action recognition. Existing methods tend to model the spatiotemporal feature blocks in an integrate-separate-integrate form, such as appearance-and-relation network (ARTNet) and spatiotemporal and motion network (STM). However, with blocks stacking up, the rear part of the network has poor interpretability. To avoid this problem, we propose a novel architecture called spatial temporal relation network (STRNet), which can learn explicit information of appearance, motion and especially the temporal relation information. Specifically, our STRNet is constructed by three branches, which separates the features into 1) appearance pathway, to obtain spatial semantics, 2) motion pathway, to reinforce the spatiotemporal feature representation, and 3) relation pathway, to focus on capturing temporal relation details of successive frames and to explore long-term representation dependency. In addition, our STRNet does not just simply merge the multi-branch information, but we apply a flexible and effective strategy to fuse the complementary information from multiple pathways. We evaluate our network on four major action recognition benchmarks: Kinetics-400, UCF-101, HMDB-51, and Something-Something v1, demonstrating that the performance of our STRNet achieves the state-of-the-art result on the UCF-101 and HMDB-51 datasets, as well as a comparable accuracy with the state-of-the-art method on Something-Something v1 and Kinetics-400.

Keywords: Action recognition, spatiotemporal relation, multi-branch fusion, long-term representation, video classification.

Citation: Z. W. Xu, X. J. Wu, J. Kittler. STRNet: Triple-stream spatiotemporal relation network for action recognition. *International Journal of Automation and Computing*, vol.18, no.5, pp.718–730, 2021. http://doi.org/10.1007/s11633-021-1289-9

1 Introduction

Pattern recognition[1] has been attracting a lot of interest since the advent of the computer. The booming development of machine learning[2] and deep learning[3] has injected new vitality to this subject. In the field of computer vision, deep learning methods play an indispensable role in most visual tasks, such as image classification[4–6], object detection[7, 8], semantic segmentation[9, 10], video classification[11–22], etc. Along with the emergence of diverse deep architectures, AlexNet[4], visual geometry group (VGG)[23], residual network (ResNet)[5], densely connected convolutional network (DenseNet)[24], the accuracy of recognition tasks is continually pushed to new heights. However, for video-based action recognition tasks, it is different from a single static picture that contains only spatial information. The temporal relation[15] provided by a video is even more significant for recognition. To exploit the information in the time dimension, the traditional method with hand-crafted features[25–27] achieved a considerable accuracy in early years, at the expense of consuming a lot of effort in video data processing. More recent approaches focus on deep networks, which can be grouped into four categories: 1) two-stream convolutional neural networks (CNN)[15, 28], 2) 3D CNN[17, 29], 3) 2D CNN with temporal models such as long short-term memory (LSTM)[30, 31], 4) skeleton-based architecture[32–34]. Two-stream CNN is a popular method which can make full use of both appearance and motion information. It performs well on action recognition tasks.

However, this method requires additional processing to extract optical-flow[35] information in advance. The two data streams are trained separately resulting in the training process taking at least twice as long compared to systems without optical-flow computation. To avoid this drawback, 3D CNN architectures have been proposed to obtain spatiotemporal features directly from sequential RGB frames. But they are computationally expensive, and their performance is worse than the two-stream net-

Research Article
Manuscript received October 30, 2020; accepted February 5, 2021; published online March 23, 2021
Recommended by Associate Editor De Xu
Colored figures are available in the online version at https://link.springer.com/journal/11633

works. Some researchers find that 2D CNNs already perform well on learning spatial features, so they attempt to develop other methods to capture temporal information. For example, recurrent neural network (RNN) is typically applied to LSTM[36], which can remember and learn a long span on temporal feature representation over several input frames. Skeleton-based methods have gradually become more prominent in recent years. They focus attention on specific human action, posture and gesture, but these approaches lack the capacity of learning appearance information and the interaction between human and object. Meanwhile, the video datasets for action recognition, such as Kinetics[11], atomic visual actions dataset (AVA)[37] and Something-Something[38], are becoming larger and larger. All of them contain complex scenes and action information.

It is notable that the architectures of most state-of-the-art methods consist of building blocks in which the spatial and temporal features can be learned simultaneously, such as ARTNet[39] and STM[21]. However, with the block stacking up, the latter blocks cannot learn the spatiotemporal and motion features explicitly. For one thing, the pixel-level addition and 1×1 convolution operation make the information abundant but tangled. For another, with the iteration of the integrate-separate-integrate form, the rear part of the network has a low interpretability for the spatiotemporal or motion relation feature representation. Based on this observation, we propose a straightforward end-to-end architecture named the spatiotemporal relation Network (STRNet), to learn more elaborate spatial temporal and motion details in an explicit way. As shown in Fig. 1, this triple branch architecture consists of an appearance branch, an enhanced motion branch and a temporal relation branch. The appearance branch is aimed at learning spatial features using 2D ConvNet. The motion branch is designed to learn refining spatiotemporal features, and the relation branch focuses on understanding the temporal relation from multiple frames. Furthermore, we conduct an artful fusion strategy to integrate multi-branch feature representations for the final classification.

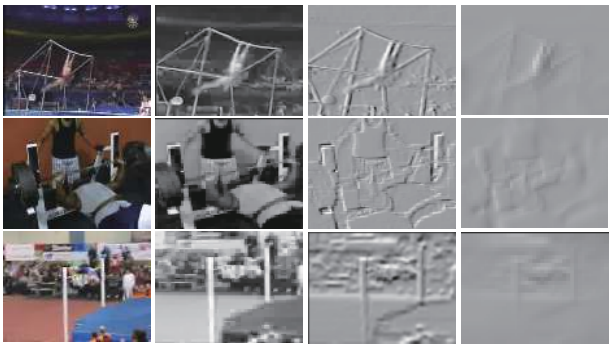

Fig. 2 Feature visualization of STRNet. The first column is the input frames. The second column is the feature maps of Stem. The third column is the fusion feature maps of stage 3. The last column is the output of spatiotemporal with relation feature maps of stage 5. We rescale the feature maps into original size for good comparison.

Our STRNet is generalized and functional to cope with all kinds of video-based action recognition, including scene-related or temporal-related videos. Although we just use RGB frames as input, we achieve a similar or favorable score compared to state-of-the-art methods on most datasets. Furthermore, our STRNet can adapt to diverse short-term or long-term videos and performs well. As shown in Fig. 2, we visualize feature maps of our STRNet.

Our main contribution can be summarized as follows: 1) We propose STRNet, which is designed to learn comprehensive spatial-temporal representations especially enhancing the motion information in an explicit way. 2) We develop an efficient way to model temporal relations between the sequential frames and it brings about a considerable improvement on temporal-related tasks. 3) We optimize the network architecture in an expressive and

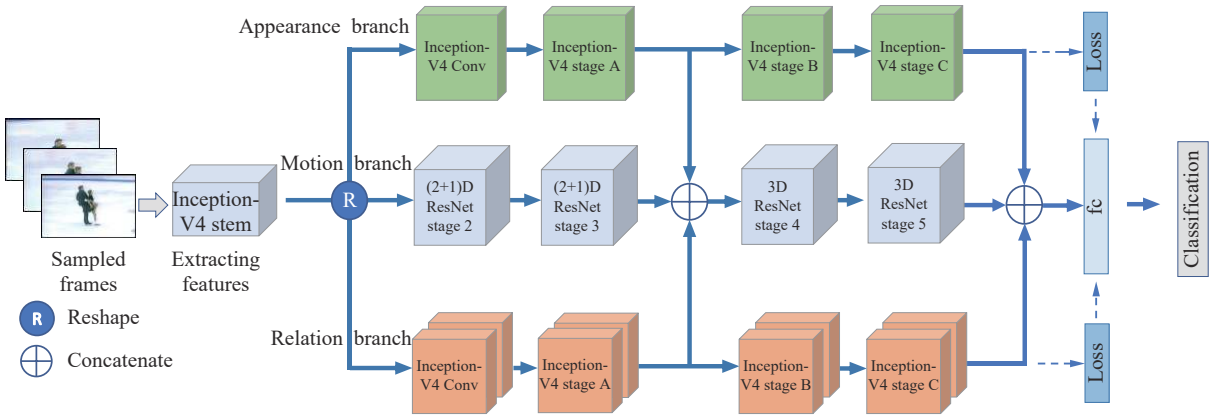

Fig. 1 Architecture overview of STRNet. Our STRNet consists of three individual branches that focus on learning appearance, motion and temporal relation information, respectively. For comprehensively representing the information of the whole video, we apply two-stage fusion and separable (2+1)D convolution to reinforce the feature learning. Finally, we apply a decision level weight assignment to adjust the classification performance.

interpretable manner.

2 Related work

Action recognition with deep learning. Since the emergence of CNN[40] in the task of image classification, many researchers have tried to design effective architectures for action recognition and video classification. Karpathy et al.[18] first applied deep networks with different temporal fusion strategies in a large-scale and noisily-labeled dataset (Sports-1M[18]). Simenyan and Zisserman[15] designed a pioneering two-stream architecture containing two parts of ConvNet: spatial stream ConvNet and temporal stream ConvNet. It sets up a benchmarking flag in the direction of separately modeling spatial and temporal features. Wang et al.[16] proposed a temporal segment network (TSN). It samples video frames sparsely in order to capture long range dependence, and identifies the best score when simultaneously using RGB frames, optical flow and warped flow as the input. 3D convolutional neural networks firstly proposed by Baccouche et al.[41] Tran et al.[17] investigated 3D CNN on learning spatiotemporal features further. Thereafter, several variants of 3D models were put forward. Carreira et al.[11] proposed a new two-stream inflated 3D CNN by expanding 2D CNNs. Sun et al.[42] attempted factorizing 3D ConvNet into spatial (2D) and temporal (1D) direction. The P3D[43] and R(2+1)D[20] models are similar in nature. Wang et al.[39] explored the stream relationship on the basis of multiplicative interaction theory[44], and used 3D convolution to explicitly model the structure like two-stream.

Our work basically focuses on designing a strong feature learning architecture to capture motion information and temporal relation information. We combine the previous outstanding works and propose novel strategies to achieve better performance.

Long-term video representation and learning relations. To deal with the problem of partial observation, some approaches expanded the temporal receptive field of the sampling window. However, increasing the temporal length of the input has two major drawbacks: 1) It requires expensive computational resource. 2) For a long-range video, it is hard to capture the whole visual information of the video. TSN[16] proposed a smart sampling strategy which can cover the whole video by randomly sampling one frame in each segment. Although the samples spread over a long span of the video, the network deals with the sampled frames independently.

Modeling or learning correspondence of frames is also an important mission in computer vision. For action recognition tasks, motion relation tends to be an indispensable factor to the recognition. Early approaches to model relations between two images would involve concatenation, 3D convolution, multiplicative interactions and so on. Recently, some action recognition methods, such as temporal relational reasoning network (TRN)[45], graph convolutional network (GCN)[46], have focused on a temporal relation of the video. But none of them are able to satisfy the requirements of long-term video action recognition as well as the motion relation characteristics simultaneously. We explore an efficient way to solve the problem of motion consistency over a long temporal span.

Attention mechanism. In the past several years, attention models[47–49] have been widely used in various types of deep learning tasks, such as natural language processing, image recognition and voice recognition. It is one of the core technologies of deep learning technology that deserves the most attention and in-depth understanding. Attention for videos may take different forms, including gating, second order pooling[50], cross-model attention[51], self-attention[52]. We propose a brand-new attention method called joint attention, which is applied after a multi-branch output to enhance the classification score.

Our work incorporates three main improvements compared to the previous methods: 1) We propose an explicit way to enhance the motion relation only using RGB frame representation. 2) We propose a novel method to model relations of consecutive frames with a weight adjustment mechanism. As an auxiliary branch, it shows a good combination with motion branch to model spatiotemporal action features. 3) Additionally, to represent the spatiotemporal features more concretely and robustly, we artfully use separated (2+1)D convolution to optimize the primary 3D CNN.

3 Method

In this section, we will introduce our STRNet and the difference from the previous spatial-temporal feature learning networks. For the following triple-branch structure, we will discuss the effect of each branch, and analyze the performance of different combinations.

3.1 Spatial temporal relation network

Our STRNet is designed for efficient spatiotemporal and relation modeling. As shown in Fig. 1, the STRNet consists of three branches to learn appearance, enhanced motion and temporal relation information respectively. After sampling the video, the data is firstly fed into the Inception V4 network[53] to extract the essential feature maps. These feature maps would comprehensively represent image features according to channels.

Appearance branch. For the most scene-related action recognition task, the final classification result depends on the appearance representation to a great extent. Hence, we apply a robust 2D convolutional network to exploit visual semantic characteristics of individual frames. Specifically, we use the Inception V4 architecture for this purpose, due to its multi-scale property as well as

the efficient propagation.

Enhanced motion branch. Empirically, the performance of the 3D convolution network is superior to 2D convolution networks in the aspect of extracting motion information from sequential feature maps, such as convolutional 3D (C3D)[17], inflated 3D convolutional network (I3D)[11]. Inspired by the R(2+1)D network[20], we opt for decomposing spatial and temporal modeling into two separate tasks. By increasing the optimization efficiency, our motion branches can extract rich spatiotemporal features. A separable (2+1)D ResBlock is incorporated to enhance the spatiotemporal feature learning, which consists of 2D convolution filters of size $1\times d\times d$ and temporal convolutional filters of size $t\times 1\times 1$ along with ReLU and batch normalization layers. Specifically, we organize the motion branch into two sections. Based on experience and probabilistic analysis[54], the former part of motion branch, which is before branch fusion (mentioned in Section 3.2), is conducted by (2+1)D ResBlocks. And the latter consists of 3D ResBlocks. It is explicable that in early layers, our network learns the appearance and edge information as well as short time scale motion information using the (2+1)D ResBlocks. Furthermore, the operation of multi-branch fusion assembles both appearance and temporal relation information which is crucial for modeling motion characteristics. With these basics, we employ the 3D convolution operation in the latter part of our motion branch. Because the 3D convolution network is well equipped to model spatiotemporal features and in the deep layers it is more beneficial to learn abstract semantic features of action[54].

Compared to the original 3D convolution networks, our newly combined branch reveals two advantages. Firstly, although keeping the same number of parameters as the 3D networks, we bring in two nonlinear operations i.e. spatial 2D convolution and temporal 1D convolution with a ReLU activation layer. These measures improve the spatial and temporal expressiveness of the network. Secondly, the (2+1)D blocks increase the complexity of the network but relax the interaction of layers so that the network turns out to be more flexible and easier to optimize.

Relation Branch. For our STRNet, the relation branch is the newly proposed scheme and the key part to model comprehensive spatiotemporal characteristics. A key problem in video-based action recognition is how to master the continuously changed motion information. Firstly, we evaluate four tentative approaches to model the relationship between two sequential images: 1) element-wise subtraction, 2) element-wise addition, 3) element-wise product, 4) matrix-multiplication. The reasons why we choose these four methods are shown as follows. 1) Element-wise subtraction: In an intuitive sense, the instantaneous motion lies in the difference between two adjacent frames. Recently, some experiments and papers[16, 21, 55] show that pixel-wise subtraction can obtain high frequency information which can indicate the position of the motion in one frame. So, it is a candidate for representing the temporal relationship. 2) Element-wise addition: As opposed to subtraction, addition operation keeps the low frequency information[55], it is good for the scene and object recognition. And it can capture the static relation between frames. 3) Element-wise product: That is Hadamard product, it can reflect the geometric distance of the matrix. So, we also take it as an experimental item. 4) Matrix-multiplication: It can be regarded as computing the correlation of the output image with a transformed version of the input image. Furthermore, the similarity of adjacent frames reflects the amplitude of the motion. Matrix multiplication can be more convenient to calculate the degree of similarity. Based on these assumptions, we test each of these four methods respectively. After the experiment, we adopt the matrix-multiplication technique as the final option because of its interpretability as well as the best performance. Specifically, the relationship between the two frames is defined as (1).

$$\tilde{x}_i = x_i^{\mathrm{T}} x_{i+1} \tag{1}$$

where x_i and x_{i+1} are adjacent feature maps and i is the index of the time sequence, thus ($i\in\{1, 2, \cdots, T-1\}$). In addition, we define $x_T = x_T^{\mathrm{T}} x_T$, i.e., self-correlation. The schema of an enhanced temporal relation unit is shown in Fig. 3.

However, sparse sample may result in great variations between a pair of frames. For sequential frames, the higher similarity score implies the smaller variation. Enlightened by the optical flow theory[35], one of the pos-

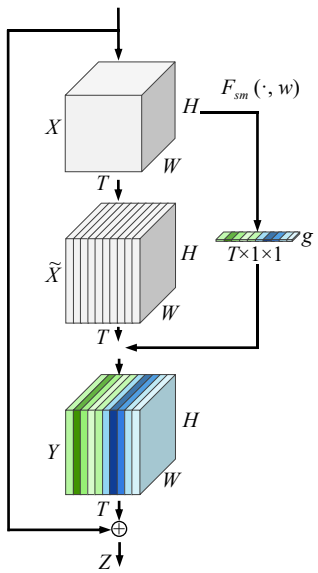

Fig. 3 The schema of building relation unit, where X denotes the original inputs of the sequential feature maps, and $\tilde{X}$ denotes the calculated relation maps. The function $F_{sm}(*)$ is to calculate the similarity measurement. And g denotes the similarity weight vector and Y denotes the final relation response maps.

tulated conditions is that the motion is tiny, so that the relation characteristics between frames can be modeled accurately. In other words, great variations imply weak correlation. Hence, we should pay more attention to those with strong relationship and consistency.

Based on this observation, we design an adaptive method to adjust the weight of the representation of the relation between sequential frames. Here, we induce a similarity measurement calculation f which indicates the similarity of two feature maps. Formally, a statistic $s \in \mathbf{R}^{T\times H\times W}$ is generated by f, such that the i-th element of s is calculated by (2).

$$s_i = f(x_i, x_{i+1}) = \theta(x_i)^{\mathrm{T}}\phi(x_{i+1}) \tag{2}$$

where $\theta(x_i) = W_\theta x_i$ and $\phi(x_{i+1}) = W_\phi x_{i+1}$ $(i\in\{1, 2, \cdots, T-1\})$ are two embeddings. Refer to (3), the similarity weight vector $g \in \mathbf{R}^{T\times 1\times 1}$ can be deduced by function F_{sm}.

$$g = F_{sm}(X, W) = \sigma(GAP(f(x, W))) \tag{3}$$

where GAP denotes global average pooling and σ denotes softmax activation function, which are used to squeeze and normalize the similarity measurement. As a consequence, the final output Z of our relation block can be written as (4).

$$Z = X + Y = X + \widetilde{X}g. \tag{4}$$

The detailed flowchart is shown in Fig. 3. As such, our work provides a new insight into temporal attention in the space-time video, induced by an action recognition task in computer vision.

3.2 Fusion and optimization

How to effectively and rationally utilize the learned appearance and temporal relation information is also a key issue for our task. This demands us to explore the potential value of spatiotemporal features. We propose a 2-stage fusion strategy to combine the two information modalities. Specifically, we apply two lateral connections[56] in the middle and the end of network to fuse the appearance and relation information into motion branch as shown in Fig. 1. In other words, the spatial features and the temporal relation features are merged into a spatiotemporal feature map. We make full use of the mid-level detailed information and high-level semantic information. In this way, the network gains the complementary information conveyed by multi-branch fusion.

To enhance the feature expression as well as to optimize the network propagation, we investigate the merits of separable (2+1) D convolution in video-based action recognition. Analyzing the specialties of our STRNet, we apply (2+1) D blocks in the early step of motion branch before the first-stage fusion layer, because the separable (2+1) D convolution is of high pertinence to model spatiotemporal features in early layers. Moreover, this type architecture facilitates the network optimization.

3.3 Joint attention

Thanks to our multi-branch structure, the network can learn different types of video feature representation. Different branches may produce different effects. For example, if the task has explicit background and foreground, the appearance branch will make a great difference to the final classification. In addition, for complex human-human or human-object interactive actions, it needs to explore the deeper motion and relation information. So how to combine the most useful feature information from the multi-branch processing is an important issue. For this purpose, we propose a joint attention mechanism to focus on discriminative classification result from each branch. In our model, we adopt a horizontal classification strategy. According to different contributions to the final classification, we calculate the weight score of each individual branch output. Thus, we define the final loss function as (5).

$$L = L_{\mathrm{all}} + \alpha L_a + \beta L_m + \gamma L_r \tag{5}$$

where L_a, L_m and L_r are appearance, motion and relation branch loss respectively, α, β, and γ are hyper-parameters. The experiment shows that it has a positive effect on the final result. We define the loss function L_* with cross entropy as (6).

$$L_* = -t_j \log(y_j) \tag{6}$$

where t_j means the ground-truth of class j.

3.4 Backward-propagation

We collect three losses from the triple-stream forward-propagation and combine them to obtain the final loss. Then the final weighted loss acts on each branch to adjust the network parameters as backward-propagation.

3.5 Instantiation

Our STRNet is generic, and it can be instantiated with different backbones[5, 23, 53]. In this subsection, we describe our instantiations of the network architectures.

As shown in Table 1, we use 3D ResNet-34[57] to construct the motion branch, and Inception V4[53] to build the appearance branch and the relation branch separately. We use temporally strided 3D units as a substitute for 2D convolutional block in appearance and relation branch. For the convenience of fusion, the size of input is denoted as $N\times C\times T\times 224\times 224$. We strictly define the

Table 1 An instantiation of our STRNet. The dimensions of kernels are denoted by $[T\times H\times W, C]$ for spatial-temporal and channel sizes. For computational efficiency, we reduce the number of channels of the motion branch. We only apply two temporal down sampling to keep the integrated temporal feature.

Layer name	Output size	Motion branch	Layer name	Output size	Appearance & relation branch
Stem	$T\times 56\times 56$	$1\times 3\times 3, 64$	Stem	$T\times 56\times 56$	$1\times 3\times 3, 64$
Stage 2	$T\times 56\times 56$	$\begin{bmatrix} 1\times 3\times 3, 64 \\ 3\times 1\times 1, 64 \\ 1\times 1\times 1, 64 \end{bmatrix}\times 3$	Convolution	$T\times 56\times 56$	$1\times 3\times 3, 192$
Stage 3	$T\times 28\times 28$	$\begin{bmatrix} 1\times 3\times 3, 64 \\ 3\times 1\times 1, 64 \\ 1\times 1\times 1, 128 \end{bmatrix}\times 4$	Inception-A	$T\times 28\times 28$	$\begin{bmatrix} 1\times 1\times 1, 96 \\ 1\times 3\times 3, 96 \end{bmatrix}\times 4$
Stage 4	$\frac{T}{2}\times 14\times 14$	$\begin{bmatrix} 3\times 3\times 3, 128 \\ 1\times 1\times 1, 256 \end{bmatrix}\times 6$	Inception-B	$\frac{T}{2}\times 14\times 14$	$\begin{bmatrix} 1\times 1\times 1, 128 \\ 1\times 1\times 7, 256 \\ 1\times 7\times 1, 256 \end{bmatrix}\times 7$
Stage 5	$\frac{T}{4}\times 7\times 7$	$\begin{bmatrix} 3\times 3\times 3, 128 \\ 1\times 1\times 1, 512 \end{bmatrix}\times 3$	Inception-C	$\frac{T}{4}\times 7\times 7$	$\begin{bmatrix} 1\times 1\times 1, 256 \\ 1\times 1\times 3, 256 \\ 1\times 3\times 1, 256 \end{bmatrix}\times 3$
Average pool, concatenate, dropout, fully connected layer					

stride of Inception V4 Stem so that we can apply suitable fusion in every stage.

Specifically, concatenation is used as our fusion method after stage 3 and stage 5. Furthermore, to keep the integrated temporal feature, we only apply two temporal down sampling operations in stage 4 and stage 5 corresponding to Inception-B and Inception-C in Table 1. Finally, we adopt global average pooling and a fully connected layer to calculate the classification score of each branch. Then we get the final loss through joint attention method.

4 Experiments

In this section we first introduce four popular and challenging datasets in the field of action recognition. We then present the details of the implementation and the experimental results to show the generalization of our approach. We evaluate and compare our proposed method with the baseline and state-of-the-art methods. We further investigate the superiority of our method from different aspects. The baseline method in our experiment is ECO-Full (efficient convolutional network for online video understanding) where we replace the backbone with ResNet-34 for fair comparison.

4.1 Datasets

We evaluate the performance of our STRNet on four action recognition benchmarks: Something-Something v1[38], Kinetics-400[11], UCF-101[58], and HMDB-51[59]. According to their space-time properties, these datasets can be divided into two categories: 1) Scene-related datasets including UCF-101, HMDB-51, and Kinetics-400 & 600 in which the appearance of objects and scenes are the most discriminative information for estimating the label of the action class. Meanwhile, the numbers of sampled frames and the sequence of the context play a relatively weak role in these datasets, because the temporal relation is not so important for final action classification. 2) Temporal-related datasets, including Something-Something v1 & v2, where there is a strong correlation between the context frames. To acquire a high recognition accuracy, we must take the temporal relationship into significant consideration. In view of the above video characteristics, our proposed method can simultaneously model the spatial and temporal information in an efficient way.

The UCF-101 is a popular and representative dataset in action recognition. It contains 101 pre-defined classes and 13 320 video clips. The HMDB-51 is composed of 51 action categories including 6 766 video clips. The Kinetics is a large and challenging dataset which has two released versions, i.e., Kinetics-400 and Kinetics-600. The Kinetics-400 contains 400 action classes and each class has at least 400 videos. The Something-Something v1 is a large video collection with detailed description labels which contain temporal relationships including even causal relationship based on human-object interactions. The dataset contains 174 classes including 108 499 videos.

4.2 Implementation details

Our approach is implemented in the Pytorch framework and all networks are trained on 4 GEFORCE RTX 2 080 Ti GPUs. In the following, we will describe the implementation details of our method.

Training. We train our STRNet with the same strategy as mentioned in ECO due to its efficiency. The

video is split into T subsections written S_i, where i=1, 2, ⋯, T of the same interval. We randomly select one frame as a sample in each subsection. We collect these T frames as input in order to cover a long-range temporal sequence from all frames and strengthen the robustness of video variations. For data augmentation, we apply randomly fixed-corner cropping and scale-jittering to the sampled frames. Then, we resize the processed frames into 224×224 for convolutional convenience. The size of the input is N×C×T×224×224, where N is the batch size, C denotes the number of channels (usually 3 for the first layer) and T is the number of sampled frames. For the Kinetics-400 datasets, we train our STRNet from scratch. We initialize the weight parameters with the Xavier method[60] which is known to be effective in most networks. The initial learning rate is 0.01 and decreases by a factor of 10 at 60th and 80th epoch. We fine-tune the Kinetics pre-trained model on UCF-101, HMDB-51 and Something-Something datasets. The learning rate starts with 0.001 and is decayed by a factor of 10 at the 15th and 30th epoch. We set the momentum of 0.9 and weight decay to 1×10^{-4} for all training steps. We use the synchronized SGD training. The default number of sampled frames is 16. After experimental verification, the hyper-parameters α, β and γ are set to 0.5, 0.8 and 0.5 respectively on scene-related datasets, i.e., UCF-101 and Kinetics. While α, β and γ are set to 0.5, 0.5 and 0.8 respectively on temporal-related datasets. That is to say, we magnify the impact of temporal relation branch on final classification.

4.3 Ablation study

The construction of the network. To study our STRNet's performance in all aspects, we choose two well-performing models (ARTNet[39] and STM[21]) as baselines. For the purpose of fairness and efficiency, we train UCF-101 from scratch as a control experiment. The ARTNet and the STM network proposed independent blocks which can learn appearance and relation information in an integrated manner. However, with the blocks stacking up, the latter blocks cannot extract the spatial-temporal and motion features explicitly on account of their mixture in an early stage. For this reason, we explore an explicit and relatively independent way to learn the appearance, motion and relation information through three branches. Meanwhile, imitating the STM block, we build the STR block, which can model the appearance, motion and relation information in a specific block. The experimental results are shown in Table 2 that our STRNet with three branches outperforms other models.

The proposed multiple branch complementary fusion strategy. The challenging factors of the action recognition task include scene, object and motion information. For the most scene-related datasets, the model may often acquire a high accuracy from the 2D appearance features (i.e., scene and object). However, exploring the motion and relation information improves recognition accuracy as well. In this case, we experimentally evaluate several complementary fusion strategies. Similar to TSN[16] and SlowFast[56] networks, we attach a cross-branch connection between each two branches for every residual block. It is worth mentioning that we choose the motion branch as the main branch, then we fuse the appearance and relation branch into the motion branch simultaneously. The two-stage fusion refers to the lateral connection that locates in the middle and the end of the motion branch, graphic representation is shown in Fig. 1. Likewise, the full-stage fusion means that we apply lateral connection at every stage of the convolution. The experimental results are shown in Table 3. The reason why the full-stage fusion gets inferior performance is that the superfluous fusion renders the multi-type features heterogeneous, so the pertinence of each individual branch goes unrewarded. In contrast, our two-stage fusion not only explicitly learns the characteristics of each branch, but also integrates the appearance, motion and relation information effectively in a complementary manner.

Table 2 Comparison with different constructions of the network

Method	Accuracy (%)
SMART block (ARTNet)	58.2
STM block	59.8
STR block	59.5
STR branch	**62.8**

Table 3 Results of different fusion strategies in our STRNet

Method	Accuracy (%)
Without fusion	62.8
Two-stage fusion	**63.2**
Full-stage fusion	62.0

Network optimization. Consider fusing the appearance and the relation information in the motion branch. In other words, it improves the ability of characteristic expression, but aggravates the network optimization burden. We design two variants of our main motion branch. The first is (2+1)D and 3D combination form, the second is a full (2+1) D form. Specifically, in the first scheme, we replace only the stage 2 and stage 3 block of 3D ResNet-34 with a separable (2+1)D block to process the feature maps before the fusion step. In another scheme, we replace all 3D blocks with (2+1)D blocks. The experimental results in Table 4 indicate that our improved (2+1)D joint 3D performs better than the full (2+1)D or full 3D. In theory, the optimization becomes easier by separating the 3D convolution into spatial and temporal components. The separable (2+1)D convolution has strong ad-

aptive ability to constitute spatiotemporal feature representation. While after our fusion step, we apply 3D blocks for subsequent feature extraction. Because the 3D convolution network is well equipped to model spatiotemporal features and in the deep layers, it is more beneficial to learn abstract semantic feature of action.

Study on the contribution of individual branch. In order to verify the effectiveness of our model, we separate our STRNet in pairs for experiment. The results are shown in Table 5, it illustrates that our motion branch makes a major contribution to the feature learning. That is the reason why we choose motion branch as our main branch. However, it seems like the relation branch has little effect on the result. The reason is that the UCF-101 is a scene-related dataset, but our relation branch focuses more attentions on the temporal relation of the video sequence. More result analyses will be argued for in the next part.

Experiment on hyperparameters. Due to our multi-branch structure, the network can learn various types of video feature representations. Each branch actually has a different influence on classification. In addition, the effect of each branch differs from different characteristics datasets. Based on this assumption, we set three weight hyperparameters for verification. Through the experiment, we find that on scene-related datasets, such as UCF-101 and HMDB-51, when we magnify the motion loss and reduce temporal loss relatively, it is beneficial to the final classification. Whereas, the performance goes the best when we magnify the relation loss and reduce motion loss on temporal-related dataset, e.g., Something-Something v1. This also demonstrates the validity of our relation branch from the perspective of prior probability. The detailed statistics are shown in Table 6. Moreover, to make our model more universal, we consider the motion branch as the main force, the relation and appearance branch as support branch. Thus, the default values of hyperparameters α, β and γ are set to 0.5, 0.8 and 0.5 respectively.

Table 4 Results of different main branch structure of the network

Method	Accuracy (%)
3D ResNet	63.2
(2+1)D ResNet	63.9
3D&(2+1)D ResNet	64.3
(2+1)D&3D ResNet	**64.7**

Table 5 Results of different branching combinations of the network

Method	Accuracy (%)
Appearance + Motion	63.2
Appearance + Relation	59.5
Motion + Relation	63.0
Full	**64.7**

Effectiveness of the relation branch. As discussed above, our STRNet comprehensively learns the spatiotemporal and motion relation information. In this part, we will discuss how the temporal relation branch works and why it performs significantly. We compare the result of the model with and without the relation branch in Table 7 on the Something-Something v1 dataset. It shows that the relation branch brings a significant improvement (44.0% VS. 40.7%).

4.4 Results on Kinetics dataset

Based on the above research work, we evaluate our STRNet on the Kinetics-400 dataset. Table 8 shows the result of STRNet and other competing methods on the Kinetics-400 dataset. From the evaluation results, we find that most actions of Kinetics can be recognized by scene and objects even from one static frame of video. Therefore, we can manually adjust the parameters to achieve acceptable accuracy referred to in (5).

4.5 Results on UCF-101 and HMDB-51

We apply our STRNet on the UCF-101 and HMDB-51 datasets to compare its performance with the state-of-the-art methods. Following the official evaluation metrics, we test our methods over three splits and report the average results in Table 9. In view of the experience of the previous best performing approaches, we apply transfer learning on these datasets. Specifically, we first pre-train our model on Kinetics-400 for 40 epochs. Then, we finetune the network on UCF-101 and HMDB-51 datasets. The results indicate that the pre-trained model indeed significantly improves the performance on the small dataset. Compared to two-stream methods such as I3D and TSN, although our method is slightly worse, two-stream methods need to extract optical-flow which induces redundant computation, whereas our method only uses RGB frames as input.

Table 6 Experiments on hyperparameters

α	β	γ	UCF-101	S-S v1(%)
0.8	0.5	0.5	63.8	43.2
0.5	0.8	0.5	**64.7**	43.5
0.5	0.5	0.8	64.0	**44.0**
0	0	0	63.1	42.8

Table 7 Demonstrating the effectiveness of the relation branch; trained on Something-Something v1 from scratch.

Method	Accuracy (%)
STRNet without Relation-branch	40.7
STRNet	**44.0**

Table 8 Performance of our STRNet on Kinetics-400 dataset compared with state-of-the-art methods. For a fair comparison, we only use RGB input and train from scratch on Kinetics.

Method	Flow	Backbone	Frames	FLOPs	Top-1(%)	Top-5(%)
Spatio-temporal channel correlation network (STC)[61]		ResNet-101	16	–	68.7	88.5
TSN RGB[16]		BNInception	16	80G	69.1	88.7
TSN two-stream	√	BNInception	16	–	73.9	91.1
I3D[11]		Inception V1	64	359G	71.4	89.3
I3D two-stream	√	Inception V1	64	–	74.2	91.3
Spatiotemporal-separable 3D convolution (S3D)[62]		Inception V1	64	–	72.2	90.6
ARTNet[39]		3D ResNet-18	16	34G	69.2	88.3
SlowFast[56]		3D ResNet-101	16+8	234G	79.8	93.9
Temporal shift module (TSM)[63]		ResNet-50	16	65G	74.7	90.7
STM[21]		RseNet-50	16	67G	73.7	91.6
ECO[57]		BNInception +3D ResNet-18	92	267G	70.7	89.4
STRNet(ours)		Inception V4+3D ResNet-34	16	103G	75.0	92.1

Table 9 Comparison with state-of-the-art methods on the UCF-101 and HMDB-51 datasets. The accuracy is reported as average over three splits. The model is pre-trained only on Kinetics and performs pretty well. And we only use RGB frames as input.

Method	Flow	Backbone	Pre-train	UCF-101(%)	HMDB-51(%)
C3D[17]		3D VGG-11	Sports-1M	82.3	51.6
TSN RGB[16]		BNInception	ImageNet + Kinetics	91.1	–
TSN two-stream	√	BNInception	ImageNet + Kinetics	97.0	–
I3D[11]		Inception V1	ImageNet	95.1	74.3
I3D two-stream	√	Inception V1	ImageNet	98.8	80.7
ARTNet[39]		3D ResNet-18	Kinetics	94.3	70.9
ECO[57]		BNInception +3D ResNet-18	Kinetics	94.8	72.4
TSM[63]		ResNet-50	ImageNet + Kinetics	94.5	70.7
STM[21]		RseNet-50	Kinetics	96.0	72.7
STRNet(ours)		Inception V4+3D ResNet-34	Kinetics	**96.7**	**73.1**

4.6 Results on Something-Something

Something-Something v1 is a large and challenging dataset with densely-labeled video clips. As a temporal-related dataset, the video mainly consists of human-object interactions whose sequential order is strictly directional. Meanwhile, some of the videos are confusing even possessing a causal relationship, such as "Pretending to pour something out of something, but something is empty." and "Putting something that cannot roll onto a slanted surface, so it stays where it is." As a consequence, most action recognition methods cannot reach a very high accuracy on this challenging dataset. Our STRNet is designed to simultaneously model the spatial and temporal relationship. Table 10 displays the results of our method and state-of-the-art methods. Without the optical flow stream, our method achieves a state-of-the-art result. Compared to ECO method, our STRNet exhibits a 4.3% improvement only with 16 frames inputs. Our approach achieves the best performance on top-1 test and top-5 validation sets.

4.7 Evaluation on COIN dataset

The COIN[64] dataset consists of 11 827 videos related to 180 different tasks. It is currently the largest dataset for comprehensive instructional video analysis. It also can be applied for action recognition mission. The comprehensive instructional video analysis (COIN) dataset has a hierarchical dictionary. We choose the second level "Task" as the label to generalize on action recognition mission. We analyze the COIN dataset and relevant experiment. The length of the video ranges from 2 min to 10 min, and most videos have descriptive subtitles and narrative frames which are the disturbance term for motion and relation modelling. In addition, these instructional videos contain lots of shot cuts of the camera. Thus, we apply cluster sampling on a video rather than random sampling in experiments to capture the effective action. Table 11 shows the result of our STRNet and other competing methods on the COIN dataset which

Table 10 Comparison with state-of-the-art methods on Something-Something v1 dataset. We pre-train our model on Kinetics and we only use RGB frames as input.

Method	Backbone	Pre-train	Frames	Top-1 val (%)	Top-5 val (%)	Top-1 test (%)
TRN[45]	BNInception	ImageNet	8	34.3	–	33.6
S3D[62]	Inception V1	ImageNet	64	48.2	78.7	42.0
I3D RGB[11]	Inception V1	ImageNet + Kinetics	32	41.6	72.2	–
TSN RGB[16]	BNInception	Kinetics	16	19.7	46.6	–
ECO[57]	BNInception +3D ResNet-18	Kinetics	92	46.4	–	42.3
STM[21]	RseNet-50	ImageNet	16	50.7	80.4	43.1
TSM[63]	ResNet-50	ImageNet + Kinetics	16	46.8	76.1	-
STRNet(ours)	Inception V4+3D ResNet-34	Kinetics	16	50.7	**80.6**	**43.5**

Table 11 Results on the COIN dataset

Method	Accuracy (%)
TSN[16]	88.0
ECO[57]	88.3
STRNet	**89.1**

demonstrates the advantage on accuracy of our network.

5 Conclusions

In this paper, we propose a novel architecture called STRNet, which is designed to learn comprehensive spatiotemporal and motion relation features from videos. One of our key innovations is that we construct the relation branch via utilizing the correlation and similarity of the contexts. Compared to previous stacking-block based networks, our method has an intuitive sense of interpretability. Besides, our STRNet takes advantage of the 1D, 2D and 3D convolutional networks for spatiotemporal feature learning. The effectiveness of the proposed STRNet is verified by experiment on several challenging benchmark datasets on action recognition. However, high accuracy comes at the expense of large parameters, also our network is of poor portability due to the specific multi-branch architecture. For a future study, we plan to make our STRNet more light-weight by knowledge distillation. We also plan to apply our STRNet to other pattern recognition tasks based on videos.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. U1836218, 62020106012, 61672265 and 61902153), the 111 Project of Ministry of Education of China (No. B12018), the EPSRC Programme FACER2VM (No. EP/N007743/1) and the EPSRC/MURI/Dstl Project under (No. EP/R013616/1.)

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*, New York, USA: Springer, 2006.

[2] D. Michie, D. J. Spiegelhalter, C. C. Taylor. *Machine Learning, Neural and Statistical Classification*, Englewood Cliffs, USA Prentice Hall, 1994.

[3] Y. LeCun, Y. Bengio, G. Hinton. Deep learning. *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ACM, Lake Tahoe, USA, pp. 1097−1105, 2012.

[5] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Dep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770−778, 2016. DOI: 10.1109/CVPR.2016.90.

[6] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 1–9, 2015. DOI: 10.1109/CVPR.2015.7298594.

[7] J. W. Han, D. W. Zhang, G. Cheng, N. A. Liu, D. Xu. Advanced deep-learning techniques for salient and category-specific object detection: A survey. *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018. DOI: 10.1109/Msp.2017.2749125.

[8] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 779−788, 2016. DOI: 10.1109/CVPR.2016.91.

[9] H. Noh, S. Hong, B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 1520−1528, 2015. DOI: 10.1109/ICCV.2015.178.

[10] E. Shelhamer, J. Long, T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017. DOI: 10.1109/TPAMI.2016.2572683.

[11] J. Carreira, A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 6299−6308, 2017. DOI: 10.1109/CVPR.2017.502.

[12] X. F. Ji, Q. Q. Wu, Z. J. Ju, Y. Y. Wang. Study of human action recognition based on improved spatio-temporal features. *International Journal of Automation and Computing*, vol. 11, no. 5, pp. 500–509, 2014. DOI: 10.1007/s11633-014-0831-4.

[13] L. M. Wang, Y. Qiao, X. O. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 4305–4314, 2015. DOI: 10.1109/CVPR.2015.7299059.

[14] X. L. Wang, A. Farhadi, A. Gupta. Actions ~ Transformations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 2658–2667, 2016. DOI: 10.1109/CVPR.2016.291.

[15] K. Simonyan, A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ACM, Montreal, Canada, pp. 568–576, 2014.

[16] L. M. Wang, Y. J. Xiong, Z. Wang, Y. Qiao, D. H. Lin, X. O. Tang, L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 20–36, 2016. DOI: 10.1007/978-3-319-46484-8_2.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 4489–4497, 2015. DOI: 10.1109/ICCV.2015.510.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F. F. Li. Large-scale video classification with convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Columbus, USA, pp. 1725–1732, 2014. DOI: 10.1109/CVPR.2014.223.

[19] B. W. Zhang, L. M. Wang, Z. Wang, Y. Qiao, H. L. Wang. Real-time action recognition with enhanced motion vector CNNs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 2718–2726, 2016. DOI: 10.1109/CVPR.2016.297.

[20] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 6450–6459, 2018. DOI: 10.1109/CVPR.2018.00675.

[21] B. Y. Jiang, M. M. Wang, W. H. Gan, W. Wu, J. J. Yan. STM: SpatioTemporal and motion encoding for action recognition. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 2000–2009, 2019. DOI: 10.1109/ICCV.2019.00209.

[22] Z. G. Tu, H. Y. Li, D. J. Zhang, J. Dauwels, B. X. Li, J. S. Yuan. Action-stage emphasized spatiotemporal VLAD for video action recognition. *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2799–2812, 2019. DOI: 10.1109/TIP.2018.2890749.

[23] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. [Online], Available: https://arxiv.orglabs/1409.1556, 2014.

[24] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 4700–4708, 2017. DOI: 10.1109/CVPR.2017.243.

[25] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, vol. 64, no. 2–3, pp. 107–123, 2005. DOI: 10.1007/s11263-005-1838-7.

[26] H. Wang, C. Schmid. Action recognition with improved trajectories. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Sydney, Australia, pp. 3551–3558, 2013. DOI: 10.1109/ICCV.2013.441.

[27] L. M. Wang, Y. Qiao, X. O. Tang. MoFAP: A multi-level representation for action recognition. *International Journal of Computer Vision*, vol. 119, no. 3, pp. 254–271, 2016. DOI: 10.1007/s11263-015-0859-0.

[28] X. L. Song, C. L. Lan, W. J. Zeng, J. L. Xing, X. Y. Sun, J. Y. Yang. Temporal-spatial mapping for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 748–759, 2020. DOI: 10.1109/Tcsvt.2019.2896029.

[29] S. W. Ji, W. Xu, M. Yang, K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012. DOI: 10.1109/TPAMI.2012.59.

[30] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 4694–4702, 2015. DOI: 10.1109/CVPR.2015.7299101.

[31] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, K. Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 2625–2634, 2015. DOI: 10.1109/CVPR.2015.7298878.

[32] S. J. Yan, Y. J. Xiong, D. H. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, USA, pp. 7444–7452, 2018.

[33] C. Wu, X. J. Wu, J. Kittler. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshop*, IEEE, Seoul, Korea, pp. 1740–1748, 2019. DOI: 10.1109/ICCVW.2019.00216.

[34] H. S. Wang, L. Wang. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4382–4394, 2018. DOI: 10.1109/TIP.2018.2837386.

[35] B. K. P. Horn, B. G. Schunck. Determining optical flow. *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, 1981. DOI: 10.1117/12.965761.

[36] H. Sak, A. W. Senior, F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, Singapore, pp. 338–342, 2014.

[37] C. H. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Q. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, J. Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, IEEE, Salt Lake City, USA, pp. 6047–6056, 2018. DOI: 10.1109/CVPR.2018.00633.

[38] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE Proceedings of International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 5843–5851, 2017. DOI: 10.1109/ICCV.2017.622.

[39] L. M. Wang, W. Li, W. Li, L. Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 1430–1439, 2018. DOI: 10.1109/CVPR.2018.00155.

[40] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[41] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt. Sequential deep learning for human action recognition. In *Proceedings of the 2nd International Workshop on Human Behavior Understanding*, Springer, Amsterdam, The Netherlands, pp. 29–39, 2011. DOI: 10.1007/978-3-642-25446-8_4.

[42] L. Sun, K. Jia, D. Y. Yeung, B. E. Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 4597–4605, 2015. DOI: 10.1109/ICCV.2015.522.

[43] Z. F. Qiu, T. Yao, T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 5533–5541, 2017. DOI: 10.1109/ICCV.2017.590.

[44] R. Memisevic. Learning to relate images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1829–1846, 2013. DOI: 10.1109/TPAMI.2013.53.

[45] B. L. Zhou, A. Andonian, A. Oliva, A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 803–818, 2018. DOI: 10.1007/978-3-030-01246-5_49.

[46] R. H. Zeng, W. B. Huang, C. Gan, M. K. Tan, Y. Rong, P. L. Zhao, J. Z. Huang. Graph convolutional networks for temporal action localization. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 7093–7102, 2019. DOI: 10.1109/ICCV.2019.00719.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Uszkoreit, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ACM, Long Beach, USA, pp. 5998–6008, 2017.

[48] H. Z. Chen, G. H.Tian, G. L. Liu. A selective attention guided initiative semantic cognition algorithm for service robot. *International Journal of Automation and Computing*, vol. 15, no. 5, pp. 559–569, 2018. DOI: 10.1007/s11633-018-1139-6.

[49] T. V. Nguyen, Z. Song, S. C. Yan. STAP: Spatial-temporal attention-aware pooling for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 77–86, 2015. DOI: 10.1109/Tcsvt.2014.2333151.

[50] X. Long, C. Gan, G. De Melo, J. J. Wu, X. Liu, S. L. Wen. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7834–7843, 2018. DOI: 10.1109/CVPR.2018.00817.

[51] X. Zhang, Q. Yang. Transfer hierarchical attention network for generative dialog system. *International Journal of Automation and Computing*, vol. 16, no. 6, pp. 720–736, 2019. DOI: 10.1007/s11633-019-1200-0.

[52] X. L. Wang, R. Girshick, A. Gupta, K. M. He. Non-local neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7794–7803, 2018. DOI: 10.1109/CVPR.2018.00813.

[53] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, AAAI, San Francisco, USA, pp. 4278–4284, 2017.

[54] Y. Z. Zhou, X. Y. Sun, C. Luo, Z. J. Zha, W. J. Zeng. Spatiotemporal fusion in 3D CNNs: A probabilistic view. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9829–9838, 2020. DOI: 10.1109/CVPR42600.2020.00985.

[55] H. S. Su, J. Su, D. L. Wang, W. H. Gan, W. Wu, M. M. Wang, J. J. Yan, Y. Qiao. Collaborative distillation in the parameter and spectrum domains for video action recognition. [Online], Available: https://arxiv.org/abs/2009.06902, 2020.

[56] C. Feichtenhofer, H. Q. Fan, J. Malik, K. M. He. Slowfast networks for video recognition. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 6201–6210, 2019. DOI: 10.1109/ICCV.2019.00630.

[57] M. Zolfaghari, K. Singh, T. Brox. ECO: Efficient convolutional network for online video understanding. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 695–712, 2018. DOI: 10.1007/978-3-030-01216-8_43.

[58] K. Soomro, A. R. Zamir, M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. [Online], Available: https://arxiv.org/abs/1212.0402, 2012.

[59] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre. HMDB: A large video database for human motion recognition. In *Proceedings of International Conference on Computer Vision*, IEEE, Barcelona, Spain, pp. 2556–2563, 2011. DOI: 10.1109/ICCV.2011.6126543.

[60] X. Glorot, Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, JMLR, Sardinia, Italy, pp. 249–256, 2010.

[61] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, L. Van Gool. Spatio-temporal channel correlation networks for action classification. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 284–299, 2018. DOI: 10.1007/978-3-030-01225-0_18.

[62] S. N. Xie, C. Sun, J. Huang, Z. W. Tu, K. Murphy. Rethinking spatiotemporal feature learning for video understanding. [Online], Available: https://arxiv.org/abs/1712.04851, 2017.

[63] J. Lin, C. Gan, S. Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 7082−7092, 2019. DOI: 10.1109/ICCV.2019.00718.

[64] Y. S. Tang, J. W. Lu, J. Zhou. Comprehensive instructional video analysis: The COIN dataset and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. DOI: 10.1109/TPAMI.2020.2980824.

Zhi-Wei Xu received the B. Eng. degree in computer science and technology from Harbin Institute of Technology, China in 2017. He is a postgraduate student at School of Artificial Intelligence and Computer Science, Jiangnan University, China.

His research interests include computer vision, video understanding and action recognition.

E-mail: zhiwei_xu@stu.jiangnan.edu.cn
ORCID iD: 0000-0003-1472-431X

Xiao-Jun Wu received the B. Sc. degree in mathematics from Nanjing Normal University, China in 1991. He received the M. Sc. and the Ph. D. degrees in pattern recognition and intelligent systems from Nanjing University of Science and Technology, China in 1996 and 2002, respectively. He is currently a professor in artificial intelligent and pattern recognition at the Jiangnan University, China.

His research interests include pattern recognition, computer vision, fuzzy systems, neural networks and intelligent systems.

E-mail: wu_xiaojun@jiangnan.edu.cn (Corresponding author)
ORCID iD: 0000-0002-0310-5778

Josef Kittler received the B. A. degree in electrical science tripos, Ph. D. degree in pattern recognition, and D. Sc. degree from University of Cambridge, UK in 1971, 1974, and 1991, respectively. He is a Distinguished Professor of machine intelligence at Centre for Vision, Speech and Signal Processing, University of Surrey, UK. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook *Pattern Recognition: A Statistical Approach* and over 700 scientific papers. His publications have been cited more than 66 000 times (Google Scholar). He is series editor of *Springer Lecture Notes on Computer Science*. He currently serves on the Editorial Boards of *Pattern Recognition Letters*, *Pattern Recognition and Artificial Intelligence*, and *Pattern Analysis and Applications*. He also served as a member of the Editorial Board of *IEEE Transactions on Pattern Analysis and Machine Intelligence* during 1982−1985. He served on the Governing Board of the *International Association for Pattern Recognition* (IAPR) as one of the two British representatives during the period 1982-2005, and President of the IAPR during 1994−1996.

His research interests include robotics, feedback control systems, and control theory.

E-mail: j.kittler@surrey.ac.uk
ORCID iD: 0000-0002-8110-9205