Neural Computing and Applications (2020) 32:14593–14602
https://doi.org/10.1007/s00521-020-05144-7

S.I. : DEEP LEARNING APPROACHES FOR REALTIME IMAGE SUPER RESOLUTION (DLRSR)

Spatiotemporal saliency-based multi-stream networks with attention-aware LSTM for action recognition

Zhenbing Liu[1] · Zeya Li[1] · Ruili Wang[2,3] · Ming Zong[2,3] · Wanting Ji[2,3]

Received: 6 November 2019 / Accepted: 17 June 2020 / Published online: 7 July 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract

Human action recognition is a process of labeling video frames with action labels. It is a challenging research topic since the background of videos is usually chaotic, which will reduce the performance of traditional human action recognition methods. In this paper, we propose a novel spatiotemporal saliency-based multi-stream ResNets (STS), which combines three streams (i.e., a spatial stream, a temporal stream and a spatiotemporal saliency stream) for human action recognition. Further, we propose a novel spatiotemporal saliency-based multi-stream ResNets with attention-aware long short-term memory (STS-ALSTM) network. The proposed STS-ALSTM model combines deep convolutional neural network (CNN) feature extractors with three attention-aware LSTMs to capture the temporal long-term dependency relationships between consecutive video frames, optical flow frames or spatiotemporal saliency frames. Experimental results on UCF-101 and HMDB-51 datasets demonstrate that our proposed STS method and STS-ALSTM model obtain competitive performance compared with the state-of-the-art methods.

Keywords Spatiotemporal saliency · Multi-stream · Attention-aware · LSTM · Action recognition

1 Introduction

Human action recognition is a process of labeling video frames with action labels [10, 29, 41, 52]. It has a wide range of applications in real life such as intelligent surveillance, virtual reality (VR), video retrieval, intelligent human–computer interaction and shopping behavior analysis.

Conventional handcrafted feature-based human action recognition methods cannot fully extract efficient and robust features from videos, especially when there are complex clutter backgrounds in the videos such as target occlusion, illumination variation and camera movement. To address this challenge, deep convolutional neural network (CNN)-based human action recognition methods have been developed, which can be categorized into three categories: (i) two-stream convolutional neural network-based methods [10, 41, 50], (ii) 3D convolutional neural network-based methods [8, 16, 47] and (iii) recurrent neural network-based methods. Typically, a two-stream convolutional neural network consists of two streams: a spatial stream and a temporal stream. The spatial stream is used to capture the appearance information from a video, while the temporal stream is used to capture the motion information from the video. Different from two-stream convolutional neural networks, 3D convolutional neural networks can simultaneously learn the spatial and temporal information from multiple consecutive video frames. Recurrent neural network (RNN)-based methods can capture the temporal long-term dependency relationships between consecutive video frames, which is widely used for temporal sequence tasks such as machine translation, speech recognition, natural language processing and action

✉ Ming Zong
M.Zong@massey.ac.nz

1 School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China

2 School of Natural and Computational Sciences, Massey University, Auckland, New Zealand

3 School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China

recognition. Long short-term memory (LSTM) is a widely adopted RNN to avoid the vanishing gradient problem existing in traditional RNNs, which usually adopt hand-crafted features or CNN extracted features as inputs.

However, the above methods have not particularly considered the effect of clutter backgrounds in the videos. For human action recognition, the clutter backgrounds impose a negative effect on recognition accuracy. To solve this problem, we propose a novel spatiotemporal saliency-based multi-stream ResNets (STS for short) for human action recognition, which combines three different streams including a spatial stream, a temporal stream and a spatiotemporal saliency stream. Given a video, the spatial stream utilizes the RGB frames of the video as input, and the temporal stream utilizes the optical flow frames of the video as input. The spatiotemporal saliency maps, which are obtained by a geodesic distance-based video segmentation method [51], are used as the input of the spatiotemporal saliency stream. This can capture the spatiotemporal object foreground information in the video and suppress the background information.

Further, in order to capture the temporal long-term dependency relationships between consecutive video frames, we propose a novel spatiotemporal saliency-based multi-stream ResNets with attention-aware LSTM (STS-ALSTM for short) for action recognition. The proposed STS-ALSTM model can capture the abstract appearance features, motion features and spatiotemporal saliency features through the STS multi-stream model. Then these three different extracted CNN features are fed into three corresponding stacked attention-aware LSTMs which can capture long-term dependency relationships in the temporal dimension. Lastly, an averaging fusion is adopted for the outputs of three different LSTM streams. Note that the preliminary results of this research have been reported in a conference paper in [58]. Compared with the paper [58] which proposed the STS multi-stream model, we further propose the STS-ALSTM model which extends our original STS multi-stream model by connecting an attention-aware LSTM. We also perform more extensive experiments to compare our proposed STS multi-stream model and STS-ALSTM model with related CNN-based or LSTM-based methods for action recognition.

In summary, the contributions of this paper include: (i) We propose a novel spatiotemporal saliency-based multi-stream ResNet (STS) for human action recognition, which consists of a spatial stream, a temporal stream and a spatiotemporal stream. The spatial stream is responsible for capturing appearance information from raw RGB video frames, the temporal stream is responsible for capturing motion information from optical flow frames, and the spatiotemporal saliency stream is responsible for capturing the spatiotemporal foreground information of video objects from spatiotemporal saliency maps. (ii) The novel spatiotemporal saliency stream can reduce the background interference in videos and provide the spatiotemporal object foreground information for human action recognition. (iii) Based on STS, we propose a novel spatiotemporal saliency-based multi-stream ResNets with attention-aware LSTM (STS-ALSTM) for action recognition, use multi-stream ResNets as a deep CNN feature extractor and then input the extracted features into three individual attention-aware LSTMs, which can capture the temporal long-term dependency relationships between consecutive video frames. (iv) An averaging fusion is adopted for the outputs of the three LSTM streams.

The rest of this paper is organized as follows. Section 2 presents related work. The proposed methods are presented in detail in Sect. 3. Section 4 shows the results of conducted extensive experiments. Section 5 provides the conclusions of the paper.

2 Related works

2.1 Two-stream network-based methods and 3D CNN-based methods

Recently, two-stream-based 2D convolutional neural networks have been widely applied for human action recognition. Simonyan et al. [41] first proposed a two-stream CNN architecture, in which spatial and temporal neural networks were developed to capture spatial and temporal information of videos separately, and the output of these two networks was combined by late fusion. Wang et al. [50] proposed the temporal segment network (TSN) with four types of input modalities, which was based on the idea of long-range temporal video structure modeling. Feichtenhofer et al. [10] proposed spatiotemporal residual networks (ST-ResNet) to add residual connections between different layers and learned spatiotemporal features by connecting the appearance channel and motion channel. Wang et al. [54] developed a spatiotemporal pyramid network to fuse the spatial and temporal features. A spatiotemporal compact bilinear operator was adopted to enable unified modeling of various fusion strategies. Jing et al. [55] combined multiple streams with dynamic images, optical flow frames and raw frames as input to improve the performance of action recognition. Liu et al. [25] proposed a multi-stream neural network by using RGB frames, dense optical flow frames and gradient maps as the input, where different streams were responsible for capturing various appearance and motion feature information.

Different from two-stream networks, 3D convolution can process multiple consecutive images at the same time, and 3D convolution neural networks have the ability to extract temporal information between video frames [13, 17, 32, 33, 36]. Ji et al. [16] firstly developed a 3D CNN model that provided multiple channels from adjacent input frames and performed 3D convolution for each channel. Du et al. [47] proposed convolutional 3D (C3D) which used multi-frames as an input of the network. Diba et al. [8] developed temporal 3D CNN (T3D) by deploying a 3D temporal transition layer (TTL) instead of a transition layer in DenseNet [15]. Qiu et al. [37] developed a residual learning model by using different convolution filters and proposed the Pseudo-3D Residual Net (P3D ResNet). Yang et al. [56] developed an asymmetric 3D convolutional deep model and proposed a multi-source-enhanced input method to decrease the computational cost. Li et al. [22] developed a spatiotemporal deformable 3D convolutions by using an attention mechanism and exploited temporal and spatial dependencies. Liu et al [24] developed a HDS-SP descriptor for skeleton-based human action by using a better viewpoint. 3D CNN-based networks need training much more parameters and cost expensive computation compared with 2D CNN-based networks [47].

2.2 RNN-based methods and others

Since videos consist of a series of consecutive video frames, RNN-based methods have been widely used for action recognition. Baccouche et al. [2] proposed to use 3D-CNN to extract abstract features from raw video frames input, and then, a LSTM was adopted for classification. Ng et al. [57] proposed to use CNNs to extract features from raw video frames or optical flow frames and then fed the outputs of a CNN into a LSTM for classification. Donahue et al. [9] proposed long-term recurrent convolutional networks (LRCNs) for activity recognition, which used CNNs to extract features and use LSTM networks for recognition. Cheng et al. [6] proposed a two-stream attention-based LSTM architecture, which utilized a visual attention mechanism for human action recognition. Mahshid et al. [27] proposed extended LSTM units to perceive the motion data and extract motion features through a spatiotemporal component for action recognition. Amin et al. [48] proposed to use convolutional neural networks to extract deep features and to use deep bidirectional LSTM networks to learn the sequential features.

In addition to the development of the above three kinds of methods for human action recognition, some research contributes to the related fields (such as data input, model architecture and fusion) to address the challenges in human action recognition. Kar et al. [18] developed AdaScan to dynamically pool the key informative frames and proposed a pooled feature vector for human action recognition. Sun et al. [46] proposed a compact motion representation which can be embedded in any existing CNN-based video action recognition framework with a slight additional cost. Xie et al. [55] combined top-heavy model design, temporally separable convolution and spatiotemporal feature gating together to improve the performance of action recognition. Shamsolmoali et al. [38] developed two residual multiple instance learning (MIL) models for the human pose estimation task by using generation and discriminator of the identical architecture. Sun et al. [43] developed a network that maintained high-resolution representations through the whole pose estimate process. Si et al. [40] developed an attention-enhanced graph convolutional LSTM network (AGC-LSTM) and used the attention mechanism to get information of key joints. To reduce noises in human motion vectors and capture fine motion details, Shou et al. [39] developed a lightweight generator network and achieved a more discriminative motion cue (DMC) representation. The usage of two-stream-based methods in real-world applications requires low latency; Crasto et al. [5] developed a linear combination of the feature-based loss and the standard cross-entropy loss training method for action recognition.

3 The proposed methods: STS and STS-ALSTM

In this section, we first introduce the spatiotemporal saliency map generated by [51] in Sect. 3.1. Then we propose a novel spatiotemporal saliency-based multi-stream ResNet (STS) for human action recognition in Sect. 3.2. After that, we propose a novel spatiotemporal saliency-based multi-stream ResNets with attention-aware LSTM (STS-ALSTM) for action recognition in Sect. 3.3. Finally, Sect. 3.4 describes the training process and the fusion strategy of the proposed methods.

3.1 Spatiotemporal saliency map

Video object segmentation method is a process of extracting objects from videos, which is widely utilized in many visual-related tasks and applications [31, 35, 50]. For human action recognition, video object segmentation has the ability to segment foreground human objects from complex background in all video frames. Inspirited by a geodesic distance-based video segmentation method [51], which distinguish the foreground objects from surrounded background areas according to the corresponding spatiotemporal edge values, this paper generates spatiotemporal saliency maps from videos using this technique.

The procedure of the proposed method can be summarized as the following steps: (i) obtaining a superpixel set for the input video frames by using a k-means clustering method [1]; (ii) obtaining a spatial edge probability map by using an edge detection method [21]; (iii) obtaining the temporal gradient magnitude of optical flow frames [4]; (iv) computing the spatial edge probability of each superpixel to obtain spatial superpixel edge maps; (v) computing the temporal gradient magnitude of each superpixel to obtain temporal superpixel optical flow magnitude maps; (vi) obtaining spatiotemporal edge probability maps by combing the spatial superpixel edge maps and the temporal superpixel optical flow magnitude maps; and (vii) obtaining spatiotemporal saliency maps from the spatiotemporal edge probability maps by calculating the probability of foreground objects based on their geodesic distance.

The spatiotemporal saliency maps generated by the geodesic distance-based video segmentation method [51] is shown in Fig. 1. The spatiotemporal saliency maps contain both foreground information and edge information of human objects, which provides rich prior spatiotemporal knowledge for human action recognition.

3.2 STS model

The framework of our proposed spatiotemporal saliency-based multi-stream ResNet (STS) model is illustrated in Fig. 2. The STS model consists of three streams: a spatial stream, a temporal stream and a spatiotemporal saliency stream. The spatial stream is responsible for capturing appearance information from raw RGB video frames, the temporal stream is responsible for capturing motion information from optical flow frames, and the spatiotemporal saliency stream is responsible for capturing the spatiotemporal foreground information of video objects from spatiotemporal saliency maps. The neural networks for the spatial stream, the temporal stream and the spatiotemporal saliency stream are trained individually. Then the outputs of the softmax layers of the three streams are averaged for fusion to form a final softmax score for human action recognition.

3.3 STS-ALSTM model

The framework of our proposed spatiotemporal saliency-based multi-stream ResNets with attention-aware LSTM (STS-ALSTM) model is illustrated in Fig. 3. Similar to the proposed STS model, the STS-ALSTM model consists of three CNN streams and three LSTM streams. Similar to the STS model, the CNN part contains: a spatial stream with RGB video frames as input, a temporal stream with optical flow frames as input and a spatiotemporal saliency stream with spatiotemporal saliency maps as input. We use them as a deep CNN feature extractor and then input the extracted three different CNN features into three individual attention-aware LSTMs to capture the temporal long-term dependency relationships between consecutive video frames, optical flow frames or spatiotemporal saliency frames. Specifically, we split the input frames of each individual stream into multiple sequences and then put every short sequence of frames into the proposed CNN encoder to generate each sequence with 1D vectors. Finally, the attention-aware LSTM model can take the 1D vectors to synthesize temporal information for action recognition.

3.3.1 Deep CNN feature extractor

Recently, various CNNs have proved their abilities in extracting both spatial and temporal features from videos, especially residual networks [11, 14, 20, 30]. Different from other deep neural networks using multiple stacked layers $F(x)$ to approximate the desired underlying mapping $H(x)$, residual networks consider using multiple stacked

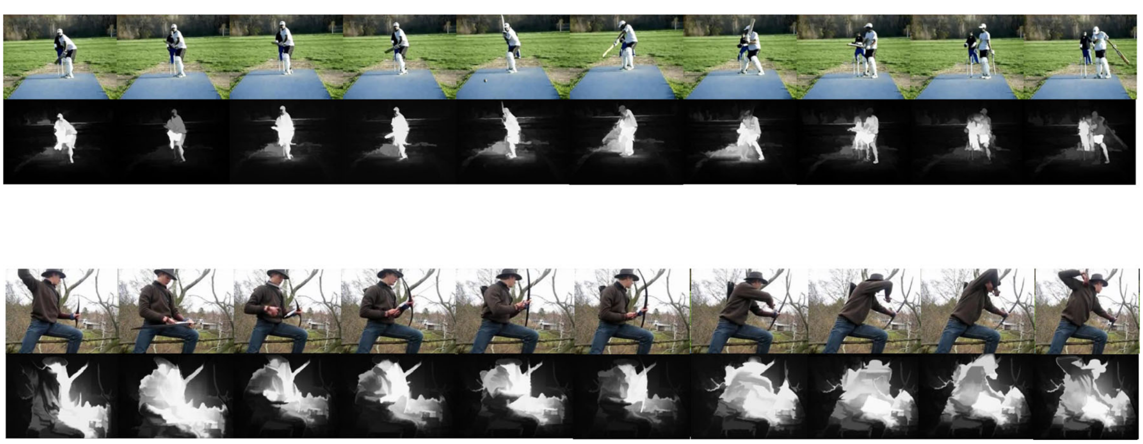

Fig. 1 Spatiotemporal saliency maps generated by the geodesic distance-based video object segmentation method [51]. The top row shows 10 consecutive RGB frames sampled with a fixed time interval in the "Cricket Shot" and "Archery" videos from UCF-101 dataset [42], and the second row illustrates their corresponding spatiotemporal saliency maps

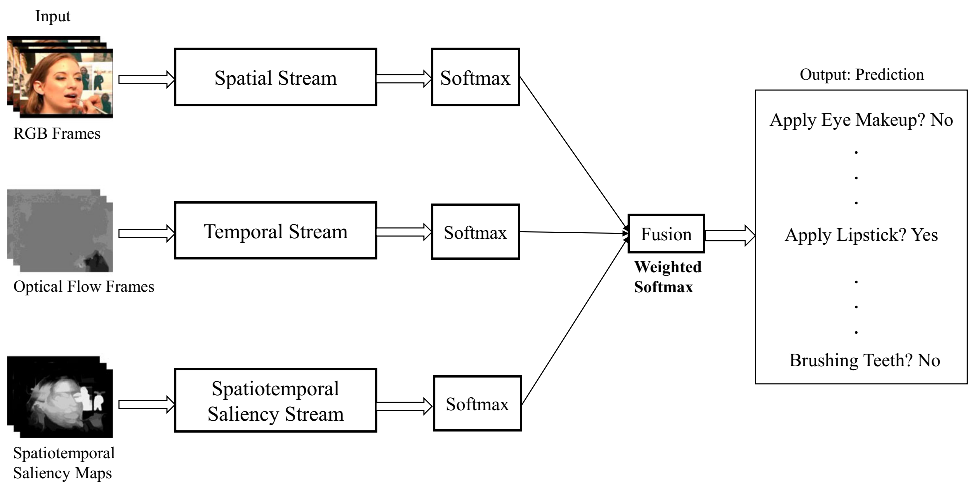

Fig. 2 Framework of our proposed STS human action recognition model. It consists of a spatial stream with RGB video frames as input, a temporal stream with optical flow frames as input and a spatiotemporal saliency stream with spatiotemporal saliency maps as input

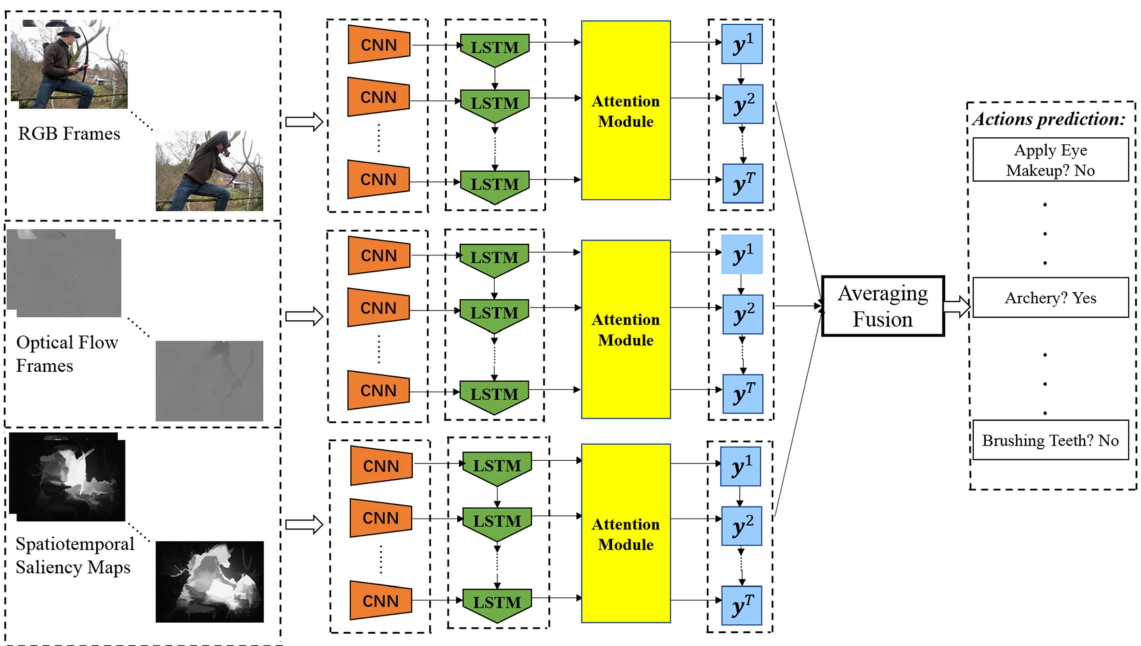

Fig. 3 Framework of the proposed STS-ALSTM model

layers $F(x)$ to approximate a residual mapping $H(x) - x$. In this paper, we use ResNet to encode each 2D frame $x^{(t)}$ into a 1D vector $r^{(t)}$ by using $f_{cnn}(x^{(t)}) = r^{(t)}$, which presents the output of the last fully connected layer.

3.3.2 Attention-aware LSTM

Recurrent neural networks, especially LSTM networks, are widely applied to many tasks such as text generation, machine translation and speech recognition [3, 26, 34, 44, 53]. In this paper, we use a LSTM network as a decoder to process the obtained vector $r^{(t)}$. The LSTM network utilizes the 1D vector $r^{(t)}$ as the input and outputs another 1D sequence $h^{(t)}$. Since a typical LSTM unit mainly includes an input activation function, a single memory cell and three gates (i.e., an input gate i_t, a forget gate f_t and an output gate o_t), we set $\sigma(x) = (1 + e^{-x})^{-1}$ as the sigmoidal nonlinearity that can map the input data into the interval [0,1], and set $\varphi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\sigma(2x) - 1$ as the hyperbolic tangent nonlinearity that can map the input data into the interval $[-1, 1]$. The related formulas of the LSTM unit are shown as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{3}$$

$$c_t = f_t \odot c_t + i_t \odot g_t \tag{4}$$

$$h_t = o_t \odot \varphi(c_t) \tag{5}$$

where W_{xi} is the relevant weight matrixes between layers; b_i is the bias; c_t is the memory cell unit that is a summation of the previous memory cell unit $c_{(t-1)}$ modulated by the forget gate f_t and the input modulation gate g_t modulated by the input gate i_t; h_t is the hidden unit; and $\odot$ is the element-wise product with the gate value.

The attention module utilized in this paper is illustrated in Fig. 4. We combine both the current target hidden state h_t and the context vector c_t to produce an attention-aware hidden state $h_t^{'}$. Thus, the attention-aware hidden state can be defined as follows:

$$h_t^{'} = tanh(W_c[c_t; h_t]) \tag{6}$$

The attention vector c_t which depends on previous hidden states $\{h_1, \ldots, h_{(T_x)}\}$ can be defined as follows:

$$c_t = \sum_{j=1}^{T_x} a_{ij} h_j \tag{7}$$

The weight a_{ij} of each target hidden state h_j is formulated as follows:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{8}$$

$$\beta = \tanh(w_{xt}x_t + w_{x(t-1)}x_{t-1} + b_t) \tag{9}$$

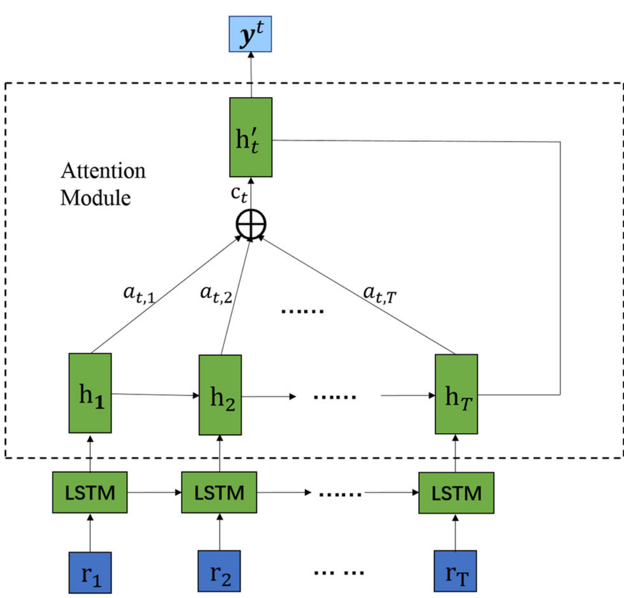

Fig. 4 Attention module at each time step t. The model infers a variable length alignment weight vector $a_{(i,j)}$ based on the attention vector c_t and the current target hidden state h_t, and y^t is the attention-aware output at the final attention-aware target hidden state $h_t^{'}$

where $e_{ij} = \beta(h_{(i-1)}^{'}, h_j)$ is a scoring model for determining the matching degree between the input at position j and the output at position i, β is the relevance coefficient and b_t is the bias parameter.

The attention aware of a hidden state vector $h_t^{'}$ is then fed through the softmax classifier for categorical predictions as follows:

$$P(y_i|y_1, \ldots, y_{i-1}, r) = softmax(W_s, h_t^{'}) \tag{10}$$

3.4 The training process and the fusion strategy

The proposed two models all consist of three different streams: a spatial stream, a temporal stream and a spatiotemporal saliency network. We train the three streams separately to extract the appearance information, motion information and spatiotemporal saliency information from videos. All training details are summarized as follows.

The spatial stream with RGB frames as input provides the basic appearance characteristics of the video, which is the most important stream in the action recognition process [50]. The input of the spatial stream consists of multiple RGB frames obtained in a random sampling interval from the extracted video frames. Similar to the temporal segment network [50] training strategy, we randomly select ten video frames from a video for representing the video. Then a consensus among the selected frames is derived as the video-level prediction. We input the ten video frames separately into each CNN and calculate the losses; then, these losses will be added as the final loss for backpropagation. The output is fed into a stack of attention-aware LSTMs; finally, the attention-aware hidden state vector is then fed through the softmax classifier for categorical predictions.

The temporal stream with optical flow frames as input provides the motion information of the action, which has been crucial for action recognition. We use the optical flow estimation [4] method to obtain optical flow frames from the raw RGB frames of videos. Different from the input of the spatial stream, we randomly select a series of stacked optical flow frames from the optical flow frames as the input of the temporal stream and use each CNN to extract features and each attention-aware LSTM to predict categories.

The spatiotemporal saliency stream with spatiotemporal saliency maps as input provides the spatiotemporal object foreground information and reduces the background interference. We utilize a geodesic distance-based video segmentation method [51] to obtain the spatiotemporal saliency maps from the RGB frames and optical flow frames. Similar to the input of the spatial stream, we randomly select ten frames from the spatiotemporal saliency maps, and we input these gray images separately into every CNN to extract spatiotemporal saliency foreground objects features, and the output is fed into a stack of attention-aware LSTMs to finish decoding. The attention-aware outputs for categorical predictions represent the output of the spatiotemporal saliency stream.

Since the proposed two models all consists of three streams (i.e., a spatial stream, a temporal stream and a spatiotemporal saliency network) as inputs, the outputs of these three streams need to be fused together to integrate the spatial information and temporal information of videos. In this paper, we utilize an averaging fusion [6, 27] as the fusion strategy, and then, the fusion results can be used for recognition.

4 Experiments

4.1 Datasets

We evaluate the performance of our proposed model on UCF-101 [42] and HMDB-51 [19] datasets. The UCF-101 dataset consists of 101 action categories with 13320 video clips. The HMDB-51 dataset includes 6849 video clips divided into 51 action categories, and each category contains a minimum of 101 video clips. We use the pre-provided training/test split of the UCF-101, which divides the UCF-101 dataset into 9537 training videos and 3783 testing videos. Similarly, we use the pre-provided training/test split of the HMDB-51, which contains about 3750 training videos and 3099 test videos.

4.2 Implementation details

We use Pytorch to implement our proposed model and train the model on 4 Nvidia GTX 2080Ti GPUs. We set the learning rate to 0.001 and use a mini-batch size of 32. We adopt 101-layer ResNet (ResNet-101 for short) for feature extraction of the spatial stream, the temporal stream and the spatiotemporal saliency stream. We first use the pre-trained ResNet-101 on the ImageNet dataset, which is a large-scale hierarchical image database containing more than 1 million images [7], as the spatial stream model parameter initialization. Then we fine-tune the pre-trained ResNet-101 on the UCF-101 and HMDB-51 datasets. For the temporal stream, by averaging the weight value across RGB channels and replicating this value by the channel number of motion stream input, we use ImageNet pre-trained weights and modify the weights of the first convolution layer pre-trained on ImageNet from (64, 3, 7, 7) to (64, 20, 7, 7), which contains 10 x-channel and 10 y-channel optical flow frames. Similar to the spatial stream, we use the pre-trained ResNet-101 on ImageNet and fine-tune the spatiotemporal saliency stream. And then, the output of the last fully connected layer is fed into a stack of attention-aware LSTMs to finish categorical prediction.

4.3 Comparison with different input

The loss scores and classification accuracies of two-stream method and our STS multi-stream method for human action recognition on the UCF-101 dataset are illustrated in Fig. 5. As shown in Fig. 5, we found that the loss value of the two-stream convolutional neural network is smaller than that of the STS multi-stream residual neural network, but as the number of iterations increases, the loss of the multi-stream residual neural network does not change significantly and gradually stabilizes at a certain level. The numerical value indicates that the parameter optimization of the network model has reached a better state. At the same time, it can be found that when the recognition accuracy of the two methods reaches 80%, the proposed

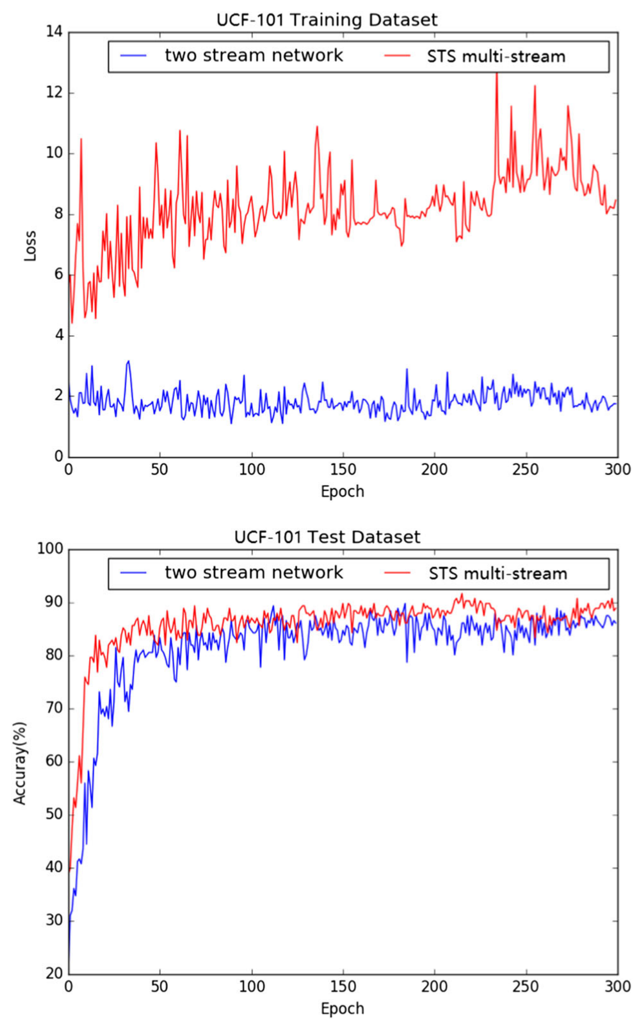

Fig. 5 Loss scores and accuracies of two comparison methods on the UCF-101 dataset

Table 1 Accuracy of different modalities on the UCF-101 dataset

Input	STS (%)	STS-ALSTM (%)
RGB	81.3	82.3
Optical flow	79.7	83.5
RGB + optical flow	87.2	89.3
RGB + spatiotemporal saliency	82.5	84.5
Optical flow + spatiotemporal saliency	80.4	86.1
RGB + optical flow + spatiotemporal saliency	90.1	92.7

Table 2 Accuracy of different modalities on the HMDB-51 dataset

Input	STS (%)	STS-ALSTM (%)
RGB	50.1	52.6
Optical flow	55.6	57.1
RGB + optical flow	60.5	63.8
RGB + spatiotemporal saliency	53.7	56.7
Optical flow + spatiotemporal saliency	57.8	57.9
RGB + optical flow + spatiotemporal saliency	62.4	64.4

method reduces the number of iterations by more than 10 times compared with the two-stream method. STS multi-stream residual neural networks we proposed can achieve recognition accuracy of more than 90%, and the network model always remains stable. From the above analysis, the STS multi-stream network based on spatiotemporal saliency is superior to the two-stream convolutional neural network in model training and can maintain a stable recognition rate.

The experimental results are reported in Tables 1 and 2. It is obvious that the accuracy of the input with two modalities (such as RGB frames + optical flow frames) is higher than the input with a single modality (such as RGB frames) on both UCF-101 dataset and HMDB-51 datasets. Further, by using our STS-ALSTM multi-stream model, we can find that the input with optical flow frames and spatiotemporal saliency improves 2.6% and 0.8% than the input with only optical flow frames on UCF-101 and HMDB-51 datasets, respectively. The addition of spatiotemporal saliency stream can provide the spatiotemporal object foreground information and reduce the background interference, which is beneficial for action recognition. A similar phenomenon can be verified when we use RGB frames and spatiotemporal saliency maps as the input of STS-ALSTM multi-stream model, the input with RGB frames and spatiotemporal saliency improves 2.2% and 4.1% than the input with only RGB frames on UCF-101 and HMDB-51 datasets, respectively. When we fuse all these three streams of our STS-ALSTM multi-stream model, we can obtain the best accuracy of 92.7% and 64.4% on UCF-101 and HMDB-51 datasets, respectively. By using STS-ALSTM multi-stream model, the input with all three modalities improves 3.4% and 0.6% than the input with RGB frames and optical flow frames on UCF-101 and HMDB-51 datasets, respectively, which demonstrates that the spatiotemporal saliency stream can further provide effective supplementary information for improving the performance of action recognition.

4.4 Comparison with state of the art

Table 3 compares the experimental results of the proposed STS multi-stream method and other state-of-the-art methods for human action recognition. The proposed multi-stream model is superior to iDT+HSV [28], C3D [47], deeper temporal net [12], two-stream [41], FstCN [45], TDD+FV [49], scLSTM [52], VideoLSTM [23], L2STM [44] and STS model [58]. Especially compared with other two-stream-based models such as Two-stream [41] and

Table 3 Comparison of our method based on multi-stream with the state-of-the-art methods on the UCF-101 and HMDB-51 datasets

Methods	UCF-101 (%)	HMDB-51 (%)
iDT+HSV [28]	87.9	61.1
C3D [47]	85.2%	–
Deeper temporal net [12]	84.9	–
Two-stream [41]	88.0	54.9
FstCN [45]	88.1	59.1
TDD+FV [49]	90.3	63.2
scLSTM [52]	84.0	55.1
VideoLSTM [23]	89.2	56.4
L2STM [44]	93.6	66.2
STS	90.1	62.4
STS-ALSTM	92.7	64.4

two-stream + LSTM [57], our proposed multi-stream STS-ALSTM model obtains better performance since the spatiotemporal saliency stream can provide the spatiotemporal object foreground information and capture long-term dependency relationships in the temporal dimension.

5 Conclusion

In this paper, we propose a novel spatiotemporal saliency-based multi-stream ResNet and a novel spatiotemporal saliency-based multi-stream ResNet with attention-aware LSTM for action recognition; these two models consist of three complementary streams: a spatial stream with RGB frames as input, a temporal stream with optical flow frames as input and a spatiotemporal saliency stream with spatiotemporal saliency maps as input. Compared with conventional two-stream-based models and LSTM-based models, the proposed methods STS can provide the spatiotemporal object foreground information and reduce the background interference, which has been verified effective for human action recognition, and the STS-ALSTM multi-stream model can further capture long-term dependency relationships between consecutive video frames. Experimental results demonstrate that our proposed STS-ALSTM multi-stream model achieves the best accuracy compared with the input with single modality or two modalities. In the future, we will further explore sharing information between different streams to improve the performance of human action recognition.

Acknowledgements This study is supported by the National Natural Science Foundation of China (Grant Nos. 61562013, 61906050), the Natural Science Foundation of Guangxi Province (CN) (2017GXNSFDA198025), the Study Abroad Program for Graduate Student of Guilin University of Electronic Technology (GDYX2018006), the National Natural Science Foundation of China (Grant 61602407), the Natural Science Foundation of Zhejiang Province (Grant LY18F020008), the China Scholarship Council (CSC) and the New Zealand China Doctoral Research Scholarships Program.

References

1. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34(11):2274–2282
2. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. In: International workshop on human behavior understanding. Springer, pp 29–39
3. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv:1409.0473
4. Brox T, Bruhn A, Papenberg N, Weickert J (2004) High accuracy optical flow estimation based on a theory for warping. In: European conference on computer vision. Springer, pp 25–36
5. Crasto N, Weinzaepfel P, Alahari K, Schmid C (2019) Mars: motion-augmented rgb stream for action recognition. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
6. Dai C, Liu X, Lai J (2020) Human action recognition using two-stream attention based LSTM networks. Appl Soft Comput 86:105820
7. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255
8. Diba A, Fayyaz M, Sharma V, Karami AH, Arzani MM, Yousefzadeh R, Van GL (2017) Temporal 3d convnets: new architecture and transfer learning for video classification. arXiv:1711.08200
9. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
10. Feichtenhofer C, Pinz A, Wildes R (2016) Spatiotemporal residual networks for video action recognition. In: Conference on neural information processing systems, pp 3468–3476
11. Gong W, Qi L, Xu Y (2018) Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment. Wirel Commun Mobile Comput. https://doi.org/10.1155/2018/3075849
12. Han Y, Zhang P, Zhuo T, Huang W, Zhang Y (2018) Going deeper with two-stream convnets for action recognition in video surveillance. Pattern Recognit Lett 107(2018):83–90
13. Hara K, Kataoka H, Satoh Y (2017) Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 3154–3160
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
15. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
16. Ji S, Wei X, Yang M, Kai Y (2012) 3d Convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
17. Jing L, Ye Y, Yang X, Tian Y (2017) 3d Convolutional neural network with multi-model framework for action recognition. In: 2017 IEEE international conference on image processing (ICIP), pp 1837–1841
18. Kar A, Rai N, Sikka K, Sharma G (2017) Adascan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3376–3385
19. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision, pp 2556–2563
20. Lei P, Todorovic S (2018) Temporal deformable residual networks for action segmentation in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6742–6751
21. Leordeanu M, Sukthankar R, Sminchisescu C (2012) Efficient closed-form solution to generalized boundary detection. In: European conference on computer vision. Springer, pp 516–529
22. Li J, Liu X, Zhang M, Wang D (2020) Spatio-temporal deformable 3d convnets with attention for action recognition. Pattern Recognit 98(2020):107037

23. Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CGM (2018) Videolstm convolves, attends and flows for action recognition. Comput Vis Image Underst 166:41–50
24. Liu J, Wang Z, Liu H (2020) Hds-sp: a novel descriptor for skeleton-based human action recognition. Neurocomputing 385:22–32
25. Liu X, Yang X (2018) Multi-stream with deep convolutional neural networks for human action recognition in videos. In: International conference on neural information processing. Springer, pp 251–262
26. Luong M-T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv:1508.04025
27. Majd M, Safabakhsh R (2020) Correlational convolutional LSTM for human action recognition. Neurocomputing 396:224–229
28. Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput Vis Image Underst 150:109–125
29. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
30. Qi L, Dai P, Jiguo Y, Zhou Z, Yanwei X (2017) "Time-location-frequency"—aware internet of things service selection based on historical records. Int J Distrib Sens Netw 13(1):1550147716688696
31. Qi L, Dou W, Chen J (2016) Weighted principal component analysis-based service selection method for multimedia services in cloud. Computing 98(1–2):195–214
32. Qi L, Wang R, Chunhua H, Li S, He Q, Xiaolong X (2019) Time-aware distributed service recommendation with privacy-preservation. Inf Sci 480:354–364
33. Qi L, Xu X, Dou W, Yu J, Zhou Z, Zhang X (2016) Time-aware IoE service recommendation on sparse data. Mobile Inf Syst 2016:12. https://doi.org/10.1155/2016/4397061
34. Qi L, Yu J, Zhou Z (2017) An invocation cost optimization method for web services in cloud environment. Sci Program 2017:9. https://doi.org/10.1155/2017/4358536
35. Qi L, Zhang X, Dou W, Chunhua H, Yang C, Chen J (2018) A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment. Future Gener Comput Syst 88(2018):636–643
36. Qi L, Zhang X, Dou W, Ni Q (2017) A distributed locality-sensitive hashing-based approach for cloud service recommendation from multi-source data. IEEE J Sel Areas Commun 35(11):2616–2624
37. Qiu Z, Yao T, Mei T (2017) Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE international conference on computer vision, pp 5533–5541
38. Shamsolmoali P, Zareapoor M, Zhou H, Yang J (2020) Amil: adversarial multi instance learning for human pose estimation. ACM Trans Multimedia Comput Commun Appl (TOMM) 16(1s):1–23
39. Shou Z, Lin X, Kalantidis Y, Sevilla-Lara L, Rohrbach M, Chang S-F, Yan Z (2019) Dmc-net: generating discriminative motion cues for fast compressed video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1268–1277
40. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1227–1236
41. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp 568–576
42. Soomro K, Zamir AR, Shah M (2012) Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
43. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5693–5703
44. Sun L, Jia K, Chen K, Yeung D-Y, Shi BE, Savarese S (2017) Lattice long short-term memory for human action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2147–2156
45. Sun L, Jia K, Yeung D-Y, Shi BE (2015) Human action recognition using factorized spatio-temporal convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4597–4605
46. Sun S, Kuang Z, Sheng L, Ouyang W, Zhang W (2018) Optical flow guided feature: a fast and robust motion representation for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1390–1399
47. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
48. Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW (2017) Action recognition in video sequences using deep bi-directional lstm with cnn features. IEEE Access 6:1155–1166
49. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4305–4314
50. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van GL (2018) Temporal segment networks for action recognition in videos. IEEE Trans Pattern Anal Mach Intell 41(11):2740–2755
51. Wang W, Shen J, Porikli F (2015) Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3395–3402
52. Wang X, Gao L, Song J, Shen H (2016) Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition. IEEE Signal Process Lett 24(4):510–514
53. Wang Y, Huang M, Zhao L et al (2016) Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615
54. Wang Y, Long M, Wang J, Yu PS (2017) Spatiotemporal pyramid network for video action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 1529–1538
55. Xie S, Sun C, Huang J, Tu Z, urphy K (2018) Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV), pp 305–321
56. Yang H, Yuan C, Li B, Yang D, Xing J, Weiming H, Maybank SJ (2019) Asymmetric 3d convolutional neural networks for action recognition. Pattern Recognit 85(2019):1–12
57. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
58. Zhenbing L, Zeya L, Ming Z, Wanting J, Ruili W, Yan T (2019) Spatiotemporal saliency based multi-stream networks for action recognition. In: Asian conference on pattern recognition, Springer, Singapore, pp 74–84

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