Mutually Reinforced Spatio-Temporal Convolutional Tube for Human Action Recognition

Haoze Wu[1*], **Jiawei Liu**[1*], **Zheng-Jun Zha**[1†], **Zhenzhong Chen**[2] and **Xiaoyan Sun**[3]
[1]National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, University of Science and Technology of China
[2]School of Remote Sensing and Information Engineering, Wuhan University
[3]Intelligent Multimedia Group, Microsoft Research Asia
{wuhaoze, ljw368}@mail.ustc.edu.cn, zhazj@ustc.edu.cn, zzchen@whu.edu.cn, xysun@microsoft.com

Abstract

Recent works use 3D convolutional neural networks to explore spatio-temporal information for human action recognition. However, they either ignore the correlation between spatial and temporal features or suffer from high computational cost by spatio-temporal features extraction. In this work, we propose a novel and efficient Mutually Reinforced Spatio-Temporal Convolutional Tube (MRST) for human action recognition. It decomposes 3D inputs into spatial and temporal representations, mutually enhances both of them by exploiting the interaction of spatial and temporal information and selectively emphasizes informative spatial appearance and temporal motion, meanwhile reducing the complexity of structure. Moreover, we design three types of MRSTs according to the different order of spatial and temporal information enhancement, each of which contains a spatio-temporal decomposition unit, a mutually reinforced unit and a spatio-temporal fusion unit. An end-to-end deep network, MRST-Net, is also proposed based on the MRSTs to better explore spatio-temporal information in human actions. Extensive experiments show MRST-Net yields the best performance, compared to state-of-the-art approaches.

1 Introduction

Human action recognition aims to recognize human actions by the visual appearance and motion dynamics of the involved humans and objects in video sequences. It is a fundamental yet challenging task due to the temporal dynamic of video content, low resolution and background interference, *etc.* Considerable efforts have been investigated for decades, and existing approaches could be roughly divided into two categories: 2D CNN based methods and 3D CNN based methods.

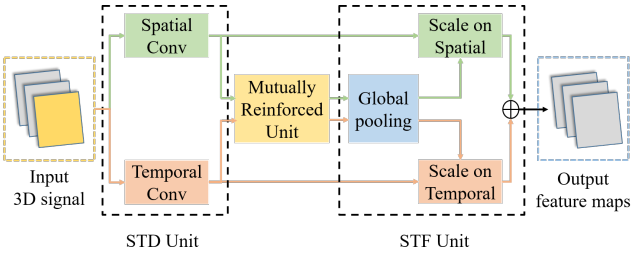


Figure 1: The illustration of our proposed MRST. The MRST mainly consists of three units: the STD unit, the mutually reinforced unit and the STF unit.

The 2D CNN based methods use either Two-stream CNNs [Simonyan and Zisserman, 2014] or CNN-Long Short-Term Memory (CNN-LSTM) [Donahue *et al.*, 2015]. The Two-stream CNNs use two CNN branches to extract spatial features from RGB frames and temporal features from stacked optical flow, respectively. The Two-stream CNNs capture temporal dependencies with hand-crafted optical flow information, which suffers from expensive computational cost and cannot model the correlation between spatial and temporal information. The CNN-LSTM networks connect a LSTM layer to the bottom of the CNNs. They firstly extract appearance features with CNNs and then model temporal information from the video frames by a LSTM layer in tandem, which loses plenty of useful information.

Jointly modeling spatio-temporal information via 3D CNNs is a natural and efficient approach for human action recognition. However, the basic 3D CNN (C3D) [Tran *et al.*, 2015] suffers from the high computational complexity of 3D convolution operation. Some works [Tran *et al.*, 2018; Xie *et al.*, 2018; Zhou *et al.*, 2018] attempt to design improved architectures with the basic 3D convolution. Nevertheless, these methods still suffer from high computational cost using 3D convolutional operation. Other works [Sun *et al.*, 2015; Qiu *et al.*, 2017] decompose the 3D convolution into two separate convolutions, i.e., a 2D spatial convolution plus a 1D temporal convolution and thus significantly reduce the model size. However, these methods ignore the correlation between spatial and temporal information. Actually, the joint explo-

[*]Equal contribution
[†]Corresponding author

ration of them could offer a comprehensive representation of videos and thus enhance the accuracy of action recognition. Human actions in videos contain both appearance information and motion information. For example, in a human playing basketball video, exploring the spatial and temporal information jointly can pay more attention to the human and ball's appearance information rather than the surroundings' since the human and ball are in continuous movement.

In this paper, we propose a Mutually Reinforced Spatio-Temporal Convolutional Tube (MRST) towards robust and accurate human action recognition. It factorizes the 3D inputs into spatial and temporal representations, mutually enhances both of them by exploiting the interaction of spatial and temporal information, and fuses the enhanced representations to obtain effective spatio-temporal features, while reducing the complexity. Specifically, the MRST consists of a spatio-temporal decomposition unit, a mutually reinforced unit and a spatio-temporal fusion unit. An illustration of our MRST is shown in Fig.1. The spatio-temporal decomposition unit extracts spatial and temporal features with a 2D convolution and a 1D convolution, respectively. The mutually reinforced unit learns the correlation between spatial and temporal information by four fully connected layers, and utilizes the correlation to reinforce the discriminative capability of the spatial features and temporal features. The spatio-temporal fusion unit selectively emphasizes informative spatial and temporal features by a global pooling layer and a sigmoid layer and fuses them to get the effective spatio-temporal feature maps. In addition, we design three types of MRSTs according to the different order of spatial and temporal information enhancement, $i.e.$, MRST-P, MRST-S and MRST-T. A novel deep network, MRST-Net, is also proposed based on the MRST to better explore spatio-temporal information in human actions. Experiment results show our proposed deep MRST-Net achieves state-of-the-art performance on three challenging action recognition datasets, Kinetics-400, UCF-101 and HMDB51 with only RGB inputs.

2 Related Work

Human action recognition is one of the core computer vision tasks and has been studied for decades. Here we outline work involving deep features and classify them into two categories: 2D CNN and 3D CNN based approaches, according to the convolutions used in feature learning.

2D CNN based. To explore the spatio-temporal information in human actions, the two-stream architecture is first proposed in [Simonyan and Zisserman, 2014] where two 2D CNNs are applied to the appearance (RGB frames) and motion (stacked optical flow) domains, respectively. Features from the two modalities are fused at the final stage and the two-stream architecture achieves high video recognition accuracy. Based on this architecture, several mechanisms are proposed for a better fusion of the two branch networks over the appearance and motion [Karpathy *et al.*, 2014; Feichtenhofer *et al.*, 2017; Tran and Cheong, 2017; Wang *et al.*, 2018a]. On the other hand, early attempts which incorporate LSTM with traditional features have shown the potential of the CNN-LSTM architecture for modeling spatio-temporal information in action recognition [Donahue *et al.*, 2015]. LSTM networks are employed to utilize the frame-level features of 2D CNNs to explicitly model spatio-temporal relationships. The work[Hu *et al.*, 2018] extracts different channels' attention to further improve the quality of representations produced by a network.

3D CNN based. The 3D CNN for action recognition is first presented in [Ji *et al.*, 2013] to learn discriminative features along both spatial and temporal dimensions. Later, the C3D feature along with the corresponding 3D CNN architectures are presented in [Tran *et al.*, 2015]. Since they use 3D convolution kernels to model both spatial and temporal information rather than 2D kernels which just model spatial information, more complex relations between appearance and motion can be learned and captured. The Res3D[Tran *et al.*, 2017] makes one step further by taking the advantage of residual connections to ease the learning process. Similarly, I3D[Carreira and Zisserman, 2017] is proposed to use the Inception network[Szegedy *et al.*, 2015] as the backbone network rather than residual networks to learn video representations. Recently, there are many frameworks proposed to improve 3D convolution[Tran *et al.*, 2018; Xie *et al.*, 2018; Zhou *et al.*, 2018; He *et al.*, 2018]. However, all of the methods still suffer from an order of magnitude more computational cost than their 2D competitors due to the newly added temporal dimension, which makes the models difficult to train and unpractical in real-world applications. To decrease the number of parameters, these works[Sun *et al.*, 2015; Qiu *et al.*, 2017] decompose a 3D convolution kernel into a combination of a 2D spatial kernel and a 1D temporal kernel.

3 MRSTs and Deep MRST Network

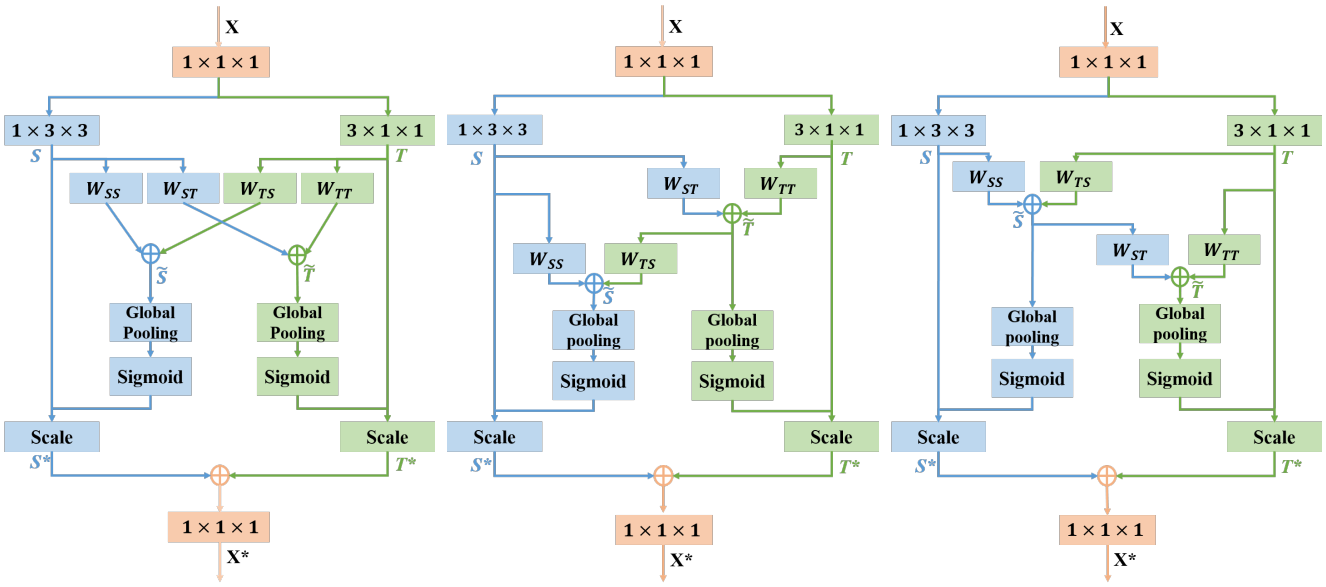
In this section, we first give a detailed description of the three types of MRSTs. Each MRST consists of a spatio-temporal decomposition (STD) unit, a mutually reinforced unit and a spatio-temporal fusion (STF) unit. We then present our robust and efficient deep network, MRST-Net for human action recognition.

3.1 MRSTs

We design the three types of MRSTs according to the different order of spatial and temporal information enhancement, as shown in Fig.2. Each MRST consists of a spatio-temporal decomposition (STD) unit, a mutually reinforced unit and a spatio-temporal fusion (STF) unit. The STD unit aims to decompose the 3D input signal into spatial features and temporal features with a 2D convolution and a 1D convolution. The mutually reinforced unit learns the correlation between spatial and temporal features by four fully connected layers, and utilizes the correlation to mutually enhance the discriminative capability of the spatial features and temporal features. The STF unit selectively emphasizes informative spatial and temporal features and suppresses useless ones and fuses them together to obtain the efficient spatio-temporal features.

Spatio-Temporal Decomposition Unit

The spatio-temporal decomposition (STD) unit is designed to reduce the high computational complexity of 3D convolution.



(a) MRST-P (b) MRST-S (c) MRST-T

Figure 2: The detailed architecture of the three types of MRSTs, which selectively emphasize informative spatial and temporal features by the guidance of the learned correlation between spatial and temporal information. (a) MRST-P: a parallel structure, which reinforces the spatial and temporal features simultaneously. (b) MRST-S: a series structure, which first reinforces the temporal features and then reinforces the spatial features. (c) MRST-T: a series structure, which first reinforces the spatial features and then reinforces the temporal features.

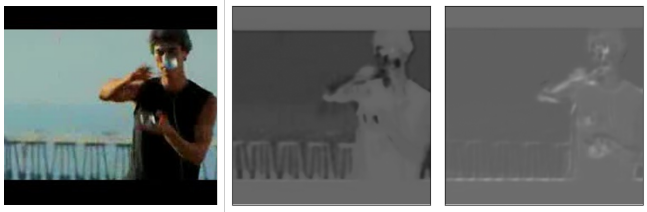


Figure 3: Visualization of the spatial and temporal feature maps. The left picture is a frame extracted from the juggling balls video in UCF-101; The middle picture is a spatial feature map of the left frame, which contains mainly spatial appearance; The right picture is a temporal feature map of the left frame and its adjacent frames, which contains mainly temporal motion.

It decomposes the 3D convolutional kernel into a $1 \times 3 \times 3$ convolutional kernel and a $3 \times 1 \times 1$ convolutional kernel. The $1 \times 3 \times 3$ convolutional kernels are used to extract spatial features $\mathbf{S}$ and the $3 \times 1 \times 1$ convolutional kernels are used to extract temporal features $\mathbf{T}$. The number of a STD unit's parameters is an order of magnitude less than that of an original 3D convolutional kernel, which could reduce the model size enormously. We additionally employ two $1 \times 1 \times 1$ convolutions at both ends of the path, which are applied to reduce and restore the channel dimensions respectively, in order to decrease the overall computational cost of MRSTs.

Mutually Reinforced Unit

The mutually reinforced unit is designed to generate the reinforced spatial and temporal representation, consisting of four fully connected layers. It takes the primitive spatial features $\mathbf{S}$ and temporal features $\mathbf{T}$ as input and produces the reinforced spatial information $\tilde{\mathbf{S}}$ and temporal information $\tilde{\mathbf{T}}$. As shown in Fig.3, we find that the primitive spatial features mainly concentrate on the objects with surrounds, such as the human appearance and the baluster. Besides, the primitive temporal features mainly concentrate on the objects which are in continuous motion, such as the human's arms and the ball. Based on the above observation, we develop a mutually reinforced unit to use the primitive spatial and temporal features to generate reinforced spatial and temporal representation. With the mutually reinforced unit, for spatial features, we can pay more attention to the objects which are in continuous motion. As for the temporal features, we can concentrate more on objects' surrounds. Specially, in Fig.3, we reinforce spatial features by suppressing useless features, which are irrelevant to sports, such as the baluster. In the meantime, we reinforce temporal features by learning the appearance of motion, such as the human's arms and the ball.

For a 3D input signal $\mathbf{X}$, the primitive spatial features $\mathbf{S}$ and temporal features $\mathbf{T}$ generated by the STD unit have the same size $L \times H \times W \times C$, where L, H, W, C refers to the length, height, width and the number of channels, respectively. We propose three different MRST mutually reinforced units which differ in the order of getting the reinforced spatial and temporal information, named as MRST-P unit, MRST-S unit and MRST-T unit, respectively. Detailed structures are provided as following:

(1) MRST-P unit: As shown in Fig.2(a), the MRST-P unit adopts a parallel structure, which reinforces the spatial and temporal features simultaneously. It consists of four ful-

ly connected networks. The temporal-to-spatial (TS) fully connected network is designed to learn the temporal features' guiding effect on spatial features. The spatial-to-spatial (SS) fully connected network learns the impact of spatial features on itself. Then the output of TS fully connected network and SS fully connected network are added together as the reinforced spatial information $\tilde{\mathbf{S}}$. Similarly, the spatial-to-temporal (ST) fully connected network and the temporal-to-temporal (TT) fully connected network are designed to learn the spatial features' guiding effect and temporal features' impact on the temporal features, respectively. The reinforced temporal information $\tilde{\mathbf{T}}$ are obtained by adding the output of ST and TT fully connected layer. The formulation of obtaining the reinforced information is shown as follows:

$$\begin{aligned} \tilde{\mathbf{S}} &= \mathbf{W}_{SS} \cdot \mathbf{S} + \mathbf{W}_{TS} \cdot \mathbf{T} \\ \tilde{\mathbf{T}} &= \mathbf{W}_{ST} \cdot \mathbf{S} + \mathbf{W}_{TT} \cdot \mathbf{T} \end{aligned} \tag{1}$$

where $\mathbf{W}_{SS}$, $\mathbf{W}_{ST}$, $\mathbf{W}_{TS}$ and $\mathbf{W}_{TT}$ refer to the parameters of the four fully connected layers, respectively.

(2) MRST-S unit: As shown in Fig.2(b), the MRST-S unit employs a series structure, consisting of four fully connected layers. First of all, a ST fully connected network and a TT fully connected network are designed to learn the spatial features' guiding effect and temporal features' impact on the temporal features, respectively, for producing the reinforced temporal information $\tilde{\mathbf{T}}$. After that, the reinforced spatial information $\tilde{\mathbf{S}}$ are obtained based on the primitive spatial features $\mathbf{S}$ and the reinforced temporal information $\tilde{\mathbf{T}}$. The formulation of the MRST-S unit is shown as follows:

$$\begin{aligned} \tilde{\mathbf{T}} &= \mathbf{W}_{ST} \cdot \mathbf{S} + \mathbf{W}_{TT} \cdot \mathbf{T} \\ \tilde{\mathbf{S}} &= \mathbf{W}_{SS} \cdot \mathbf{S} + \mathbf{W}_{TS} \cdot \delta(\tilde{\mathbf{T}}) \end{aligned} \tag{2}$$

where δ refers to the ReLU function.

(3) MRST-T unit: As shown in Fig.2(c), the MRST-T unit utilizes a series structure and consists of four fully connected layers. Firstly, a TS fully connected network and a SS fully connected network are designed to learn the temporal features' guiding effect and spatial features' impact on the spatial features, respectively, for producing the reinforced spatial information $\tilde{\mathbf{S}}$. Afterwards, the reinforced temporal information $\tilde{\mathbf{T}}$ is obtained based on the primitive temporal features $\mathbf{T}$ and the reinforced spatial information $\tilde{\mathbf{S}}$. The formulation of the MRST-T unit is is shown as follows:

$$\begin{aligned} \tilde{\mathbf{S}} &= \mathbf{W}_{SS} \cdot \mathbf{S} + \mathbf{W}_{TS} \cdot \mathbf{T} \\ \tilde{\mathbf{T}} &= \mathbf{W}_{ST} \cdot \delta(\tilde{\mathbf{S}}) + \mathbf{W}_{TT} \cdot \mathbf{T} \end{aligned} \tag{3}$$

Spatio-Temporal Fusion Unit

The spatio-temporal fusion (STF) unit aims to selectively emphasize informative spatial and temporal features, which can be interpreted as a means of biasing the allocation of available computational resources towards the most informative components of features, and fuses them to obtain effective spatio-temporal features.

For any given transformation $\mathbf{U} = \mathbf{F}_{tr}(\mathbf{X})$, where $\mathbf{X}$ is a three-dimensional input and $\mathbf{F}_{tr}$ refers to a 3D convolutional operator. We suppose $\mathbf{U} \in \mathbb{R}^{L \times H \times W \times C}$, and we can write it as $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_C]$, where $\mathbf{u}_c$ refers to the parameters of the c-th channel in $\mathbf{U}$. We first use the global average pooling operation to generate channel-wise statistics. Formally, a statistic $\mathbf{z} \in \mathbb{R}^C$ is generated by shrinking $\mathbf{U}$ through its spatio-temporal dimensions $L \times H \times W$, such that the c-th element of $\mathbf{z}$ is calculated by:

$$z_c = \mathbf{F}_{gp}(\mathbf{u}_c) = \frac{1}{L \times H \times W} \sum_{k=1}^{L} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(k, i, j) \tag{4}$$

To ensure that multiple channels are allowed to be emphasised, we follow the global pooling operation with a sigmoid function.

$$\mathbf{s} = \sigma(\mathbf{z}) \tag{5}$$

where σ refers to the sigmoid function. The final output of the STF unit $\mathbf{X}^*$ is obtained by rescaling the transformation output $\mathbf{U}$ with the activation:

$$\mathbf{x}_c^* = \mathbf{F}_{scale}(\mathbf{u}_c, s_c) = s_c \cdot \mathbf{u}_c \tag{6}$$

where $\mathbf{X}^* = [\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_C^*]$, and the scale function refers to channel-wise multiplication between the scalar s_c and the feature map $\mathbf{u}_c$. Particularly, in our MRSTs, the STF unit takes the spatial features $\mathbf{S}$, the temporal features $\mathbf{T}$, the reinforced spatial information $\tilde{\mathbf{S}}$ and the reinforced temporal information $\tilde{\mathbf{T}}$ as inputs and the final output can be calculated by:

$$\begin{aligned} \mathbf{S}^* &= \mathbf{F}_{scale}(\mathbf{S}, \sigma(\mathbf{F}_{gp}(\tilde{\mathbf{S}}))) \\ \mathbf{T}^* &= \mathbf{F}_{scale}(\mathbf{T}, \sigma(\mathbf{F}_{gp}(\tilde{\mathbf{T}}))) \\ \mathbf{X}^* &= \mathbf{S}^* + \mathbf{T}^* \end{aligned} \tag{7}$$

where σ refers to the sigmoid function, and $\mathbf{F}_{gp}$ represents the global average pooling operation, which is shown in Eq.4.

3.2 Deep MRST Network

We propose a simple yet efficient MRST-Net by stacking the MRST blocks. The proposed MRST-Net contains an initial convolutional block, five residual convolutional blocks and a LSTM [Hochreiter and Schmidhuber, 1997] layer. The initial convolutional block consists of six convolutional layers, which extracts primitive features. Each of the five residual convolutional blocks contains several MRSTs with Batch Normalization (BN) and Rectified Linear Units (ReLU). The main idea of the residual convolutional block is to learn the additive residual function with reference to the unit inputs which is realized through a shortcut connection, instead of directly learning unreferenced non-linear functions [He *et al.*, 2016]. Moreover, the Max-Pooling operation is performed with the initial convolutional block and each residual convolutional block, reducing the size of the feature maps (All the Max-Pooling layers' kernel size are $1 \times 2 \times 2$ and strides are (1,2,2)). We also add a Max-Pooling layer and a reshape operation after the conv6_x residual block to get the LSTM input. The LSTM layer progressively takes each time's feature output from the last residual block as input and decides whether to retain or discard the features from the current time and previous ones. Finally, a fully connected layer is designed to predict the classification of the input video sequence. More details of the network architecture are provided in Table 1.

MRST-Net		
layer	output size	kernel size
conv1	$L \times 112 \times 112$	$1 \times 7 \times 7, 64 \quad 3 \times 1 \times 1, 64$ $(1 \times 1 \times 1, 64) \times 4$
conv2_x	$L \times 56 \times 56$	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 64 \quad 3 \times 1 \times 1, 64 \\ (1 \times 1 \times 1, 64) \times 4 \\ 1 \times 1 \times 1, 128 \end{bmatrix} \times 3$
conv3_x	$L \times 28 \times 28$	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 64 \quad 3 \times 1 \times 1, 64 \\ (1 \times 1 \times 1, 64) \times 4 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 4$
conv4_x	$L \times 14 \times 14$	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 128 \quad 3 \times 1 \times 1, 128 \\ (1 \times 1 \times 1, 128) \times 4 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 12$
conv5_x	$L \times 7 \times 7$	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \quad 3 \times 1 \times 1, 256 \\ (1 \times 1 \times 1, 256) \times 4 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 11$
conv6_x	$L \times 4 \times 4$	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \quad 3 \times 1 \times 1, 256 \\ (1 \times 1 \times 1, 256) \times 4 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 3$
Max Pool	$L \times 2048$	$1 \times 2 \times 2$, stride (1,2,2) and reshape
LSTM	$L \times 1024$	1024 hidden units
average temporal pooling, fc layer with softmax		

Table 1: Architecture of the deep MRST-Net. It has one initial convolutional block and five residual convolutional blocks and a LSTM layer. The Max-Pooling layers after the first layer of each convolutional block are omitted for simplification. The details of residual convolutional blocks are shown in brackets, next to the number of times each block is repeated in the stack. The dimensions of filters and outputs are given by time, height, and width.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. Three well-known benchmarks, i.e., Kinetics-400[Kay *et al.*, 2017], UCF-101[Soomro *et al.*, 2012], and HMDB-51[Kuehne *et al.*, 2013], are included in the evaluations. The large-scale Kinetics-400 benchmark dataset consists of approximately 300, 000 videos from 400 action categories. UCF-101 consists of 13,320 manually labelled videos from 101 action categories. HMDB51 is collected from various sources, e.g. web videos and movies, which proves to be realistic and challenging. It consists of 6,766 manually labelled clips from 51 categories. Both UCF101 and HMDB51 are provided with 3 splits for training and testing and we report the accuracy by averaging over all 3 splits.

Implementation details. Our data augmentation includes random clipping on both spatial (firstly resizing the smaller video side to 256 pixels, then randomly cropping a 224×224 patch) and temporal (randomly picking the starting frame among those early enough to guarantee a desired number of frames). We average clip predictions which are uniformly sampled from the long video sequence to obtain the video predictions. Batch normalization is applied to all convolutional layers. We use the Adam Gradient Descent optimizer with an initial learning rate of $1e^{-4}$ to train the MRST-related networks from scratch. The drop out ratio and weight decay rate are set to 0.5 and $5e^{-5}$. We set the initial learning rate is $1e^{-5}$ and utilize the gradient descent optimizer with a momentum of 0.9 to train our MRST-related networks initialized with the ImageNet-1k and Kinetics-400 pre-trained model.

method	UCF101	HMDB51	size
C3D	85.3%	58.5%	321M
STD	86.9%	60.8%	209M
STD(with STF)	88.8%	63.8%	213M
MRST-P	91.2%	67.2%	211M
MRST-S	90.8%	66.4%	211M
MRST-T	**92.2%**	**68.9%**	211M

Table 2: Ablation study. Performance comparison to C3D, STD, STD with STF on UCF101 and HMDB51. They use the same network structure as C3D and they are all pre-trained on Kinetics-400. The complexity is measured using model size.

We employ the higher drop out ratio of 0.9 and the weight decay rate of $5e^{-4}$ to prevent over-fitting.

4.2 Ablation Study

To demonstrate the effectiveness of each component of the proposed MRSTs, we conduct a series of ablation experiments on UCF-101 and HMDB51 datasets. We also make a comparison with C3D[Tran *et al.*, 2015], the baseline 3D CNN. We choose C3D as our baseline since it is the most direct way to demonstrate the effectiveness of our final MRSTs. For fairness, all architectures use the same backbone (with 8 convolutional layers, 5 max-pooling layers, and 2 fully connected layers) and the same input $16 \times 3 \times 224 \times 224$ (16 is the number of frames in each input clip). Moreover, they are all pre-trained on the Kinetics-400 datasets.

Table 2 shows the comparison results in terms of the Top-1 classification accuracy and the model size on both UCF-101 and HMDB51 datasets. From Table 2, we can observe that our three MRSTs significantly outperform the baseline approach C3D. We also find that compared to the STD unit, the STD unit with STF improves accuracy by 1.9% on UCF-101 and 3% on HMDB51, which can prove the efficiency of attention recalibration. In addition, all the three MRSTs outperform the STD (with STF) unit, which strongly demonstrates the effectiveness of the mutually reinforced unit. Meanwhile, we find that among all three MRSTs, the MRST-T yields the best performance. Compared to the MRST-P, the MRST-T increases accuracy by 1% and 1.7% on UCF-101 and HMDB51 respectively. And compared to the MRST-S, the MRST-T improves accuracy by 1.4% on UCF-101 and 1.7% on HMDB51. We conjecture that the human action recognition task relies more on temporal features, namely the objects which are in continuous motion. Moreover, some objects with surround in spatial features may interfere with extracting effective attentions. As for some tasks rely more on spatial features, such as the video person re-identification task, the MRST-S may get better performance than the others. We also evaluate the model size in all the six architectures. It is clear that the model size of MRST is much smaller than the C3D and is close to the STD model size, which proves that our MRSTs consume less computational resources.

4.3 Comparison to the State-of-the-Art Methods

We further demonstrate the advances of our proposed MRST-Net in comparison with state-of-the-art methods for action recognition. All methods use only RGB input. Based on the

Method	Backbone	Input×clips number	Kinetics-400	#Params	FLOPs
C3D[Carreira and Zisserman, 2017]	-	$[16 \times 3 \times 224 \times 224] \times 1$	56.1%	79.0M	296.7G
LRCN[Carreira and Zisserman, 2017]	-	$[25 \times 3 \times 224 \times 224] \times 1$	63.3%	9.0M	41.5G
Nonlocal-I3d[Wang *et al.*, 2018b]	ResNet50	$[128 \times 3 \times 224 \times 224] \times 1$	67.3%	35.3M	145.7G
ARTNet[Wang *et al.*, 2018a]	ResNet18	$[16 \times 3 \times 112 \times 112] \times 25$	69.2%	35.2M	25.7G
I3D-RGB[Carreira and Zisserman, 2017]	BN-Inception	$[All \times 3 \times 256 \times 256] \times 1$	71.1%	12.7M	544.4G
StNet[He *et al.*, 2018]	ResNet101	$[25 \times 15 \times 256 \times 256] \times 1$	71.4%	52.2M	310.5G
R(2+1)D-RGB[Tran *et al.*, 2018]	ResNet34	$[32 \times 3 \times 112 \times 112] \times 10$	72.0%	63.8M	152.4G
S3D[Xie *et al.*, 2018]	BN-Inception	$[All \times 3 \times 224 \times 224] \times 1$	72.2%	8.8M	518.6G
MF-Net[Chen *et al.*, 2018]	-	$[16 \times 3 \times 224 \times 224] \times 50$	72.8%	8.0M	11.1G
MRST-T(ours)	ResNet101*	$[16 \times 3 \times 224 \times 224] \times 20$	**74.1%**	31.7M	99.6G

Table 3: Performance comparison with the state-of-the-art results on Kinetics-400 with only RGB frames as inputs. Here, "All" means using all frames in a video. #Params means the total number of model parameters. FLOPs means floating point operations.

results in section 4.2, we use the MRST-T (which gets the best performance among three MRSTs) and ResNet101* (compared to the base ResNet101, we add a spatial pooling layer and change the channel numbers of some layers, as shown in Table 1.) structure as backbone. In order to control the number of parameters of the whole network, we use the input clip size as $16 \times 3 \times 224 \times 224$. We uniformly sample 20 clips per video and average these 20 clip predictions to obtain the video prediction. Related results on Kinetics-400, UCF101 and HMDB51 are shown in Tables 3 and 4, respectively.

Results on Kinetics-400. Table 3 shows the performance comparison of our proposed MRST-T-Net (pre-trained on ImageNet-1k) against nine state-of-the-art methods in terms of Kinetics-400 Top-1 classification accuracy and the total number of parameters and the FLOPs. The compared methods use only RGB inputs and they have almost the same backbone. The proposed MRST-T-Net achieves 74.1% Top-1 classification accuracy, and the total number of parameters is 31.7M, and the FLOPs is 99.6G. We can see that our method surpasses existing methods, improving the 2nd best compared method MF-Net by 1.3% at Top-1 classification accuracy. Moreover, our MRST-T-Net achieves significant performance improvement compared to the baseline C3D by 18% at Top-1 classification accuracy. Besides, the total number of parameters of our proposed MRST-T-Net is no more than half of the C3D. Compared to the same backbone structure ResNet101-StNet[He *et al.*, 2018], our network has fewer parameters. We can also see that the computational cost (FLOPs) of MRST-T-Net is lower than those of most existing 3D CNN based methods. The comparisons indicate that the MRST-T-Net can learn and represent spatio-temporal features much more efficiently and accurately than other methods.

Results on UCF-101 and HMDB51. We also experiment with fine-tuning MRST-T-Net (pre-trained on ImageNet-1k and Kinetics-400) on UCF-101 and HMDB51 to evaluate the generality and robustness. From Table 4, we can observe that the proposed MRST-T-Net outperforms almost all the existing state-of-the-art methods with only RGB inputs on both UCF-101 and HMDB51, obtaining 96.5% Top-1 classification accuracy on UCF-101 and 75.4% Top-1 classification accuracy on HMDB51. In addition, MRST-T-Net boosts the baseline method C3D by 14.2% at Top-1 accuracy on UCF-101 and 23.8% at Top-1 accuracy on HMDB51. The only method that has better performance on UCF-101 than ours is the R(2+1)D-34 layers network, but our proposed model require fewer parameters and less computational cost.

Method	UCF101	HMDB51
Two-stream[Simonyan and Zisserman, 2014]	73.0%	40.5%
C3D[Tran *et al.*, 2015]	82.3%	51.6%
ST-ResNet-50[Feichtenhofer *et al.*, 2017]	82.3%	48.9%
ST-ResNet-152[Feichtenhofer *et al.*, 2017]	83.4%	46.7%
TSN[Wang *et al.*, 2016]	85.7%	54.6%
Res3D[Tran *et al.*, 2017]	85.8%	54.9%
P3D ResNet[Qiu *et al.*, 2017]	88.6%	-
MiCT-Net[Zhou *et al.*, 2018]	88.9%	63.8%
ARTNet[Wang *et al.*, 2018a]	94.3%	70.9%
I3D-RGB[Carreira and Zisserman, 2017]	95.6%	74.8%
MF-Net[Chen *et al.*, 2018]	96.0%	74.6%
R(2+1)D-34-RGB[Tran *et al.*, 2018]	**96.8%**	74.5%
MRST-T(ours)	96.5%	**75.4%**

Table 4: Action recognition accuracy on UCF-101 and HMDB51, averaged over three splits. The top part of the table refers to related methods with the Sports-1M pre-trained, the lower part refers to related methods with the Kinetics-400 pre-trained.

5 Conclusion

In this work, we propose a mutually reinforced Spatio-Temporal Convolutional Tube (MRST) for human action recognition. It decomposes 3D inputs into spatial and temporal representations, mutually enhances both of them by exploiting the interaction of spatial and temporal information, and fuse them for extracting effective spatio-temporal features, while reducing the computational complexity. Experiment results show that the MRST-T yields the best performance among three MRSTs and significantly outperforms traditional 3D CNNs for action recognition. Moreover, the MRST-T-Net achieves the best performance on three datasets, Kinetics-400, UCF-101 and HMDB51 in comparison to state-of-the-art approaches, indicating that the proposed MRST-T-Net is general and efficient.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants 61622211 and 61620106009 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

References

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[Chen *et al.*, 2018] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *ECCV*, pages 352–367, 2018.

[Donahue *et al.*, 2015] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015.

[Feichtenhofer *et al.*, 2017] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, pages 4768–4777, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2018] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Liming Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. *arXiv preprint arXiv:1811.01549*, 2018.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jurgen Schmidhuber. *Long short-term memory*, volume 9. MIT Press, 1997.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[Ji *et al.*, 2013] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–231, 2013.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[Kuehne *et al.*, 2013] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *HPCSE*, pages 571–582. Springer, 2013.

[Qiu *et al.*, 2017] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5533–5541, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[Sun *et al.*, 2015] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *ICCV*, pages 4597–4605, 2015.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[Tran and Cheong, 2017] An Tran and Loong-Fah Cheong. Two-stream flow-guided convolutional attention networks for action recognition. In *ICCV*, pages 3110–3119, 2017.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[Tran *et al.*, 2017] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017.

[Tran *et al.*, 2018] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.

[Wang *et al.*, 2016] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016.

[Wang *et al.*, 2018a] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *CVPR*, pages 1430–1439, 2018.

[Wang *et al.*, 2018b] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[Xie *et al.*, 2018] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

[Zhou *et al.*, 2018] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Mict: Mixed 3d/2d convolutional tube for human action recognition. In *CVPR*, pages 449–458, 2018.