Information Sciences 606 (2022) 864–876

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Spatial-temporal interaction learning based two-stream network for action recognition

Tianyu Liu [a], Yujun Ma [b], Wenhan Yang [a], Wanting Ji [c], Ruili Wang [b],*, Ping Jiang [a],*

[a] College of Mechanical and Electronical Engineering, Hunan Agricultural University, Changsha, China
[b] School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand
[c] School of Information, Liaoning University, Shenyang, China

ARTICLE INFO

Article history:
Received 29 January 2022
Received in revised form 8 April 2022
Accepted 26 May 2022
Available online 28 May 2022

Keywords:
Action recognition
Spatial-temporal
Two-stream CNNs

ABSTRACT

Two-stream convolutional neural networks have been widely applied to action recognition. However, two-stream networks are usually adopted to capture spatial information and temporal information separately, which normally ignore the strong complementarity and correlation between spatial and temporal information in videos. To solve this problem, we propose a Spatial-Temporal Interaction Learning Two-stream network (STILT) for action recognition. Our proposed two-stream (i.e., a spatial stream and a temporal stream) network has a spatial–temporal interaction learning module, which uses an alternating co-attention mechanism between two streams to learn the correlation between spatial features and temporal features. The spatial–temporal interaction learning module allows the two streams to guide each other and then generates optimized spatial attention features and temporal attention features. Thus, the proposed network can establish the interactive connection between two streams, which efficiently exploits the attended spatial and temporal features to improve recognition accuracy. Experiments on three widely used datasets (i.e., UCF101, HMDB51 and Kinetics) show that the proposed network outperforms the state-of-the-art models in action recognition.

1. Introduction

Action recognition is a fundamental task in computer vision, which is to classify a human action in a video into a predefined category. Action recognition empowers many real-world applications such as human–machine interaction, entertainment, visual surveillance, video retrieval, and the autonomous driving car [5,7]. Different from static images, videos usually contain cluttered backgrounds, camera motions, illumination conditions and intra/inter-class variations [8,9]. Due to significant camera motions and background noises, motion features cannot be accurately extracted [10–12,22]. Other object-related issues such as pose variations and movement speeds can also be the challenges that prohibit most action recognition approaches from being used in outdoor practical environments [15].

Conventional action recognition approaches usually contain two modules which are a feature representation module and a classifier [37]. The feature representation module converts a video into a series of feature vectors and the classifier infers an action class from the generated feature vectors [37]. Hand-crafted feature representation-based approaches [27,37,40] predefine the parameters by experts. Although hand-crafted feature representation-based approaches [27,37] have achieved

* Corresponding authors.
E-mail addresses: ruili.wang@massey.ac.nz (R. Wang), 5229889@qq.com (P. Jiang).

https://doi.org/10.1016/j.ins.2022.05.092
0020-0255/© 2022 Elsevier Inc. All rights reserved.

great success in action recognition, these approaches required significant experts works and professional knowledge. However, this approach [37] cannot capture long-term duration information in videos. To capture features for motion trajectories, Yan et al. [44] developed a hierarchical spatiotemporal approach for video classification, which captured trajectories by matching corresponding scale-invariant feature transform points on adjacent frames. In addition, this kind of approach does not show good accuracy on large action recognition datasets.

Deep neural network based approaches [13,19,40,45,48,50] have obtained high attention due to the powerful ability to extract discriminative features with high abstraction and invariance from large-scale raw data. Recent deep neural network-based approaches [15,23,36,49,50] have achieved highly accurate performances on public action recognition datasets.

Recently, a two-stream convolutional neural network [30] has achieved a great performance in action recognition, which contains two separate streams (i.e., a spatial stream and a temporal stream). The spatial stream learns spatial information from consecutive RGB video frames, while the temporal stream learns motion information from consecutive optical flow frames [30]. The classification result is obtained via the fusion of the respective softmax layer. Inspired by [30], some approaches [33,35,38] proposed to adopt two independent networks to extract spatial and temporal information in videos. Although achieving better accuracy than [30], these approaches [33,35,42] learn spatial and temporal features separately. However, the natural relationship between spatial and temporal information offers complementary clues to represent the same action category [30]. Thus, the spatial and temporal information should be modelled collaboratively to boost feature learning, and an interaction method should be utilized between two streams.

The attention mechanism is first proposed for neural machine translation [1] and has become a well-known concept, which is a beneficial tool in deep neural networks. The attention mechanism is inspired by how humans focus on different areas of an image or correlated words in a sentence [1,11]. Attention mechanisms have been widely introduced to machine translation, video processing, speech classification, and many other tasks [3,16–18].

To address the issue of the lack of interaction between the two streams, we propose a spatial–temporal interaction learning two-stream network for action recognition. Our proposed network contains two streams, which extract frames features and optical flow features respectively. As shown in Fig. 1, the spatial–temporal interaction learning module is proposed to establish the interaction between two streams to generate optimized spatial attention features and temporal attention features by utilizing an alternating co-attention operation. In other words, the proposed network can jointly perform optical flow guided attention and RGB video frames guided attention.

The contributions of this paper are summarized as follows:

- We propose a novel Spatial-temporal Interaction Learning Two-stream network (STILT) for action recognition to jointly extract discriminative and complementary spatial and temporal features to enhance feature modelling.

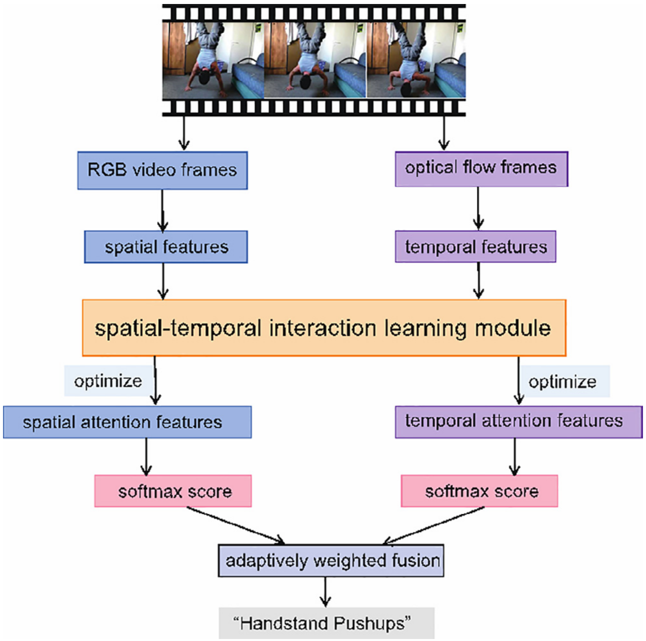

Fig. 1. The structure of the proposed spatial–temporal interaction learning two-stream (STILT) model for action recognition.

- Three standard action recognition benchmarks (i.e., UCF101, HMDB 51 and Kinetics-400) have been used for comparison experiments to display the significant performance of the proposed STILT network. The experimental results demonstrate that the STILT network achieves comparable results compared with several state-of-the-art approaches.

The remaining paper is organized as follows. In Section 2, various approaches related to ours are reviewed. Our proposed STILT network is described in Section 3. Experimental results and analyses are shown in Section 4. Finally, we make conclusions in Section 5.

2. Related works

The related works are introduced in this section. The existing action recognition approaches based on two-stream networks are reviewed in Section 2.1. Then the attention mechanism-based approaches are reviewed in Section 2.2.

2.1. Deep features based approaches for action recognition

Recently, deep learning has presented its strong ability in feature representation, which can model the discriminative and robust features in computer vision. Based on the different structures, we summarize the deep learning-based approaches into three categories, which are multi-stream networks, spatial–temporal networks, and hybrid networks. The summarizations of these three categories are shown in Table 1.

The multiple stream networks [7,9,10,33,35,39,46,49] adopt multiple CNNs to extract features from different modalities (e.g., RGB frames, optical flow frames, saliency maps and audio) in videos. Simonyan et al. [30] proposed a successful work based on the two-stream hypothesis, which contains two separate streams that learn features from RGB frames and flow frames respectively. The two-stream CNN [30] fused the prediction of the two streams directly, which ignored the long-term motion information in videos. Thus, Girdhar et al. [12] developed a spatial–temporal aggregation model for action recognition, which achieved video-level aggregated features to capture long-term dependency. The proposed pooling layer (ActionVLAD) could aggregate convolutional features among different video frames to represent the entire video. Later, Diba et al. [5] proposed a deep temporal linear encoding approach for video classification, which aggregated spatial features and temporal features from the entire video and then encoded the aggregation information into a compact low dimensional feature representation. However, the feature encoding module ignored the spatiotemporal information in videos. Thus, Duta et al. [8] proposed a spatial–temporal vector-based approach for action recognition, which built a deep feature representation module to incorporate the CNN features on the entire video. The proposed max-pooling layer organized features based on the correlation of features and then performed max-pooling on the positions of features to implement spatial–temporal encoding.

In the conventional two-stream CNN [30], there was a gap between the output of the spatial network and that of the temporal network, which indicated the two-stream CNN was incapable of distinguishing the different contributions of each stream to the final prediction. The accuracy of a temporal stream was better than the accuracy of a spatial stream. Therefore, Feichtenhofer et al. [10] developed a fusion function-based model for action recognition, which inserted fusion layers between two streams to make full use of temporal information. Later, Feichtenhofer et al. [9] proposed a spatio-temporal multiplier network for video classification, which added residual skip connections from the spatial network to the temporal network and allowed spatial features to be multiplicatively scaled by temporal features. Tran et al. [33] developed a flow-guided two-stream model for action recognition, which injected a cross-linked layer from the output of the temporal stream to the input of the spatial stream. The injected layer could help the spatial stream to concentrate on the saliency area of the action rather than the background.

Apart from video frames and optical flow, Wu et al. [41] further exploited audio information to enhance the action recognition accuracy. Wang et al. [35] developed a three-stream CNN for video classification, which consists of a spatial stream, a local temporal stream and a global temporal stream that learns frame features, flow features and motion stacked difference image (MSDI) features respectively. Later, Wang et al. [39] developed a four-stream model for video classification, which contains a video saliency stream, a spatial stream, a temporal stream, and an audio stream.

Spatial-temporal networks [5,6,13,19,28,50] were straightforward extensions of 2D CNNs as they extract temporal features by 3D convolutions. Du et al. [6] proposed a C3D model for action recognition, which only contains 5 convolutional

Table 1
Summarizations of deep networks.

Categories	Summarizations
Multi-stream networks	Utilize multiple CNNs to extract features from different modalities (e.g., RGB frames, optical flow frames, saliency maps and audio) in videos.
Spatial-temporal networks	Utilize 3D CNNs as the backbone of feature extractor, which considers both spatial and temporal information.
Hybrid networks	Adopt LSTM behind CNNs to aggregate temporal information as a hybrid network, which captures long-range dependencies.

layers, 5 max-pooling layers, 2 fully connected layers, and a softmax loss layer. One main limitation of 3D CNN is that they usually have a mass of parameters, which are constrained by the machine storage limit and computing abilities.

Another way to extract temporal features is to adopt recurrent neural networks (RNNs) after the CNNs, such as LSTM, to build hybrid networks [2,4]. These hybrid networks take the advantage of both CNNs and RNNs and thus capture long-range dependencies over video frames. Donahue et al. [4] adopted LSTM to learn the time series of frame features extracted by 2D CNNs. The recurrence properties of LSTM allowed the model to generate textual descriptions in variable lengths and recognize the actions.

However, these approaches process the feature representation of spatial and temporal information independently, which ignores the correlations between them. According to [30], the spatial information from video frames and temporal information from frame sequences are both essential in the video, which are two complementary factors to symbolize an action class. To help the spatial information and temporal information to boost each other and make use of the complementarity between spatial and temporal information, we propose a spatial–temporal interaction learning module. The proposed module can collectively extract the spatial and temporal features by the feature interaction, as well as enhance the feature representation.

2.2. Attention mechanisms

The attention mechanism was first developed for machine translation, which could allow a network to focus on the key information in each set of inputs. The calculation of the attention mechanism was quantified via learned weights and the results were generally obtained as a weighted average [1]. Bahdanau et al. [1] developed a soft attention mechanism-based model for neural language processing, which could learn the alignment weights between raw texts and target words. Cheng et al. [2] proposed a self-attention/intra-attention mechanism-based approach for machine-reading, which learned the correlation between the current words and the previous part of the sentence. The self-attention mechanism has been broadly applied in machine reading, abstractive summarization and image description generation [1–3,7,17].

Benefitting from the inherent advantages of attention mechanisms for learning sequential data, many attention mechanism-based approaches are proposed for action recognition. Sharma et al. [29] developed a visual attention mechanism-based model for action recognition, which could essentially learn which parts of the video frames were related to the action recognition and attached greater attention to them. Yan et al. [44] proposed a hierarchical multi-scale attention-based approach for action recognition, which could capture the most relevant information from the inputs and enhance the relationship between the inputs and outputs.

Later, Du et al. [7] developed a recurrent spatial–temporal attention-based approach for video classification, which could concentrate on the key action regions around the objects in video frames. A spatial–temporal attention-based approach [41] is proposed for action recognition, which could learn the deep feature representation and characterize the efficient information at both RGB frames and flow frames, respectively. Dai et al. [3] proposed an attention-based two-stream based approach for action recognition, which could focus on the key information of the video frames and pay different degrees of attention to the outputs of each CNN feature map.

Ji et al. [18] developed a cross-attention siamese model for video salient object detection, which integrated a bi-direction cross-attention module into an encoder-decoder mode under the siamese network. This type of bi-direction cross-attention mechanism tries to extract saliency coherency based on the spatial–temporal pixelwise relation between two sequential RGB video frames. This is different from the co-attention mechanism in our proposed spatial–temporal interaction learning module, which aims to learn the correlation between RGB video frames and the optical flow frames over the same action. Moreover, cross-attention was adopted in a siamese structure [18], which processed the same input modality and calculated by the element-wise production between the attended features of frame t and frame t + 1. This is different from the input features of co-attention in our two-stream network, which confronts two different types of input features (i.e., frames features and flow features) and sequentially alternates between generating frame and flow attentions.

However, the existing attention mechanism-based two-stream approaches mainly focus on key information in video frames, while ignoring the coexistence relationship between two streams. Therefore, we propose a spatial–temporal interaction learning two-stream network (STILT) for action recognition, which establishes an interaction method between two streams by adopting an alternating co-attention operation.

2.3. Weighted score fusion

In recent years, different fusion strategies have shown remarkable performances in many tasks such as video retrieval, speech recognition, and cryptosystems. The fusion strategies were largely focused on three levels, which are the feature level, score level, and decision level. According to [43], the classifier-based score fusion had more power to exploit the multisource information among these three strategies. Xu et al. [43] developed an adaptive weighted fusion approach for image classification, which could automatically determine optimal weights. Li et al. [24] proposed an adaptive weighted fusion-based semi-supervised learning model for image annotation, which used the adaptive weighted fusion to evaluate the reliability of the pseudo-labels of the unlabeled data thus the new samples can be chosen more dependably. Later, Wang et al. [34] adopted an adaptive weighted fusion strategy for object detection to learn key features of channels while dropping redundant information and outstanding target features. Inspired by these fusion strategies, we propose to adaptively learn different softmax score fusion weights of spatial and temporal information for each action class to evaluate different contributions of spatial information and temporal information to improve the performance.

3. Proposed STILT model

We present the details of our proposed spatial–temporal interaction learning two-stream (STILT) network, which contains two identical networks: a spatial network and a temporal network. As shown in Fig. 2, the input videos are first divided into RGB video frames and optical flow frames for modelling spatial and temporal features. The spatial–temporal interaction learning module utilizes an alternating co-attention mechanism to help spatial information and temporal information to guide each other. Finally, we utilize an adaptively weighted fusion to fuse the softmax score of each stream for different video categories.

In Section 3.1, we present the overall architecture of our proposed STILT network. The detailed interaction learning module is introduced in Section 3.2. The adaptively weighted fusion method is presented in Section 3.3.

3.1. Our proposed STILT

Video usually includes spatial and temporal dimensions, which are two complementary cues to represent the specific action class. As shown in [12], independently modelling spatial and temporal features is not sufficient to completely capture important clues offered by spatial and temporal information for action recognition. Therefore, a spatial–temporal interaction learning module is proposed to take the spatial and temporal features extracted by CNNs as input and processes interaction learning in an alternate mode to explore the complementary cues between spatial and temporal features. The proposed module adopts an alternating co-attention mechanism and allows the spatial stream and temporal stream to guide each other, which generates spatial attention features and temporal attention features.

As shown in Fig. 2, the proposed model is constructed based on a residual network in which we make several modifications to extract features in videos. The proposed model consists of two separate streams (i.e., a spatial stream and a temporal stream), which take RGB video frames and consecutive optical flow frames as inputs, respectively. We add a spatial–temporal interaction learning module after the last pooling layer in the original ResNet-50. The proposed spatial–temporal interaction learning module also consists of an alternating co-attention block, two fully connected layers and two softmax loss layers. Then, we adopted an adaptively weighted fusion layer to take the generated softmax score as the input to compute the final prediction score.

3.2. Spatial-temporal interaction learning module

As shown in Fig. 2, we propose a spatial–temporal interaction learning module to take the extracted spatial and temporal features as input and processes interaction learning in an alternate mode to explore the complementarity between them. The module is consisting of an alternating co-attention block, two fully connected layers, and two softmax layers. Specifically, the

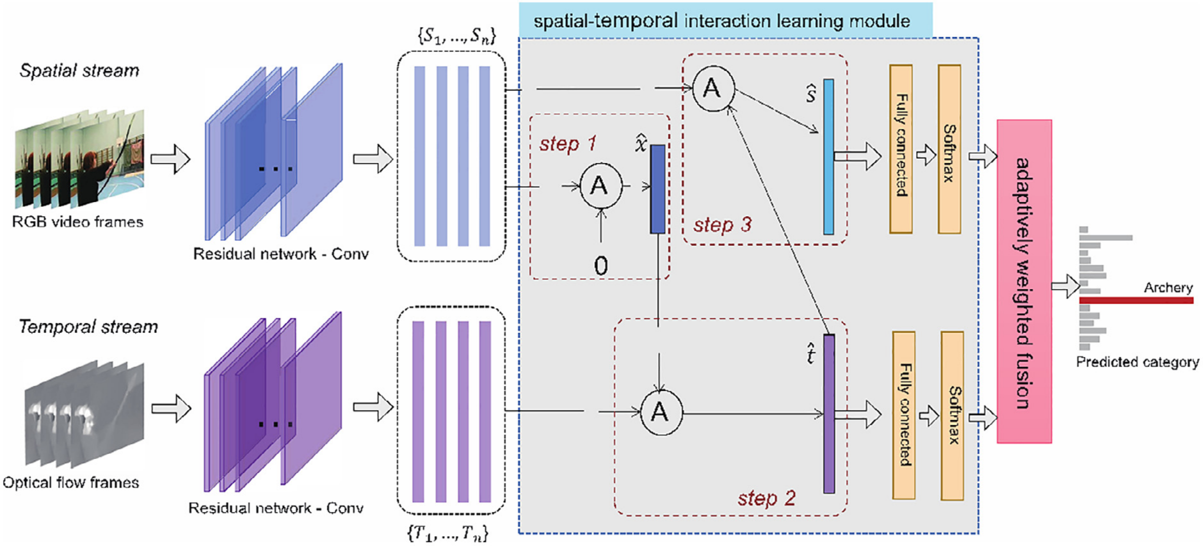

Fig. 2. The architecture of the proposed spatial–temporal interaction learning two-stream (STILT).

alternating co-attention layer is utilized to accomplish the alternate optimizations and then generate spatial attention features and temporal attention features. The two softmax layers are utilized to output classification scores of the spatial stream and the temporal stream. As shown in Fig. 2, the spatial–temporal interaction learning module allows the learned spatial information to perform the optimization of temporal information, and vice versa.

In the spatial–temporal interaction learning module, we sequentially alternate between spatial features and temporal features. Specifically, the operation includes three steps which are marked in Fig. 2:

- Summarizing the spatial features into a single vector.
- Attending to the temporal features based on the summarized vector.
- Attending to the spatial features based on the attended temporal features.

Specifically, in frame h, the spatial features $S = \{S_1, \cdots, S_n\}$ are used to perform the optimization of temporal features $T = \{T_1, \cdots, T_n\}$. An attention operation $M = \mathscr{A}(K; g)$, is defined, which takes the spatial features and temporal features K and attention guidance g, derived from frames or optical flows as inputs, and outputs the attended spatial or temporal feature vector. By adopting the spatial–temporal interaction learning module, we obtain:

$$H = tanh\left(W_k K + (W_g g)1^T\right), \tag{1}$$

$$a^k = softmax\left(w_{hk}^T H\right), \tag{2}$$

$$O = \sum a_i^k k_i, \tag{3}$$

where 1 denotes a vector with all elements of 1. W_k, $\mathrm{W_g}$ and w_{hk} are the weight parameters; a^k stands for the learned attention weights of feature matrix K. The structure of the alternating co-attention layer is shown in Fig. 2:

- In step 1, $K = S$, and g is 0.
- In step 2, $K = T$, and T is the temporal features, and the guidance g is the intermediate attended spatial feature $\widehat{x}$ from step 1.
- In step 3, $K = S$ and $g = \widehat{t}$, we use the attended temporal $\widehat{t}$ as the guidance to attend to the spatial features again.

The proposed spatial–temporal interaction learning module obtains the optimization of spatial features and temporal features by learning the powerful complementary clues between spatial information and temporal information. The optimization operation in the proposed spatial–temporal interaction learning module is summarized in Algorithm 1.

Algorithm 1: Spatial-temporal interaction learning module.

```
Input: The spatial features and temporal features learned from the RGB frames and flow frames, which are indicated as
    $S = \{S_1, \cdots, S_n\}$ and $T = \{T_1, \cdots, T_n\}$, respectively.
Output: The optimized (attended) spatial features $\widehat{s}$ and temporal features $\widehat{t}$.
Initializing optimization parameters on spatial features as $a^s$ and all the $k$ elements are set to $1/k$.
Repeat.
    Summarizing the spatial features $S$ into a matrix $L_s = \sum \mathrm{a_i^s s_i}$.
    Optimizing the temporal features applying $L_s$ by Equ. (1) and (2) and obtaining the optimization parameters $a^t$ on the temporal features.
    Summarizing the temporal features $T$ into a matrix $L_T = \sum \mathrm{a_i^t t_i}$
    Optimizing the spatial features using $L_T$, and obtaining the optimization parameters $a^s$ on the spatial features.
Until. The loss functions converge.
Return. The spatial attended features $\widehat{s} = S^{\mathrm{T}} a^s$, and the temporal attended features $\widehat{t} = T^{\mathrm{T}} a^t$.
```

3.3. Adaptively weighted fusion

According to [28], spatial information and temporal information make different contributions to each action class. For example, some action classes do not show salient motion areas, such as “laugh” and “chew”, which can be recognized only from spatial information. There are some action classes (e.g., “walk” and “jump”) that have a salient motion area and temporal information is important for recognizing them. Thus, instead of simply summing up the softmax prediction scores of the spatial stream and the temporal stream, we use adaptively weighted fusion to perform the fusion weights of the spatial stream and the temporal stream for different action classes.

Specifically, we define the softmax score of the x-th training data in the y-th action class as:

$$S^y_{x,1} = \left[{S^y_{x,1}}^T, {S^y_{x,2}}^T\right]^T \in \mathbb{R}^{2\times N}, \tag{4}$$

where N refers to the number of action classes. $S^y_{x,n} \in \mathbb{R}^{1\times N}$ indicates the softmax score of the n-th stream for x-th training data in the y-th class. The fusion weight matrix for the y-th class is $W_y = \left[w_{y,1}, w_{y,2}\right]$. The fusion weight matrix of each class is modelled independently and to obtain the W_y, the objective function is computed as:

$$argmax(W_y) = \mathrm{P}_y - \lambda M_y, \tag{5}$$

where P_y is defined as:

$$\mathrm{P}_y = \sum_{x=1}^{m_y} W_y S^y_x Y_y. \tag{6}$$

where m_y denotes the number of training data in the y-th class. $Y_y = [0, \ldots, 1, \ldots 0] \in \mathbb{R}^{N\times 1}$ means the y-th element is 1, and the others are 0. To maximize P_y, the dot product of W_y and the y-th column vector of matrix S^y_x should be maximized. The M_y is defined as:

$$M_y = \sum_{\{k=1,k\neq y\}}^{C} \sum_{x=1}^{m_k} W_y S^k_x Y_y, \tag{7}$$

which means to maximize the dot product of the matrix Y_y and the y-th column vector of S^k_x. The goal of P_y and M_y is to consider the relationship between correctly and non-correctly training samples for W_y. And λ denotes the parameter which can balance the weight of correctly and non-correctly training samples. The fusion weight is learned by linear programming easily.

When dealing with test data, the softmax score of the spatial and temporal streams are calculated and stacked at first, which is referred as:

$$S_t = \left[S^T_{t,1}, S^T_{t,2}\right]^T \in \mathbb{R}^{2\times C}. \tag{8}$$

Also, the prediction of different classes is computed by:

$$argmax(x) = W_x S_t Y_x. \tag{9}$$

For each action class, different fusion weights are considered. With adaptively weighted fusion for the spatial and temporal streams, the different contributions of spatial and temporal information for each action category can be distinguished. The final classification accuracy is decided by the highest fusion score.

4. Experiments

We evaluate our proposed spatial–temporal interaction learning two-stream (STILT) model on three widely used action recognition benchmarks (i.e., UCF101, HMDB51 and Kinetics). Besides, we will compare our proposed STILT with several state-of-the-art action recognition models containing the two-stream CNN based models and 3D CNN based models. Finally, we conduct several ablation experiments to analyze the proposed STILT model.

4.1. Datasets

UCF-101. UCF-101 [32] is a popular action recognition benchmark that includes 13 k video clips from 101 action classes (e.g., body motion, human–human interaction, and human-objects interaction). The video clips in UCF101 are selected from YouTube which includes large variations in scales, lights, backgrounds, scales, and viewpoints. For the splits of training data and test data, we inherit the provided three train-test splits. In the experimental results, we report the final average recognition accuracy over the three splits.

HMDB-51. HMDB-51 [21] is a challenging benchmark for action recognition which includes 6 k video clips categorized into 51 human action classes. HMDB51 provided three splits for training and testing. The final result is obtained by the average of the classification results over the splits of training data and test data.

Kinetics. Kinetics 400 [20] contains around 240 K training videos and 20 K validation videos which are cropped into 10 s in 400 action classes. We report top-1 recognition results on the validation datasets. Three examples of these three datasets are shown in Fig. 3.

4.2. Implementation details

This subsection presents the implementation details of the proposed STILT model architecture and experimental training details.

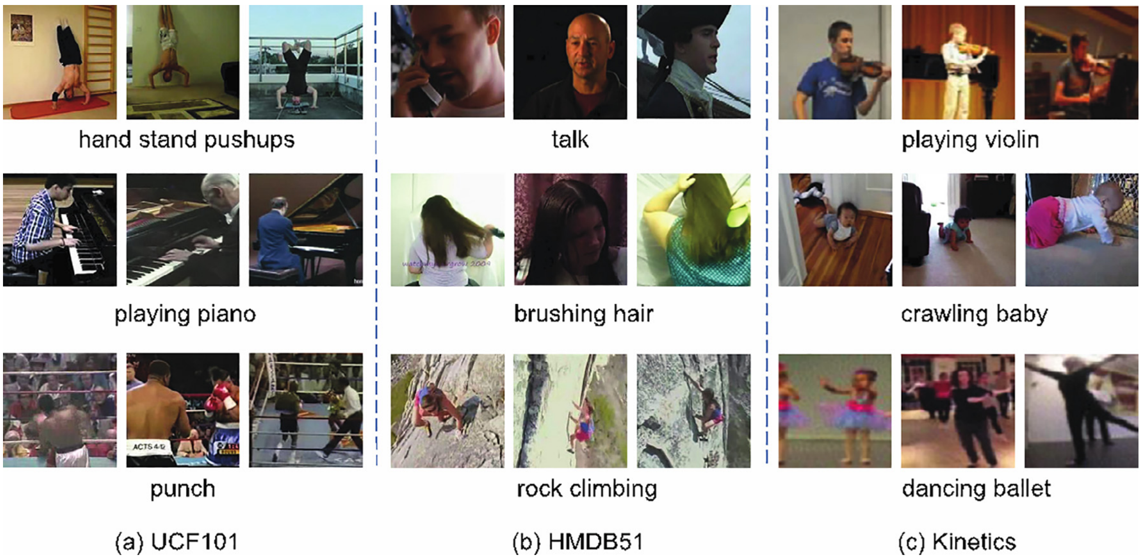

Fig. 3. Three examples of the HMDB51, UCF101 and Kinetics datasets.

Firstly, we utilize ResNet-50 [14] as a feature extractor on both spatial stream and temporal stream due to the wide applications and remarkable performance of ResNet-50 in action recognition as reported in conventional models [14]. At the beginning of the temporal stream, the optical flow frames are pre-obtained by utilizing the TVL1 method [47] and the frames are stored as JPEG images. The default optical flow stacked frame number L is set to 2 × 10 following [30]. The size of input RGB frames and flow frames is cropped into 224 × 224.

Table 2 shows the feature extractor network which is constructed based on ResNet-50 [14]. The input sizes of Conv1 are 224 × 224 × 3 and 224 × 224 × 20, which denotes the size of input RGB frames and input optical flow frames, respectively. An average pooling layer is utilized as a feature output layer. The final dimensions of both extracted spatial and temporal features are 2048. The proposed spatial–temporal interaction learning module follows the average pooling layer to optimize the features.

As shown in Fig. 3, the proposed spatial–temporal interaction learning module consists of an alternating co-attention layer, two fully connected layers and softmax layers. The detailed architecture of the spatial–temporal interaction learning

Table 2
The detailed architecture of the backbone network. The down sampling is obtained by Conv3_1, Conv4_1, and Conv5_1.

Layers	Input size	Block information	Output size
Conv1	224 × 224 × 3 224 × 224 × 20	Kernel: 7 × 7 Feature map number:64 Stride 2	112 × 112 × 64
Max-Pooling1	112 × 112 × 64	Kernel: 7 × 7 Stride 2	56 × 56 × 64
Conv2_x	56 × 56 × 64	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64 \\ 1\times1,256 \end{bmatrix}\times 3$, stride 1	56 × 56 × 256
Conv3_x	56 × 56 × 256	$\begin{bmatrix} 1\times1,128 \\ 3\times3,128 \\ 1\times1,512 \end{bmatrix}\times$ 4For the first convolutional layer: stride 1 For other convolutional layers: stride 2	28 × 28 × 512
Conv4_x	28 × 28 × 512	$\begin{bmatrix} 1\times1,256 \\ 3\times3,256 \\ 1\times1,1024 \end{bmatrix}\times 6,$ For the first convolutional layer: stride 1 For other convolutional layers: stride 2	14 × 14 × 1024
Conv5_x	14 × 14 × 1024	$\begin{bmatrix} 1\times1,512 \\ 3\times3,512 \\ 1\times1,2048 \end{bmatrix}\times 3,$ For the first convolutional layer: stride 1 For other convolutional layers: stride 2	7 × 7×2048
Average pooling	7 × 7×2048	Kernel: 7 × 7 Stride 1	1 × 1×2048

module is shown in Table 3. Specifically, the spatial–temporal interaction learning module takes spatial features and temporal features as inputs and optimizes them by a sequentially alternate method which is introduced in Algorithm 1.

In the training process, we use the Stochastic Gradient Descent (SGD) algorithm with momentum ($m = 0.9$) to train our proposed model, the batch size is set to 128 and the weight decay is set to 5×10^{-4} following [30]. We randomly generate training samples from video frames in training data for data augmentation [9]. The initial learning rate is set to 10^{-3} and is decreased with $\gamma = 0.1$ every 10 K iterations for the UCF-101 dataset and the Kinetics dataset and every 5 K iterations for the HMDB-51 dataset. All the training stages are stopped after 50 k iterations.

4.3. Model analysis and discussion

Our STILT approach consists of a spatial stream and a temporal stream. We represent the two streams and unite them as "Spatial", "Temporal" and "Two-stream", respectively. The ablation experiments are conducted to check the effect of the spatial–temporal interaction learning module and adaptively weighted fusion, respectively. In Table 3, "STIL" denotes spatial–temporal interaction learning module and "AWF" denotes adaptively weighted fusion.

Comparing the results of "Spatial", "Temporal" and "Two-stream", the video frame features, and the optical flow features are complementary in action recognition. Comparing the results of "Two-stream" and "Two-stream + STIL", we can find that our proposed spatial–temporal interaction learning module can advance the modelling of spatial and temporal features mutually and model the complementarity between the spatial information and the temporal information thus further improving action recognition accuracy. In detail, the STIL module achieves promotions of 2.6% on UCF101, achieves promotions of 6.4% on HMDB51, and achieves promotions of 2.5% on Kinetics.

Comparing the results of "Two-stream + STIL" and "Two-stream + STIL + AWF", we can observe that "Two-stream + STIL + AWF" achieve a higher classification result, which proves that it is useful to adopt the adaptively weighted fusion for spatial stream and temporal stream. In detail, the AWF method achieves promotions of 1.5% on the UCF101 dataset, achieves promotions of 3.2% on the HMDB51, and achieves promotions of 3.6% on the Kinetics dataset.

As shown in Table 4, "Two-stream + STIL + AWF" achieves the highest accuracies among the five settings, which proves that the proposed spatial–temporal interaction learning module and adaptively weighted fusion can optimize the feature representation for action recognition.

We further conduct the experiments to show the effectiveness of adaptively weighted fusion by applying different fusion strategies, which are late average fusion and early fusion. Late average fusion averages the softmax prediction scores of the spatial and the temporal stream, which is shown as "Late Ave" in Table 5. Early fusion concatenates spatial features and temporal features and trains an SVM classifier for final prediction, which is denoted as "Early" in Table 5. In Table 5, "Two-stream + STIL + Late Ave" refers to applying spatial–temporal interaction learning module, then fused by late average fusion method.

Compared to the results in Table 5, late average fusion and early fusion methods show lower results, because they do not have the ability to distinguish different contributions of spatial and temporal features for different types of action classes. "Two-stream + STIL + AWL" achieves the best recognition results among all three settings, because it can distinguish different contributions of spatial and temporal information for each action class.

4.4. Comparisons with state-of-the-art models

We compare the top-1 accuracy of our proposed STILT model with several state-of-the-art models over the three datasets (e.g., UCF101, HMDB51, and Kinetics). Several pre-trained image classification CNNs are applied as the backbone network (e.g., VGG-19 [31], ResNet-34 [14], and ResNet-50 [14]). The compared results are reported in Table 6. The ResNet-50 based model achieves the best results because of the deep convolutional layers. The compared models include the two-stream models [10,37,38], the 3D CNN models [6,28] and other very recent models [7,32,49]. The classification results are averaged over the three splits.

We can observe that the fusion strategy in [10] is very simple. However, the powerful complementarity between spatial and temporal information is still ignored. Besides, they miss the attended features, thus the features extracted by these models are not discriminative enough to recognize actions. Compared to the results of the compared models, our proposed STILT achieves the best accuracy among these state-of-the-art models on two datasets (i.e., 72.1% on HMDB51, and 72.4% on Kinetics). The proposed STILT also achieves 95.6% accuracy on UCF101, which only 0.3% less than AMFNet-C [25]. This is probably because AMFNet fused the RGB frames and flows frames before the main feature learning block. Our proposed STILT model

Table 3
The detailed architecture of the spatial–temporal interaction learning module.

Layers	Input size	Output size
Alternating co-attention layer	2048	512
Fully connected layer	512 × 1024	–
Softmax	–	1024 × the number of classes in the corresponding dataset

Table 4
Experimental results of the spatial–temporal interaction learning module. The accuracies of model “Two-stream” and model “Two-stream + STIL” are achieved by late average fusion.

Models	UCF101 (%)	HMDB51 (%)	Kinetics (%)
Spatial	81.2	51.3	53.8
Temporal	83.5	52.6	56.1
Two-stream	91.5	62.5	66.3
Two-stream + STIL	94.1	68.9	68.8
Two-stream + STIL + AWF	**95.6**	**72.1**	**72.4**

Table 5
Experimental results on UCF101 and HMDB51 with different fusion strategies.

Methods	UCF101 (%)	HMDB51 (%)	Kinetics (%)
Two-stream + STIL + Late Ave	94.1	68.9	68.0
Two-stream + STIL + Early	94.2	68.4	68.1
Two-stream + STIL + AWF	**95.6**	**72.1**	**72.4**

Table 6
Comparisons with the state-of-the-art models on UCF101, HMDB51, and Kinetics.

Models	Input Modalities	UCF101 (%)	HMDB51 (%)	Kinetics (%)
Girdhar *et al.* [11]	RGB + others	–	52.2	–
Meng *et al.* [26]	RGB + others	87.1	53.1	–
C3D [6]	RGB only + 3D CNNs	77.4	46.7	56.1
P3D-199 [28]	RGB only + 3D CNNs	89.2	62.9	–
Two-stream CNN [30]	RGB + optical flow	88.0	59.4	61.0
TDD [37]	RGB + optical flow	90.3	63.2	–
Two-stream fusion [10]	RGB + optical flow	92.5	65.4	–
TSN [38]	RGB + optical flow	94.2	69.4	–
TSN Corrnet [46]	RGB + optical flow	94.4	70.6	–
MSM-ResNets [49]	RGB + optical flow + others	93.5	66.7	–
ARTNet-Res18 [32]	RGB only	94.3	70.9	69.2
AMFNet-C [25]	RGB + optical flow	**95.9**	71.2	71.6
RSTAN (TSN) [7]	RGB + optical flow	94.6	70.5	**–**
STILT(VGG-19)	RGB + optical flow	89.9	66.3	64.8
STILT(ResNet-34)	RGB + optical flow	93.7	69.3	68.4
STILT (ResNet-50)	RGB + optical flow	95.6	**72.1**	**72.4**

can learn the powerful complementarity between video frame features and optical flow features and boost the learning by alternating co-attention. Besides, the adaptively weighted fusion can distinguish different contributions of spatial and temporal information to different action classes.

We can observe that the proposed STILT model achieves relatively lower results on the HMDB51 dataset compared with the results on UCF101. The HMDB51 dataset is very challenging because of the various lights, uncertain camera motion, and poor resolution, which leads to high intra-class variations. We report several correctly and non-correctly classified video examples of the proposed STILT model on the HMDB51 dataset in Fig. 4. We can observe the low resolution of videos and the similarity of these actions make the classification very difficult. However, our proposed STILT model shows the highest result even on the HMDB51 dataset among 12 state-of-the-art networks, which proves its universality and effectiveness.

As shown in Fig. 5, we report the confusion matrix on the top-20 most frequent action classes in the HMDB51 dataset, in which the columns and rows denote the action classes in HMDB51. Any classes outside the top-20 are grouped into a superclass labelled “other”. We can observe that our STILT model achieves remarkable performance for most actions including human–human interaction and self-body motion.

5. Conclusion

In this work, we propose a spatial–temporal interaction learning two-stream model for action recognition, which consists of a spatial–temporal interaction learning module, a spatial stream, a temporal stream, and an adaptively weighted fusion layer. The spatial stream extracts the spatial features in videos from RGB frames and the temporal stream extracts the temporal features in videos from optical flow frames. Then, the proposed spatial–temporal interaction learning module optimized both the spatial and temporal features and mutually boosted them as discriminative features for action recognition. Experiments on three popular action recognition datasets indicate that the STILT model performs the best result among over twelve state-of-the-art models.

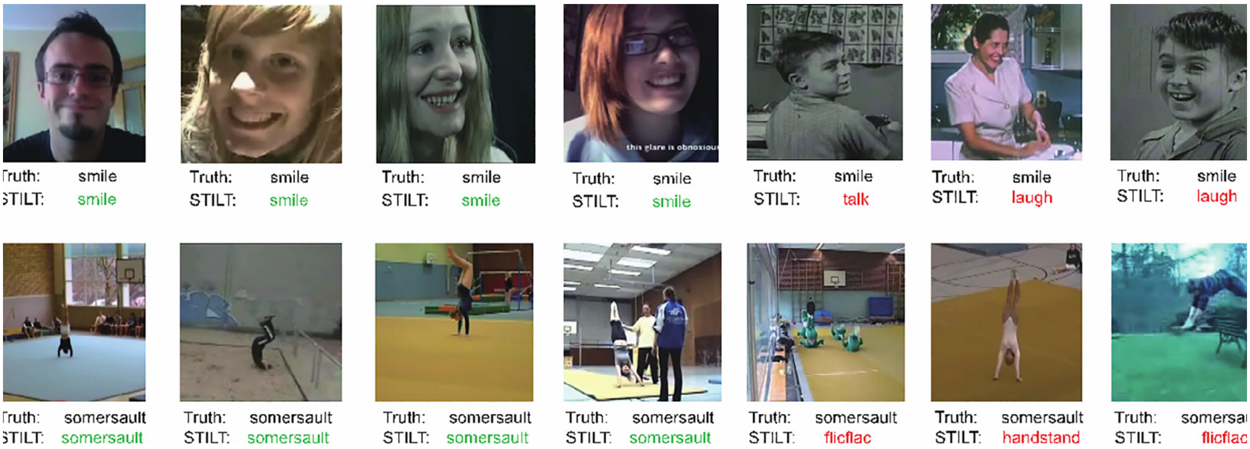

Fig. 4. Several samples of correctly and non-correctly recognition results on HMDB51. The first row lists the label (truth) of the videos, the second row shows the classified results obtained by our STILT model.

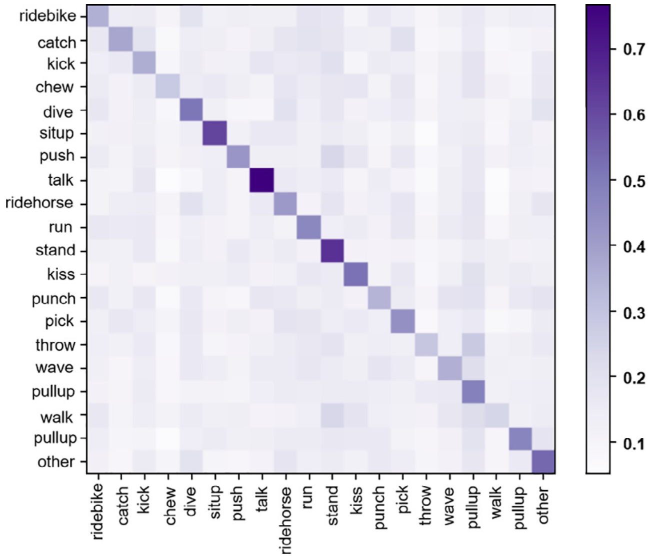

Fig. 5. The performance of our STILT model on the top-20 most frequent action categories on HMDB51. Any categories outside the top-20 are grouped into a super-class labelled “other”.

The future works will focus on two aspects. Firstly, we will concentrate on developing more efficient co-attention mechanisms and proposing more discriminative spatial–temporal representation methods. Secondly, we will utilize unsupervised learning and semi-supervised learning in the proposed model, which will promote the use of massive unlabeled videos on the Internet.

CRediT authorship contribution statement

Tianyu Liu: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Yujun Ma:** Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Wenhan Yang:** Investigation. **Wanting Ji:** Validation. **Ruili Wang:** Writing – review & editing. **Ping Jiang:** Resources.

Declaration of Competing Interest

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

Acknowledgements

This research is supported by 2022 Science and Technology grant, the Forestry Department, Hunan Province, China.

References

[1] D. Bahdanau, C. Kyunghyun, Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).

[2] J. Cheng, D. Li, L. Mirella, Long short-term memory-networks for machine-reading, International Conference on Learning Representations, 2015.

[3] C. Dai, X. Liu, J. Lai, Human action recognition using two-stream attention-based LSTM networks, Appl. Soft Comput. 86 (2020) 105820.

[4] J. Donahue, L.A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[5] A. Diba, M. Fayyaz, V. Sharma, A.H. Karami, M.M. Arzani, R. Yousefzadeh, L.V. Gool. Temporal 3d convnets: new architecture and transfer learning for video classification. arXiv:1711.08200 (2017).

[6] T. Du, L. Bourdev, R. Fergus, Learning spatiotemporal features with 3d convolutional networks Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489-4497.

[7] W. Du, Y. Wang, Y. Qiao, Recurrent spatial-temporal attention network for action recognition in videos, IEEE Trans. Image Process. 27 (3) (2018) 1347–1360.

[8] I.C. Duta, B. Ionescu, K. Aizawa, N. Sebe, Spatio-temporal vector of locally max pooled features for action recognition in videos, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3205–3214.

[9] C. Feichtenhofer, A. Pinz, R.P. Wildes, Spatiotemporal multiplier networks for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4768–4777.

[10] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.

[11] R. Girdhar, R. Deva, Attentional pooling for action recognition, Advances in Neural Information Processing Systems, 2017.

[12] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, B. Russell, Actionvlad: Learning spatio-temporal aggregation for action classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 971–980.

[13] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and ImageNet?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp 6546–6555.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[15] Y. Hsueh, W. Lie, G. Guo, Human behaviour recognition from multiview videos, Inf. Sci. 517 (2020) 275–296.

[16] W. Ji, R. Wang, A multi-instance multi-label dual learning approach for video captioning, ACM Trans. Multimedia Comput. Commun. Appl. 17 (2s) (2021) 1–18.

[17] W. Ji, R. Wang, Y. Tian, X. Wang, An attention-based dual learning approach for video captioning, Appl. Soft Comput. 117 (2022) 108332.

[18] Y. Ji, H. Zhang, Z. Jie, L. Ma, Q.M. Jonathan Wu, CASNet: A cross-attention siamese network for video salient object detection, IEEE Trans. Neural Networks Learn. Syst. 32 (6) (2021) 2676–2690.

[19] H. Kataoka, K. Hara, R. Hayashi, E. Yamagata, N. Inoue, Spatiotemporal initialization for 3D CNNs with generated motion patterns, in: 2022 IEEE Winter Conference on Applications of Computer Vision, 2022, pp. 737–746.

[20] W. Kay, J. Carreira, K. Simonyan, The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[21] H. Kuehne, H. Jhuang, E. Garrote, HMDB: a large video database for human motion recognition, in: International Conference on Computer Vision, 2011, pp. 2556–2563.

[22] S. Lee, H. Lee, C. Shin, H. Son, S. Lee. Beyond natural motion: exploring discontinuity for video frame interpolation. ArXiv abs/2202.07291 (2022).

[23] J. Li, X. Liu, M. Zhang, D. Wang, Spatio-temporal deformable 3d convnets with attention for action recognition, Pattern Recogn. 98 (2020) 107037.

[24] Z. Li, L. Lin, C. Zhang, H. Ma, W. Zhao, Z. Shi, A Semi-supervised learning approach based on adaptive weighted fusion for automatic image annotation, ACM Trans. Multimedia Comput. Commun. Appl. 17 (2021) 1–23.

[25] S. Liu, Xin Ma. Attention-driven appearance-motion fusion network for action recognition. IEEE Transactions on Multimedia (2022).

[26] L. Meng, B. Zhao, B. Chang, G. Huang, W. Sun, F. Tung, L. Sigal, Interpretable spatio-temporal attention for video action recognition, Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[27] D. Navneet, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[28] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks proceedings of the IEEE, in: International Conference on Computer Vision, 2017, pp. 5533–5541.

[29] S. Sharma, R. Kiros, R. Salakhutdinov, Action recognition using visual attention, Advances in Neural Information Processing Systems, 2015.

[30] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (2014).

[32] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402, 2012.

[33] A. Tran, L.F. Cheong, Two-stream flow-guided convolutional attention networks for action recognition, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 3110–3119.

[34] H. Wang, L. Jiang, Q. Zhao, H. Li, K. Yan, Y. Yang, S. Li, Y. Zhang, L. Qiao, C. Fu, H. Yin, Y. Hu, H. Yu, H. Yu, Progressive structure network-based multiscale feature fusion for object detection in real-time application, Eng. Appl. Artif. Intell. 106 (2021) 104486.

[35] L. Wang, L. Ge, R. Li, Y. Fang, Three-stream CNNs for action recognition, Pattern Recognition Letter 92 (2017) 33–40.

[36] L. Wang, W. Li, W. Li, Appearance-and-relation networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1430–1439.

[37] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.

[38] L. Wang, X. Yuan, W. Zhe, Y. Qiao, D. Lin, X. Tang, L.V. Gool, Temporal segment networks: towards good practices for deep action recognition, in: European Conference on Computer Vision, 2016, pp. 20–36.

[39] L. Wang, X. Yuan, M. Zong, Multi-cue based four-stream 3D ResNets for video-based action recognition, Inf. Sci. 575 (2021) 654–665.

[40] R. Wang, M. Zong, Joint self-representation and subspace learning for unsupervised feature selection, World Wide Web 21 (2018) 1745–1758.

[41] Z. Wu, X. Wang, Y. Jiang, H. Ye, X. Xue, Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, In Proceedings of the 23rd ACM International Conference on Multimedia, pp. 461-470.

[42] K. Xu, X. Jiang, T. Sun, Two-stream dictionary learning architecture for action recognition, IEEE Trans. Circuits Syst. Video Technol. 27 (3) (2017) 567–576.

[43] Y. Xu, Y. Lu, Adaptive weighted fusion: a novel fusion approach for image classification, Neurocomputing 168 (2015) 566–574.

[44] S. Yan, J.S. Smith, W. Lu, B. Zhang, Hierarchical multi-scale attention networks for action recognition, Signal Process. Image Commun. 61 (2018) 73–84.

[45] Y. Yu, Y. Gao, H. Wang, R. Wang, Joint user knowledge and matrix factorization for recommender systems, World Wide Web 21 (2018) 1141–1163.

[46] N. Yudistira, T. Kurita, Correlation net: spatiotemporal multimodal deep learning for action recognition, Signal Process. Image Commun. (2020) 115731.
[47] C. Zach, T. Pock, H. Bischof, A duality-based approach for real-time tv-l 1 optical flow, in: Joint Pattern Recognition Symposium, 2007, pp. 214–223.
[48] H. Zheng, R. Wang, W. Ji, M. Zong, W. Wong, K. Lai, H. Lv, Discriminative deep multi-task learning for facial expression recognition, Inf. Sci. 533 (2020) 60–71.
[49] M. Zong, R. Wang, X. Chen, Motion saliency based multi-stream multiplier ResNets for action recognition, Image Vis. Comput. 107 (2021) 104108.
[50] M. Zong, R. Wang, Z. Chen, M. Wang, X. Wang, J. Potgieter, Multi-cue based 3D residual network for action recognition, Neural Comput. Appl. 33 (2020) 5167–5181.